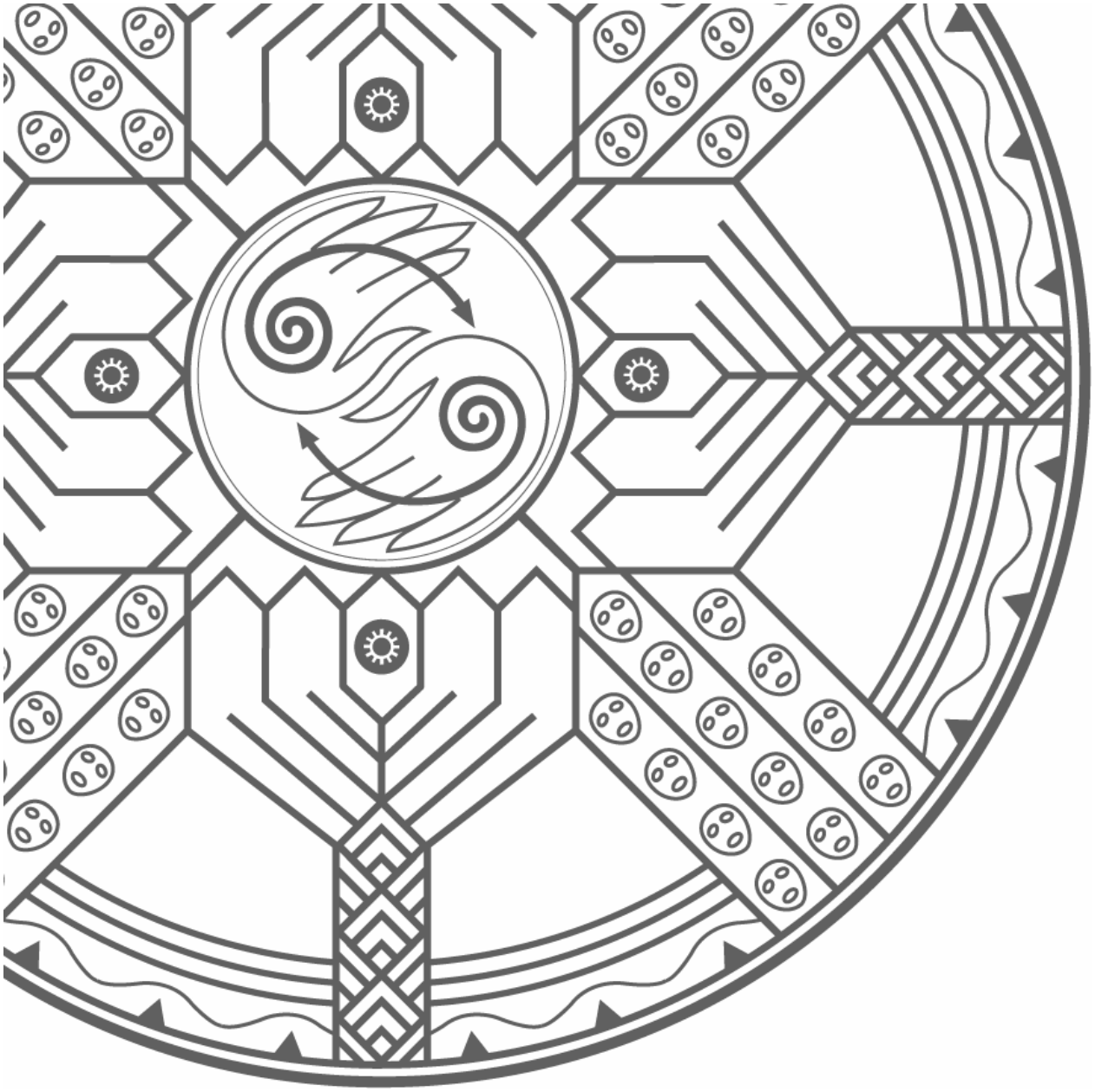


Tools for Assessing Web Site Usage

Scott Anderson, Terri Willard, Heather Creech and Deborah Bakker

IISD
INTERNATIONAL INSTITUTE FOR SUSTAINABLE DEVELOPMENT
INSTITUT INTERNATIONAL DU DÉVELOPPEMENT DURABLE

IISD contributes to sustainable development by advancing policy recommendations on international trade and investment, economic instruments, climate change, measurement and indicators, and natural resource management. By using Internet communications, we report on international negotiations and broker knowledge gained through collaborative projects with global partners, resulting in more rigorous research, capacity building in developing countries and better dialogue between North and South.

IISD's vision is better living for all—sustainably; its mission is to champion innovation, enabling societies to live sustainably. IISD receives financial support from the governments of Canada and Manitoba, other governments, UN agencies, foundations and the private sector. IISD is registered as a charitable organization in Canada and has 501 (c) (3) status in the United States.

Published by the International Institute for Sustainable Development

Printed in Canada

Copies are available from IISD.

International Institute for Sustainable Development
161 Portage Avenue East, 6th Floor
Winnipeg, Manitoba
Canada R3B 0Y4
Tel: +1 (204) 958-7700
Fax: +1 (204) 985-7710
E-mail: info@iisd.ca
Internet: http://www.iisd.org

Table of Contents

Executive Summary

Web sites are now a common component of network communications with external audiences. This paper is an examination of several tools for assessing the effectiveness of a web site. Is the knowledge of the network reaching its target audiences? Is the presentation of that knowledge usable by the audience?

The paper reviews methods employed by the Sustainable Development Communications Network (SDCN)[1] to assess use of the SD Gateway, the network's web site: log file analysis, user testing, user feedback and user surveys. The SD Gateway (http://sdgateway.net) was developed as the SDCN's primary vehicle for reaching the following objectives:

- Increasing the amount and visibility of developing country content on the Internet;
- Assisting decision-makers to make use of information from network members; and
- Attracting and engaging new audiences to promote action on sustainable development.

In our efforts to monitor our progress towards these objectives, we tested a number of evaluation tools. Our findings are based on the application of log file analysis and user assessments for the SD Gateway from 1997 to 2000.

We believe that log file analysis can no longer provide a clear, reliable picture of the number of users of a web site, including geographic distribution and sector representation. Nevertheless, log file analysis can provide useful benchmarks and comparative measures to improve site structure, management and marketing efforts. For example, does the traffic increase or remain static when a new product is launched? Are some databases or other products on a web site consistently receiving more use than others? Some very broad estimates can be made of the number of direct users (as opposed to users who access web pages through intermediary caches on their own computers or on other servers). We strongly recommend that web communicators look to additional tools to assess the use and effectiveness of a web site, including user feedback and surveys and web site testing.

Every web communicator should start with Jakob Nielsen's *Designing Web Usability: The Practice of Simplicity.*[2] Nielsen's regular e-mail newsletter, the Alertbox, can be subscribed to at http://www.useit.com/alertbox. The web site for Analog, a log file analysis software provider, also provides information on how the web works and important considerations about analysis of web site statistics and user assessment (http://www.analog.cx/docs/webworks.html).

Based on our experience, there are three reasons for investing time and resources in web site analysis:

- To improve content and navigation on your web site;
- To evaluate the impact of marketing and outreach activities to target audiences; and
- To gauge system performance and technical requirements.

[1] For a list of current SDCN member organizations, see http://sdgateway.net.
[2] Jakob Nielsen, *Designing Web Usability: The Practice of Simplicity,* New Riders Publishing, Indianapolis, 2000, ISBN 1-56205-810-X.

The five key lessons for web communicators and network managers are surprisingly simple:

1. **Analyze server log files regularly and act on findings.** Run an analysis of your web server log files monthly, with quarterly and annual totals. Once several months of data are in hand, it is possible to look for signals related to access and navigation, trends and spikes resulting from marketing efforts, and indicators of system performance. The most useful variables to monitor for internal purposes are:
   - page requests (as a rough equivalent to amount of information used);
   - visits (as a rough equivalent to number of users); and
   - PDF file downloads and subscriptions to mail lists with web archives (as an indication of interest in specific content).

   More detailed analysis of log files can help to focus attention on:
   - heavy traffic areas;
   - user access to the site via search engines and other links;
   - additional or different meta tags required for improved retrieval by search engines;
   - maintenance of broken links; and
   - server load and processing times.

2. **Site statistics can be very misleading**. While they are useful to provide signals and baseline data, one should not infer too much from site statistics. Site statistics should not be used for corporate promotional purposes—it is not possible to claim levels of traffic, geographic and sectoral reach with any definitive degree of accuracy. Moreover, site statistics should never be used to compare one institution's web performance against another's.

3. **Focus on the user**. Remember that engaging and influencing users is your ultimate goal. You can determine their interests, motivations and opinions through log-on forms, feedback forms and user surveys. User testing, though, is the only reliable way to find out how they actually use the site. Only through user testing will you get the information required to design and maintain a site that will allow users to access content as efficiently as possible.

4. **Combine methods**. A single data collection method will not provide an accurate representation of users' needs and whether those needs were met with the content on the site. Log file analysis has its place when combined with other user assessment methodologies.

5. **Remember the big picture** within your organization and for sustainable development. Web site usage information can be valuable for people with different priorities—for systems administrators, researchers, content managers, designers and marketing managers. Web site analysis will help to improve the communication of your sustainable development knowledge, with improved impact on target audiences and increased global awareness of actions needed and solutions available.

1.0 Reaching Audiences Through the Web

The Sustainable Development Communications Network (SDCN) is a knowledge network of organizations working together to test new methods of collaborative communications on sustainable development, with a particular emphasis on Internet technologies. The outputs of the SDCN are twofold: increased understanding of how to use information and communications technologies; and new content on the Internet about sustainable development from leading institutes around the world.

A knowledge network can be defined as a group of expert institutions working together on a common concern, strengthening each other's research and communications capacity, sharing knowledge bases and developing solutions that are made available for use by others outside the network.[3] The effectiveness of a knowledge network like the SDCN can be assessed by the composition and level of interest of its external audience—those individuals who are knowledgeable about the network's research and who may be willing to act upon the network's recommendations. Press releases, media events, book launches, conference presentations and e-mail are all used today to alert decision-makers to new research emanating from organizations and networks. The success of this marketing is best measured by how well it convinces the target audiences to follow up with the network and to seek additional information or advice. The measure of success is not how many people have been sent a message, but rather how many respond to the message. Users respond by e-mailing queries to the network and its member organizations, by accessing the network web site and by building a relationship with regular contact and interaction over time.

There are progressive levels to this relationship-building process. For the SDCN it begins with notifications broadcast through listservs and other web sites of new products and actions of the network. The next step is to provide information to more interested individuals, upon their request, on specific sustainable development issues. We at the SDCN track these types of direct requests in order to establish a better profile of the existing audience and how well it matches the target audience desired. However, tracking and supporting these relationships can be resource-intensive. For that reason, many organizations have begun to use the Internet web as a strategic tool to assist with this direct response level of relationship building. Marketing techniques are used to get messages out about new research and solutions, but audiences are then advised to get more information from the web site. Publications and general research results are posted on the web to ensure that many individuals have access to them.

However, while the web has greatly increased the reach of sustainable development knowledge networks and organizations, use of the web has in fact reduced the ability to analyze and monitor that reach. While we may know who orders a book from us or who attends a conference, we do not know exactly who visits our web site. In exchange for our knowledge, we learn very little about who our audiences actually are, what they really need and whether they are getting it from the web site.

[3] Heather Creech, Draft presentation on Knowledge Networks, IISD, 1999.

Consequently, organizations, networks and businesses are turning to web site analysis tools to determine who is using their web sites and what information is being accessed. The path a user takes through a web site, captured in the server log of pages accessed by the user, is often called the "clickstream." Clickstream data collection and analysis is currently estimated to be a US$200-million industry.[4]

Regardless of your organization's size, basic web site analysis can be useful for several purposes:

- **To improve content and navigation on your web site;**
- **To plan and evaluate the impact of marketing and outreach activities on target audiences; and**
- **To gauge system performance and technical requirements.**

Section 2 of this paper describes the process of log file analysis as a useful but imperfect tool for gathering and interpreting quantitative data on web site traffic. Section 3 focuses on what can be learned from detailed analysis of log files, to assist with content design, marketing and system requirements. Section 4 provides an overview of three other data collection tools and briefly describes their limitations and benefits. Section 5 describes how SDCN members applied web site usage studies to evaluate several experiments with marketing the web site. A short glossary of technical terms used in the paper is included in Appendix 1.

[4] Mubarak Dahir, "Just for Clicks," *The Standard*, May 8, 2000, <http://www.thestandard.com/article/display/0,1151,14727,00.html>.

2.0 Quantitative Data Analysis

Most managers want to know how many people visit their organization's web site, what they are reading and whether the web site is accessible to everyone. This section outlines how these questions can be answered and warns of some common pitfalls and software limitations that can be avoided while finding the answer.

- First, we identify five sources for data: counters, cookies, user log-on forms, web-based site analysis services and server log files.
- Second, we identify two important impacts on server log data: caching and IP addressing.
- Third, we review special needs for data collection and interpretation. We describe in particular three variables which we consider somewhat more meaningful measurements that can be interpreted from log file analysis software: page requests, as an equivalent to amount of information used; visits, as an equivalent to number of users; and product downloads or subscriptions (PDF file downloads, subscriptions to mailing lists and use of their archives online) as an expression of interest in specific content.
- Fourth, we review what cannot be learned from log files and limitations to log file analysis.

2.1 Counters

Many organizations that want to learn more about the usage of their Web sites begin by installing a counter on their home page. We recommend against using this tool if you want to obtain any meaningful information about who is visiting your site. Many counters can be falsified either by setting the initial number higher than zero, or by clicking "refresh" (or "reload") repeatedly to raise the number of counts. Moreover, counters as a rule do not measure information across an entire site. Rather, they only record accesses for pages on which the counter is located. Finally, depending how it is installed, a counter may subvert caching mechanisms (see Section 2.4), making it more difficult for people to access the content of the site.[5]

2.2 Cookies

A "cookie" is a command sent by a web site to a user's computer to capture and store user information on the user computer. Information captured by the cookie can later be read back by the web site from that computer. Cookies are useful for storing passwords and preferences of start pages. Microsoft and Netscape rely on cookies to allow personalized versions of their home pages. Cookies can be used to personalize shopping baskets in an e-commerce setting. They can also be used obtain information about a computer (and the computer's user) for targeted marketing purposes. Use of cookies to collect user data has become controversial. "When the information collected online through cookies is combined with outside marketing databases a rather detailed user profile can be created without the user's knowledge. The ability to create such a detailed user profile has raised many concerns regarding the protection of consumer privacy."[6] If you do intend to implement cookies on your web site in order to collect user

[5] Jeff Goldberg, "Why web usage statistics are (worse than) meaningless," Cranfield University, August 24, 2000, <http://www.cranfield.ac.uk/docs/stats/>.

[6] Linda A. Goldstein, "Is the Internet Cookie about to Crumble," *Digitrends*, Spring 2000, <http://www.digitrends.net/digitrends/dtonline/features/00/dt_spring/cookie.shtml>.

information, you should, at a minimum, advise your users that cookies are in use. You should be aware that many users will not or cannot accept cookies. In addition, software is now readily available to allow anonymous surfing by removing cookies from users' computers. Therefore the data captured by your cookie will not provide a complete picture of users.

2.3 User log-on forms

Commercial sites such as Amazon.com have instituted "sign-in" forms to collect demographic and user preference information about their customers in order to customize services. Some NGOs are now experimenting with this approach, by asking new users to fill in a form online before they can access the web site. On submission of the form, the user is assigned a password that allows them to log on to the web site. The form may request everything from an e-mail address and institutional affiliation to more detailed demographics (age, gender, education level, occupation, place of residence) and interest areas. Password control allows site managers to track new and repeat users. We do not advise that you require users to complete such a form just to track user statistics more effectively. Many users are becoming increasing protective of their privacy on the Internet. If you require a user to log on, you may alienate many people who would otherwise have read or used your information.

However, we do see real value in requesting information from users as part of building relationship with target audiences. If you wish to market new content directly to your web site users, then you might want to make a sign-in form available at various points throughout your web site. Advise your users that in exchange for their e-mail address, selected demographic details and profile of interests, you will send them notices of new products relevant to their interests. There are some basic protocols to follow should you wish to capture user demographics and profiles of interests for marketing purposes, including:

- Making explicit to users what you intend to do with the data, including whether you intend to resell or pass that data along to other organizations/companies.
- Making the data entry form optional.

2.4 Web-based site analysis services

Web-based services and tools for auditing your web site traffic are now widely available. Free traffic analysis services are becoming more common, although they provide less thorough analyses than commercial services. Some will provide detailed reports for only a single page within a site. While these kinds of services provide a basic understanding of site traffic, they do not provide sufficient information for broader site management. Several sources of commercial and free tracking and analysis services are listed in Appendix 2.

"Bobby" is one example of a very useful, web-based tool to assess the general functionality of your web site. Bobby, available at http://www.cast.org/, was created by CAST (Center for Applied Special Technology) as a web-based tool to analyze web pages for their accessibility to people with disabilities. When you log onto Bobby and enter the URL for your web site (or anyone else's, for that matter!), you will be given a detailed report on browser compatibility, download times, layout, coding and other priority considerations.

2.5 Server log files

Based on our experience, server log file analysis provides considerable information about site usage. But it is important to understand how the web functions and the inherent limitations of data recording methods, in order to interpret the site statistics in a meaningful way.

If users follow a link to a home page, their browser will send a request to the server hosting the page asking for the page in question. If the page contains graphics, separate requests are sent to the server for each graphic file. If users navigate to other pages, new requests are sent to the server for each page and each graphic. Each request sent to the server is recorded in a log file. The log file generally records the following:

- the name and size of each file requested;
- the date and time of the request;
- the address (Internet Protocol—or IP—address) of the host (computer) that is making the request;
- a code indicating whether the request was successful or not;
- the URL of the referring page to the site; and
- the agent and operating system used by the computer making the request (the agent will be a browser like Netscape Navigator or an automated program such as a search engine indexing your site).

From this data, standard log server analysis software can compile the following[7] (see Appendix 1 for more detailed definitions of terms):

- total number of requests (hits);
- number of requests by file type;
- number of page requests (separate pages viewed);
- number of bytes transferred;
- number of hosts (distinct IP addresses) that requested information from your server, and the number of requests from each;
- number of requests by status code (successful, failed, redirected, etc.);
- totals and averages over time (minutes, hours, days, weeks, months);
- number of requests by domain;
- URLs from referring pages;
- browsers and versions requesting pages; and
- requests by operating system.

This type of information can be used to assess hardware and software performance (e.g., to plan for server upgrades), content development and management (e.g., identifying which features are being accessed), the effectiveness of site marketing (e.g., assessing whether some methods of

[7] Susan Haigh and Janette Megarity, "Measuring Web Site Usage: Log File Analysis," *Network Notes #57*, National Library of Canada, August 4, 1998, <http://www.nlc-bnc.ca/pubs/netnotes/notes57.htm>.

promotion are more effective than others in drawing users to the site), or interesting user preferences that can help shape content development.

There are two options for collecting and analysing the data from the server hosting your web site:

1. **Use your ISP's (Internet Service Provider) statistical package.**
   If an ISP hosts your web site, it might provide an analysis of the traffic on your site. Not all ISPs provide this level of user support. Those that do might only provide daily log file analysis, and not monthly or annual analyses since these exceed the ISP's server capacity. You will still need to become familiar with the features of the statistical package used by your ISP in order to interpret the data for your own purposes.

2. **Use log file analysis software yourself.**
   Your ISP might provide you with the raw data of accesses logged for your site. If you host your site on your own server you can capture this data yourself. A variety of software packages can be used to analyse one's own web site traffic. While you will have control of the information, there may be a significant learning curve, particularly to understand all the technical parameters. See Appendix 2 for examples of software packages.

2.6 Impacts of caching on server log data

More sites and more people surfing the Internet create network congestion, making pages load more slowly. This is tempered somewhat by caching, a process which speeds up document loading time by automatically storing copies of requested files at different physical locations that are more readily accessed than the actual server. While caching speeds up user access to content, the process limits the validity of log files as an estimate of site traffic because once a file is cached (after a first visit), repeat visits are not recorded in the log file. A browser will preferentially access a cached file on a local hard drive, rather than sending a request to the server where the file resides. Moreover, since frequently accessed pages will be cached more often, it is difficult to gauge relative popularity of different pages.

There are several levels to caching. The first level is local. When a page is requested, a browser will look first in its cache on a local hard drive. If a file is not found there, the browser will access a second level of caching at some intermediate location *en route* to the server, such as a cache set up by an ISP. Thirdly, a national or regional server cache may be accessed before a server is sent a request for a page. These intermediary servers are called proxy servers. Caching is becoming more common. In September 1999, StatMarket estimated that about 17 per cent of Internet traffic uses proxy servers,[8] and most large commercial ISPs, such as AOL, cache pages to speed up access for their clients.[9]

[8] *Proxy Servers on the Rise,* StatMarket – A WebSideStory Production, <http://statmarket.com/SM?c=stat091099>.
[9] Doug Linder, "Interpreting WWW Statistics," University of Heidelberg, no date, <http://www.zuv.uni-heidelberg.de/webstat/rem2.htm>.

2.7 Impacts of IP addressing on estimates of traffic

Internet Protocol (IP) addresses are the distinct numbers that identify each computer. ISPs have a limited number of IP addresses that are assigned to different users when they access the Internet. Therefore, a single IP address might actually represent several different users. And a single user might be associated with several different hosts if they access the Internet at different times or from different computers or service providers (e.g., from work or from home). In addition, AOL may allocate a different IP address for different types of requests. One user requesting a single page with one text element and nine graphic elements could be identified in the log file as 10 different computers[10] accessing a single page.

2.8 Data collection considerations and preliminary interpretations

The data collection process can be set up so as to produce some useable results. A few important considerations for data collection and interpretation are:

- **Filter log data.**

As a general rule, it is good practice to configure your log file analysis software to filter out internal requests from the log data (i.e., traffic from your own organization). At the SDCN, up to mid-2000, we also filtered the log data to eliminate search engine spiders and/or crawlers. These are programs that roam the web either indexing for a search engine or checking links. A host with a disproportionately large number of page requests is likely a spider.

Table 1: SD Gateway Traffic Statistics, January 2000.

Measure	Unfiltered Data	Data filtered for internal requests, search engines
# of Requests (Hits)	446,324	
# of Page Requests	75,219	46,081
# of Hosts	10,910	7,654
# of Countries	106	98

Upon examining one month's worth of data from the SD Gateway, we found that using the unfiltered information, the number of page requests is almost 70 per cent higher, the number of hosts is over 40 per cent higher and the number of countries is about eight per cent higher. Examining the unfiltered data, three of the top five organizations listed were spiders, and accounted for over 10 per cent of pages requested.

However, with the growth in web site traffic and the constant emergence of new spiders and other agents, system administrators are finding it more difficult to filter log data. And there is the argument that spiders and other agents do represent actual intended audiences. Your system administrator should inform you about any changes made in filtering data or changes in software used for log file analysis. Such changes will affect the numbers provided and will impact your monitoring of long-term trends on your web site.

[10] Stephen Turner, "How the Web Works," May 31, 2000, *Analog*, <http://www.analog.cx/docs/webworks.html>.

- **Focus on page requests rather than hits.**
  It is important to understand the distinction between these related but often mistakenly interchanged terms. A hit is a single request for a file from a server[11] (which could be text or a graphic) while a "page request" is counted when an entire page—graphics and text[12]—is downloaded. A count of hits is a highly unreliable indicator of traffic because text and graphics are counted separately, thereby leading to a significant overestimation of traffic. For example, downloading a web page containing one text element and nine graphic elements would register as 10 "hits" when only a single page was viewed. Since the number of graphics on each page can differ substantially within a site, and between different sites, using hits to gauge traffic to and within a particular site is a meaningless exercise. Our data for the SD Gateway suggest that we received 446,324 "hits" in January 2000, but this would conceal the fact that images (gif, jpg) accounted for 83 per cent of files requested, and 19 of the top 20 requested files were graphic files. In this case, the number of hits overestimated traffic by nearly a factor of 10 over the number of page requests for filtered data. The number of page requests gives a better indication of how much information users are accessing on your site.
  .
- **Focus on visits.**
  As opposed to page requests and hits, a visit, also known as a "user session," is defined as a series of consecutive page requests from a host to a site (i.e., in one "visit" a user could view several different pages of a site.). We consider visits to indicate a very rough estimate of the number of users on the SD Gateway, although we recognize that we cannot tell how many of these visits are from repeat users. If you are tracking visits, be aware that the time period to determine the end of a visit can be customized in the software. If a user makes no requests from a particular site during a certain period of time (usually set at 30 minutes), the next hit from their IP address would constitute the beginning of a new visit. This time-out factor can lead to an overestimation of the actual number of visits. However, because of caching and proxy servers, it is more likely that the "visits" count will always underestimate the total number of users working with your content in any given period.

  When used in connection with page requests, some broad generalizations about volume of traffic over time can be made, expressed as X number of users accessing Y pages from [date] to [date].

  We prefer visits to distinct IP addresses as a reflection of the number of users, since a distinct IP address could indicate not only a person navigating your site, but also an ISP representing multiple users.

- **Count product downloads and subscriptions.**
  In addition to page requests, we look at the number of downloads for individual PDF files we have attached to web pages (e.g., full text reports, books and other documents). We also look

[11] Brad Aronson, "Measuring the Web: What Server Logs Really Tell Us," Presentation given at Web Advertising '96 Conference, New York City, October 31 – November 1, 1996, <http://www.thunderlizard.com/tlp_pdfs/aronson.pdf>.
[12] Aronson, <http://www.thunderlizard.com/tlp_pdfs/aronson.pdf>.

at subscriptions to mailing lists and accesses via the web interfaces for those lists. This provides us with an indication of the value of that content to users.

- **Amalgamate data from mirror sites.**[13]
  A mirror site is a replica of an already existing site, used to reduce network traffic or improve the availability of the original site and is useful when the original site generates too much traffic for a single server to support.[14] If you operate mirror sites, it is important to add any log file analysis results to those of the main site for a more accurate picture of how information is being accessed.

- **Design the file structure of your web site to assist with data analysis.**
  When you are (re)designing your site, consider using a directory structure that will make discerning usage in different sections of the site easier. For example, the directory structure for earlier versions of the SD Gateway was very flat, and contained almost all files (in three languages) in one folder, making the analysis of discrete parts of the site very tedious and time consuming. When it was redesigned in March 2000, the files for each online product or service were given their own sub-directory and a consistent naming system was used to indicate the various language versions. We were then better able to assess levels of traffic on different sections of the site, and could clearly differentiate the traffic on the French and Spanish versions from the English for those different sections.

2.9 Special data collection considerations for knowledge networks

Knowledge networks can be established with a common web site acting as the "hub" for the network, but each organization in the network will have its own institutional web site as well. It is difficult to compare one organization's web traffic to another's, due to wide variations in log file analysis software programs, differences in definitions and terminology, and the impact of caching. Network managers who want to compare and/or aggregate the data from the institutional web sites of all network members as well as the network hub must agree to a common standard for log file analysis.

- **Use the same log file analysis software program.**
  Definitions for traffic analysis terminology are not standard. For instance, "request" in some software packages can mean "hit" or it can mean a "page request."[15] To allow for comparative analyses between organizations or their sites, it would be best if each organization in the network used the same log file analysis software. If this is not possible, it is essential to know what software package each organization in the network is using, what the programs measure and the assumptions made in the calculations. Analysis of the same data with different software could produce different results and lead to different conclusions.

---

[13] Teresa Elms, "A Web Statistics Primer," November 23, 1999, <http://builder.cnet.com/Servers/Statistics/>.
[14] "Mirror site," *Webopedia: Online Computer and Internet Dictionary*, August 22, 2000, internet.com, <http://webopedia.internet.com/TERM/m/mirror_site.html>.
[15] Elms, <http://builder.cnet.com/Servers/Statistics/>.

- **Establish a common "time out" period for visits.**
  If the log file analysis software being used counts visits, consider having network members establish a common "time out" period for visits. This will help correlate and compare data more effectively across the network.

- **Decide whether to filter spiders.**
  A spider is an automated program that explores the web, looking for information. The most common kinds of spiders (also called crawlers, robots or bots) are the ones that index sites for search engines, collect e-mail addresses or check links. If one member does not have the capacity to filter for spiders and other agents, then no member should filter any. Otherwise, it will not be possible to compare data with any degree of accuracy across the network. If members choose to filter these agents, then establish a common list of spiders, crawlers and robots that all members will filter from their institutional log files. The same agents should also be filtered from the log files of the network hub.

- **Decide whether to filter visits from network members.**
  Network members should choose either to filter from their institutional web sites all use from member IP addresses or to filter none. Most knowledge network members view their partner institutions as valid and important users of their web sites and choose not to filter them from their log analysis. Depending on the size of the user base, organizations participating in knowledge networks might wish to run two sets of log analyses: one that filters out member use, to highlight external audiences; and one that does not filter out member use, providing a more complete picture of the volume of activity across the network.

2.10 What cannot be learned from log files

It is important to remember that log files cannot provide an accurate record of web site traffic because some information is misleading and other information is not recorded. The following points are important variables that log files cannot measure reliably:

- **User identity.**
  Unless users are required to enter a password to log into a site, their identity cannot be known, nor can any demographic information (age, occupation, etc.) be determined.

- **Exact number of users.**
  Again, requiring users to log in is the only way to determine exactly how many people are visiting your site. The use of cookies can help somewhat with this problem, but they are not entirely dependable since many users will not or cannot accept cookies. In addition, software is now readily available to allow anonymous surfing by removing cookies from users' systems.

- **A user's entry point to your site and their path through it.**
  Because of caching, a user might have viewed one or more pages before their computer sends a request to your server, which is then recorded in a log file. The log file data would incorrectly identify that this page request was the entry point to the site, rather than the middle of a session. The ability to assess a user's path through your site is limited since files

that are reloaded using the back button will be loaded from the browser's cache rather than registering a request with the server.

- **Geographic location of users.**
  We at the SDCN are interested in the percentage of users who come from developing countries and countries in transition. Although we can get some indication of the number of computers from different countries accessing the site, it is impossible to accurately classify users by their country of origin. Log analysis software identifies point of origin as the location of IP address registration. Hence, anyone accessing a site through the commercial provider AOL will be identified as coming from Virginia, USA, regardless of a user's actual location. Moreover, some software wrongly assumes that common domains like .com, .net, and .org are American in origin, further complicating the issue. It is also important to note that national and regional caches might distort the levels of traffic from different regions. A recent development that emphasizes the confusion that can come from linking geography with Internet domains is the domain .nu. It is the domain for Niue, an island nation in the South Pacific, but it can also be purchased by anyone in the world. In 1997 it was the fastest growing domain name[16] and has become popular in Northern European countries where the word "nu" means new. A final point to remember is that a large proportion (20–70 per cent) of IP addresses are usually unidentifiable by log file analysis software, further complicating interpretation.

- **Identification of users by sector.**
  A similar barrier applies to classification of users by sector (academic, NGO, business, government). For instance, non-profit organizations are not restricted to using .org for their domain names. Many non-profits prefer to use a country domain, for example:

  - Environnement et développement du tiers monde (www.enda.sn); and
  - Stockholm Environment Institute (www.sei.se).

  And, of course, country domains will not reflect sector. For instance, any Canadian organization in any sector can use the .ca domain including universities (University of Toronto – www.utoronto.ca), government (Environment Canada – www.ec.gc.ca) and business (Toyota – www.toyota.ca). For the SD Gateway we have found that approximately 20–25 per cent of page requests cannot be identified. Usually 18 or more of the top 20 organizations are ISPs, accounting for another 20 per cent of traffic. Therefore, 40–45 per cent of our traffic cannot be easily identified. Many smaller and medium-sized organizations and companies who run web pages through an ISP rather than their own server could be concealed.

- **"Stickiness" or time spent on a web site during a visit.**
  Some log file software will also track the length of a user visit. The longer a visit is, the "stickier" the site is considered to be. This notion of stickiness is becoming a more popular measure of how useful site content is. However, there are too many external factors that, in

[16] *Techmall*, "New Internet Domain Name NU Now Available for Everyone Worldwide," November 1997, <http://www8.techmall.com/techdocs/TS971110-3.html>.

some measure, invalidate the time recorded in the log file. The determination of the length of a visit can be arbitrarily customized in the log analysis software itself. Because of caching, "stickiness" could be underestimated—the more popular a page, or section of a web site, the more likely it is to be cached elsewhere. More reliable information about the usefulness of all or parts of a web site can be gathered with a user survey.

- **Files that have not been accessed.**
  Since log files only record activity on your site, it is not possible to report directly which (if any) files have not been accessed at all. Depending on the size of the site being analyzed, a list of these files could be extremely tedious to compile manually.

- **Where a user went next from your site.**
  You can't tell where users go when they leave your site. The server is not informed that someone is leaving, nor where they are going.

In summary, web site statistics should be used to indicate general trends and not absolute truth.[17]

[17] Linder, <http://www.zuv.uni-heidelberg.de/webstat/rem2.htm>.

3.0 Detailed Uses of Log File Data

There are more detailed applications of log file analysis to measure progress on achieving the objectives of a web site. In spite of the limitations of site statistics, not applying them to inform future site development activities represents an enormous missed opportunity. But variables should be monitored for internal purposes only, not for public relations applications or formal project evaluations.

3.1 Variables to monitor on a regular basis

We use log file analysis on the SD Gateway to assess the following on a regular basis:

Heavy traffic areas of the site.
Despite the fact that caching can make it difficult to gauge relative popularity of pages, this information gives us an indication of the more popular content on the site.

Table 2: Heavy traffic areas of the SD Gateway, June–October 2000.

Listing directories with at least 10 requests for pages, sorted by the number of requests for pages.

```
pages: %pages: directory
-----: ------: ---------
54103: 36.30%: /jobs/
34883: 23.41%: [root directory]
24376: 16.36%: /topics/
10638:  7.14%: /events/
 9518:  6.39%: /introsd/
 6216:  4.17%: /webring/
 5459:  3.66%: /mailinglists/
 1730:  1.16%: /webworks/
 1357:  0.91%: /livelihoods/
  543:  0.36%: /expertmode/
  165:  0.11%: /ecolegis/
   39:  0.03%: /noframe/
    5:       : [not listed: 4 directories]
```

We have been able to correlate regular promotional efforts for certain sections of the site, such as the Job Bank and the Introduction to Sustainable Development module, with the heavier traffic on those sections. Recently, we've been using this information to redesign the jobs and events databases to connect users from these more popular databases to less frequently accessed areas of the site.

From where site users arrive (e.g., through search engines or links from other sites).
These data give us a sense of which search options drive users to our site, where we are getting exposure and where we might target future marketing activities. As the data below indicate, traffic arriving from AltaVista and Google is significantly greater than traffic from MSN Excite. While this might reflect relative popularity of various search engines, nevertheless we might consider re-indexing the Gateway with Excite.

Table 3: From where users arrive on the SD Gateway, June–October 2000.

Listing the first 30 referring sites by the number of requests, sorted by the number of requests.

reqs	site
4067	http://www.altavista.com/
2515	http://google.yahoo.com/
1752	http://www.google.com/
1218	**http://iisd1.iisd.ca/**
1105	http://192.197.196.2/
1052	**http://iisd.ca/**
973	http://nt1.ids.ac.uk/
874	http://www.idealist.org/
681	http://www.worldbank.org/
517	http://dir.yahoo.com/
447	http://www.cyber-sierra.com/
427	http://www.hri.ca/
330	http://www.info-emploi.ca/
292	http://nrm.massey.ac.nz/
247	http://www.northernlight.com/
241	http://search.yahoo.com/
231	http://google.netscape.com/
229	http://search.excite.com/
222	http://search.msn.com/
222	http://www.workinfonet.ca/
219	**http://www.ecouncil.ac.cr/**
207	**http://www.webring.org/**
194	http://www.lycos.es/
180	http://ink.yahoo.com/
178	http://www.worldwatch.org/
160	http://www.wri.org/
157	http://www.ualberta.ca/
156	http://www.geocities.com/
144	**http://www.enda.sn/**
143	http://informant.dartmouth.edu/

In addition, this information can help identify whether our network members' web sites and the WebRing for the broader group of SD institutions are driving traffic to the SD Gateway. We can tell from the bolded entries in Table 3 that of all the members in the network, significant traffic to the Gateway is coming only from IISD, the Earth Council, ENDA and the WebRing. Our next step would be to verify whether the link to the Gateway on a member's web site has been broken or inadvertently removed. Additional regional marketing and promotion of the Gateway and the member organization would then be discussed with the member.

Factors other than research interests might also be reflected in the data. For example, the World Bank is in the process of building its own meta-gateway; the significant number of accesses from www.worldbank.org might indicate an interest in the structure of the Gateway rather than in the content itself.

What keywords users are entering in search engines that lead them to the site.
We look at keywords in the log files to identify general issues and concepts that are driving traffic to the SD Gateway. The keywords provide an indication of how successful search engines

are at finding our site, and the various ways the site is indexed in the search engines. If some keywords are surprising or missing, we use this information to modify meta tags in the Gateway.

Table 4: Keywords used in search engines, June–October 2000.

Listing the first [49 of] 50 queries by the number of requests, sorted by the number of requests.

```
 289: contabilidad de costos
 208: empleos
 192: organizaciones no gubernamentales
 182: contaminación
 142: sustainability
 128: sd gateway
 126: sustainable development
 113: gateway
 104: desarrollo sostenible
 102: contabilidad
  87: senegal
  70: expo 2000
  68: costos
  59: bosques
  54: grasslands
  53: medio ambiente
  51: danger signs
  46: desiertos
  45: ucrania
  41: listas de correo
  41: medio oriente
  40: développement durable
  39: desechos
  38: full cost accounting
  34: señales de peligro
  34: señales
  33: vietnam
  31: desertification
  31: nueva zelandia
  30: cache:sdgateway.net/noframe/fr_event21999.htm building the information
  30: conflict resolution
  30: causes of deforestation
  29: contaminacion
  29: sustainable development jobs
  29: africa
  28: republica checa
  28: consumismo
  28: biomas
  26: cost accounting
  25: cache:sdgateway.net/noframe/en_event31998.htm water land law workshop
  25: conflictos
  25: praderas
  24: biodiversidad
  24: sustainable development gateway
  24: désertification
  24: australie
  24: resolucion de conflictos
  24: sd jobs
  23: sdgateway
```

At the top of the preceding list are keywords in Spanish. These data provide an important internal indicator that we are achieving our objectives for the SD Gateway: to increase the amount and visibility of southern content and to attract new audiences to sustainable development. The data indicate that there is a demand for well-indexed, highly-credible information on sustainable development in Spanish. Given the greater use of Spanish over French, we might choose to expand Spanish versions of modules and other resources on the site. We will also increase our capacity-building efforts with our network members in Spanish-speaking regions in order to meet the growing demand for content in Spanish.

Also of some interest is the frequency with which users enter the phrases "sd gateway," "sustainable development gateway" and "sdgateway" (bolded in the table above). It may be that use of these phrases reflects growing name recognition for the SD Gateway, in response to our various marketing efforts over three years.

Usage patterns at different times of the year.
Table 5 shows the general trend in page requests for the SD Gateway over a two-year period. Note the consistent drop in requests in December of 1998 and 1999 followed by a strong rebound in January, a function of the holiday season in North America.

Table 5: Page requests logged for the SD Gateway, 1998–2000.

Month	Page Requests	Month	Page Requests	Month	Page Requests
		January 99	37,081	January 00	46,081
February 98	4,851	February 99	36,505		
March 98	5,017	March 99	31,052		
April 98	3,765	April 99	33,789		
May 98	4,022	May 99	26,577		
June 98	7,665	June 99	26,440		
July 98	8,860	July 99	28,187		
August 98	20,452	August 99	30,462		
September 98	22,463	September 99	36,706		
October 98	26,014	October 99	35,179		
November 98	31,282	November 99	36,486		
December 98	23,206	December 99	29,990		

Countries from where users are coming.
This gives us a very conservative and general sense of the geographic reach of our audience. One of the goals of the SDCN and the SD Gateway is to increase the visibility of content from developing and transitional countries. By doing so, we also hope to increase use of the Gateway by other organizations in those regions. Consequently, even though accesses from selected countries in Africa may be at the bottom of the percentile, we consider any access from those countries to be an indicator of success. Access from southern countries is often low compared to U.S., Canadian and European traffic. However, we have noted constant levels of traffic from countries such as Mexico, Argentina and South Africa. We have also noted significant levels of traffic from the South on the Spanish version of the Gateway. Nevertheless, as mentioned in

Section 2.7, because of the limitations of log file analysis software, only general trends can be determined, not users' exact locations.

First page accessed by users.
Log file data have also been useful to highlight that the SD Gateway home page usually accounts for only 5–10 per cent of page requests. Although the home page might really be seen more frequently due to caching, these statistics show that most users will arrive somewhere deeper in the site (i.e., most don't enter via the home page, nor will they navigate to the home page). Recognizing that users usually do not enter by the home page, we have improved our branding[18] and navigation bars across every page in the site. We are continuing to investigate ways to cross-link sections of the SD Gateway to increase traffic from one section of the site to another.

Load on the server and processing times.
This information is useful to gauge whether our server can handle the traffic to the site. Processing times tell us how quickly the server is responding to requests and gives us an idea of how fast the site is being downloaded. It is not only important that we make more content from the South available on the net, we also need to ensure that the site is designed to download quickly for those users with unreliable or expensive connectivity. In the last redesign of the SD Gateway, we set and achieved a target to reduce the user download time by 50 per cent. When using a 28,800 baud modem, the approximate download time for the text of http://sdgateway.net is less than four seconds. Total download time for text and graphics is 21 seconds.

Maintenance requirements: failure reports.
Failure reports in the log file analysis will signal problem areas on the site, such as broken links or non-functioning redirect commands deployed after site redesigns. Failure reports can help you prioritize areas of the site that need attention.

For example, we recently noticed a consistent failure report relating to the SDCN's Sustainability WebRing. The ring was established in 1998 as a major feature on the SD Gateway to link together the broader community of sustainable development institutions. We set it up through the Internet site, WebRing. From April 1999 to March 2000, www.webring.org was consistently among the top referring URLs to the SD Gateway, and never appeared in the failed referrer report of our log file analysis.

In mid-2000, WebRing was purchased by Yahoo!™. From June to October 2000, Yahoo!™ WebRing dropped by 50 per cent in the list of top referring URLs, and accounted for over 20 per cent of the known failed referrer URLs. This information was instrumental in our decision to close the ring to new members until further notice, to inform current members about technical problems with the Yahoo!™ takeover of WebRing, and to work with Yahoo!™ to resolve technical and administrative issues. As of November 2000, Yahoo!™ has corrected basic technical problems, and we have reopened the Sustainability WebRing. We will continue to monitor the Failed Referrer URLS to see whether technical problems are still preventing users from accessing the Sustainability WebRing.

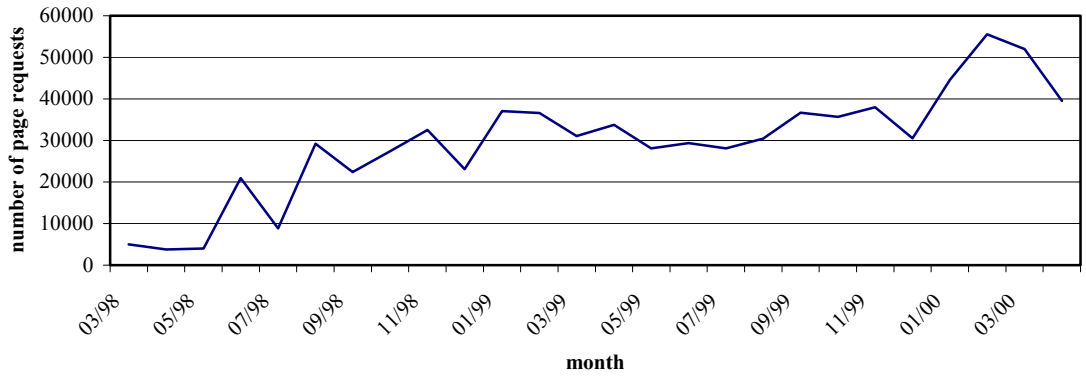
[18] We "brand" the site with the SD Gateway name and logo, the network acronym (SDCN) linked to further information about the network, and the host organization (IISD).

Most popular access times.
Again, this measurement can help predict server load and allow content managers, as well as our system administrator to highlight the best times for conducting site maintenance and updating content.

3.2 Variables to check after major redesigns of a web site

After our first major redesign of the SD Gateway, we applied log file analysis to assess the effectiveness of site changes. Over a four-month period in 1998 (May–August), the number of page requests on the SD Gateway increased five fold.

Table 6: SD Gateway traffic, in page requests, by month, February 1998 – April 2000.



This growth in usage coincided with two announcements to listservs—one about the redesigned site that removed frames in early June 1998; the other about the SD Job Bank at the end of July. The spike in usage in June clearly shows that our marketing efforts had impact. More important, the growth in page requests was sustained after August, indicating that it was not a one-time "flash in the pan" increase.We have inferred from this that the growth in usage was sustained because we produced a site that was easier to navigate, with useful products like the SD Job Bank.

We have also been able to validate our latest redesign of the SD Gateway, which restructured the original SD Primer into a more focused organization of SD Topics. The SD Primer was a directory of topics with links to member content, and in some cases, an article about the subject. In analyzing the data from April 1999 to March 2000, the SD Primer accounted for only 0.4 per cent of the traffic on the SD Gateway. This suggested that the sections were not being accessed as frequently as originally hoped. In drilling down into the data, we discovered that the full articles were very rarely accessed. After considering the resources required to write and keep the articles up to date, we determined that this was not a priority for the future development of the site. The articles were eliminated, the number of categories was reduced and their classification structure was simplified. The name of the product was also changed to SD Topics—a clearer and more easily translatable label. In the months following the redesign work, the SD Topics section jumped to over 16 per cent of the overall traffic on the site.

4.0 Other Assessment Methods

While log file analysis can indicate general trends of site usage, other data collection methods are needed to determine user motivations, preferences, and whether users are successful finding what they want on a site. The following section describes some of these qualitative methods with which we have some degree of experience.

4.1 User testing

In 1998, Nielsen estimated that approximately 90 per cent of commercial web sites are difficult to use.[19] Ease of use is critical for encouraging repeat visitors, especially for sites whose major products are research and information. While considerable time, energy and money might have been invested into your web site, if users find it confusing or can't find the information they are seeking, your efforts will largely be wasted.

User testing is the most effective means to evaluate a user's experience with your site, and will also identify any problems with site functionality. This information can and should be used to ensure that your site will have the attributes that attract repeat users (i.e., ease of use and quality content). Although user testing might seem daunting, it can take many forms and be as formal or informal as you wish. Nielsen[20] noted that useful data can be collected in as little as an hour, and that most major usability problems can be found by testing as few as five users.[21, 22]

While it is important to get a representative audience for testing, don't let this step prevent you from conducting user testing; testing staff at your organization will still provide useful results. You will still end up with usable results if you test your own staff. Moreover, it is not necessary to have a working site before embarking on user testing. Even if a site is still at the early stages of development, it is possible to test the conceptual framework and navigation using paper versions. This way, some problems can be caught early in development before significant amounts of time are spent on creating a functional prototype. Ideally, user testing will be integrated in an continual design process to improve the site through successive stages of development.

We designed our methodology based on the "Discount Usability Engineering" model described by Jakob Nielsen.[23, 24] We asked a group of test users to browse the site for five minutes before being assigned general and specific task-related questions based on site content that required different parts of the site to be used (e.g., calendar of events, using the search functions). Each

---

[19] Jakob Nielsen, "The Web Usage Paradox: Why Do People Use Something This Bad?", *Alertbox*, August 9, 1998, <http://www.useit.com/alertbox/980809.html>.
[20] Jakob Nielsen, "Survey your Users," *Users First*, February 18, 1999, <http://www.zdnet.com/devhead/stories/articles/0,4413,2211547,00.html>.
[21] Jakob Nielsen, "Cost of User Testing a Web Site," *Alertbox*, May 3, 1998, <http://www.useit.com/alertbox/980503.html>.
[22] Jakob Nielsen, "Why You Only Need to Test with 5 Users," *Alertbox*, September 1, 1997, <http://www.useit.com/alertbox/9709a.html>.
[23] Jakob Nielsen, "Guerrilla HCI: Using discount usability engineering to penetrate the intimidation barrier," 1994, <http://www.useit.com/papers/guerrilla_hci.html>.
[24] Jakob Nielsen, "Cheap Usability Tests," *Users First*, March 11, 1999, <http://www.zdnet.com/devhead/stories/articles/0,4413,2224316,00.html>.

user was also asked to "think out loud," verbally noting aspects about the site that they liked, didn't like, or found confusing or clear as they browsed and performed the tasks. Observations were made while they answered the questions. Finally users filled out a questionnaire about their impressions of the site (e.g., if the meaning of icons was clear, if navigation was easy, if there was graphic consistency and logical structure, aspects they liked the best/least). Interestingly, we found that answers provided in the survey about their impressions often conflicted with observed behaviour. For instance someone might have answered that something was easy to find, but they clearly had difficulty during the tasks.

It was humbling to witness this real world use of earlier versions of the SD Gateway, where some users had significant difficulties completing assigned tasks. We learned that something that might make sense to the designer or content developer might not make sense to users. For instance, our first experience with testing the SD Gateway revealed that users did not understand frames and how they functioned, resulting in difficult site navigation. These findings prompted us to redesign the SD Gateway without frames. In another example, we found that some link titles were unclear. This proved to be a critical barrier to successfully retrieving information, since the link in question was (wrongly) deemed irrelevant for the particular search.

Once again, this emphasizes a point made earlier—that while log file analysis can indicate general usage trends and can provide some information about usage patterns, it does not reflect a user's experience with your site. A user might be frustrated with the navigation of a site, constantly moving back and forth, or they might be lost. But because of caching, log files will not indicate these problems. User testing is the most effective means of discovering these kinds of navigational problems.

4.2 User feedback

It is becoming standard practice to have a "contact us" button on top-level web pages, often accompanied by a feedback form that provides users a means of providing comments, complaints and requests for further information on your site. While user testing provides us with a healthy reality check about the quality of design and usability of our site, user feedback forms emphasize the importance of having a customer relations element to our external contacts, an approach not commonly stressed in NGO web communications.

It is essential to provide users with a means to contact you. A feedback mechanism is an easy way to collect users' opinions about your site. It is important to pay attention to feedback received via e-mail and feedback forms on your site. Interested users who take the time to provide comments and suggestions or constructive criticism this way merit a response and consideration of their ideas. In addition to gathering substantive feedback, take note of people's geographic location. One can often gauge this from e-mail address domains or in comments made in the message body. The information can be used to identify if there might be particular issues of concern or interest to a given user group. For example, we have recently received a number of e-mails from organizations in other countries requesting permission to translate and reproduce the Introduction to Sustainable Development in local languages. The level of interest in this module might lead us to seek funds for translation and broader dissemination of several modules from the Gateway. Finally, implement a system for responding so users know their message has been received and that their comments are valued. If you choose to automate this

system, for the sake of public relations and nurturing a community of like-minded organizations, you should also allow for time and resources to supplement this with more personal responses.

On the SD Gateway, we provide a feedback form that includes an area for comment and a databased series of questions that ask users for their favourite features and opinions about the design and speed of the site. Though few in number,[25] we have received some valuable personal comments from these forms during the three years that the SD Gateway has been operational. The forms also provide an interesting contrast to conclusions made from log file analysis data. For instance, while the SD Job Bank is one of the most accessed feature according to the log files, user feedback forms suggest that other features are equally or more popular. Users who took the time to fill out the form indicated that most useful features of the SD Gateway were as follows:

1. the list of news sites (11 per cent of responses),
2. the SDCN member links and the library collection (10 per cent each);
3. the Job Bank (seven per cent of responses);
4. other links and contextual articles (six per cent);
5. electronic mailing lists (five per cent); and
6. calendar of events (two per cent).

4.3 User surveys

User feedback forms tend to be static, remaining unchanged on web sites for months or years, and are available to whomever wishes to provide feedback. User surveys are more focused means of approaching users to gather opinions about your site at a particular point in time and allow a more careful selection of the survey population.

As Nielsen notes, "[s]urveys collect data from real users while they are at your site for a real purpose [and are] brilliant for finding out:

- why people visit your site;
- whether they were satisfied with the site's ability to support them in accomplishing their goal;
- whether people like your site and its visual appearance; and
- what they might like you to do in the future (but remember that users have limited ability to predict what they will in fact want to do in the future)."[26]

Three caveats accompany this endorsement of user surveys. The first is well known to those experienced in social science research methods. Survey participants answering general opinion questions (e.g., did you find the site easy or difficult to use?) might have a selective or faulty memory of their own actions, and might skew their answers to provide information that they deem more useful to the researcher. As noted above, we encountered this tendency in our user

[25] Ninety-eight submissions in three years.
[26] Jakob Nielsen, "Survey your Users," *ZDnet Developer Users First*, February 18, 1999, <http://www.zdnet.com/devhead/stories/articles/0,4413,2211547,00.html>.

tests. This can be avoided by more careful attention to designing and interpreting survey questions. Tips for constructing effective surveys are available in general guides to planning and carrying out social science research.

The second caveat is that while surveys can determine preferences and dislikes, they can't reveal detailed user behaviour. Surveys don't reveal why or where people have difficulties with navigation, merely that they find site navigation difficult to use.

Third, good response rates are critical to ensuring survey validity. One possible way to improve the response rate is to offer incentives to people who complete the survey. One SDCN member that surveyed users provided a small gift to those completing their survey and achieved a 35 per cent response rate. Another way to ensure survey validity is to have a large enough sample size so that even a low response rate will give meaningful results. Another SDCN member successfully implemented an online survey and was able to glean enough important information about who was visiting the site, demographic information (age, gender, etc.), how frequently they visited, their opinions about the site and the utility of the content. The information was used in conjunction with log file analysis to make recommendations about areas of improvement and future evaluation.

Other tips for ensuring effective surveys include the following:

- keep surveys short (ideally to a single screen, whether on a web site or in an e-mail);
- most questions should be closed. Where the users have to supply a single fact, have them go through a checklist or state their opinion on a rating scale;
- only ask a question if the answer will make a difference to your project; and
- test a survey with three to five people before implementing it.

5.0 Using Site Usage Analysis to Assess Marketing Strategies

Information about a site's usage can be applied to plan and assess site design, identify marketing opportunities and to gauge system performance, as discussed in Sections 3 and 4. Web site analysis can also be used to evaluate the impact of marketing and outreach activities. The following section reviews our experiments with the correlation of marketing activities with changes in site usage.

Readers should keep in mind that fluctuations in usage might not only be the result of a short-term promotion, but could be due to the growth of overall Internet traffic. As of June 2000, global growth of the Internet, as measured by the number of servers, is approximately five per cent per month[27] (doubling every 18 months) The success of site improvements and marketing activities should demonstrate growth in site usage beyond this level of "background noise." In addition, success in reaching particular geographic regions must be assessed against the level of overall Internet access in those regions. Even a well designed marketing campaign might not generate 20 per cent of page requests to one's site from Africa since the continent accounts for only one per cent of the estimated total Internet users around the world.[28] Sources for global Internet access and traffic include StatMarket (www.statmarket.com), Nua Internet Survey (www.nue.ie) and The Standard (www.thestandard.com).

Three examples are provided below of the impact of marketing activities on the usage of online products and services.

5.1 Print and e-mail announcement campaign

In the case of IISD's Business and Sustainable Development web site (http://www.iisd.org/business), web site statistics were helpful to compare two different marketing strategies: e-mail press releases and a print advertisement in a widely distributed business journal. From early April 1999 onwards, regular substantive updates to site content were made, and were announced by sending press releases to targeted electronic mailing lists (listservs). In addition to these releases, in July 1999, an advertisement was placed in *Report on Business,* a quarterly magazine that accompanies one of Canada's national newspapers, the *Globe and Mail*. July's site traffic was the highest of the year, indicating that a new user body was reached with the print advertisement. While we cannot infer from this one instance that print ads work better than listservs to reach the business community, it did reinforce for us the need to understand where a target audience looks for information. In this case, using print media was a valuable component of the web site marketing strategy.

5.2 Periodic e-mail updates

On the SD Gateway, the Job Bank is one of the more popular features and is the only feature of the SD Gateway that is regularly promoted via e-mail. While these frequent e-mail updates do influence the traffic to the Job Bank, we have also found that many visitors are referred

[27] International Software Consortium, *Internet Domain Survey January 2000*, <http://www.isc.org/ds/WWW-200001/report.html>.
[28] Nua Internet Surveys, "How Many Online," <http://www.nua.ie/surveys/how_many_online/index.html>.

specifically to the Job Bank section of the site through links from other sites. It is likely that these links were created after other site administrators received a copy of the e-mail Job Bank update. Besides the e-mail updates, other factors contributing to the job bank's success include the type of content (employment, a popular and recurring interest), and the fact that the content is updated more frequently than other areas of the site, encouraging repeat visits.

5.3 Banner ads

In the latter part of 1998, we decided to experiment with buying ads on a major portal site. We chose Yahoo!™ because its popularity and appeal to a wide range of users matched our interests to reach a wider audience for sustainable development information. For a total of US$3,500 we launched an advertising campaign from September 1 to December 15, buying four keywords (développement durable, sostenible, sustainable, sustentable). Each ad consisted of two to five animated graphic images that would appear when any one of the keywords was entered as a search term (see Appendix 4 for a sample). As part of the same campaign, we also arranged for our ads to appear 23,000 times in the category of "Sustainable Development" and all its sub-categories.

Regular monitoring of Yahoo!™ revealed that our ads were not appearing throughout the category and sub-categories of "Sustainable Development" as we had understood. We found that our ads were only appearing consistently in a few sub-categories. We also learned that ads from larger clients got precedence for most of the more popular categories, relegating us to some more obscure sub-categories. This problem was not easily resolved, and required persistent communication and negotiation with Yahoo!™. Because of these problems, Yahoo!™ extended our ad campaign until the end of February and provided additional ad impressions in the "Environment and Nature" category and its sub-categories.

In total, we received over 100,000 ad impressions (including those from keyword searches and appearances in categories), resulting in 1,656 page requests to the SD Gateway. The overall clickthrough rate[29] was 1.6 per cent, though it ranged from 2.7 per cent between September and December to 1.3 per cent between January and February. This is more than triple the industry average.[30] Clickthrough rates were higher when ads were seen as a result of keywords (3.4 per cent) than when they appeared in categories (1.4 per cent).

We found that the category selection had an effect on the ability of our ads to generate traffic to the SD Gateway. About half of the ad-generated traffic to the SD Gateway (807 out of 1,656 clickthroughs) came from the broader category of "Environment and Nature," but ads in the "Sustainable Development" category were more effective (2.2 per cent clickthrough rate vs. 1.2 per cent for "Environment and Nature" category) at bringing users to the SD Gateway. Therefore, had we placed all of our ads in the narrower category of "Sustainable Development," we might have generated more visits to the SD Gateway. The exercise also taught us an

[29] The clickthrough rate is the rate at which ads were clicked on compared to the number of times the ad appeared. For example, a two per cent clickthrough rate means that out of the total number of times an ad was viewed, users clicked on the ad two per cent of the time (or conversely that 98 per cent of the time the ad showed up, users did not click).

[30] The industry average clickthrough rate for banner ads is now less than 0.5 per cent.

important lesson about the effectiveness of keyword choice over category choice since the clickthrough rate reached a high of 4.2 per cent when the ad was seen as a result of the keyword "sustainable."

We learned an important nuance about the way Yahoo!™ structures their search system. We chose the keywords we did because we wanted users searching for combinations of "sustainable" and other topics to be directed to the SD Gateway (e.g., sustainable forestry, sustainable water resource management). But we later found that users who were only using the single search term "sustainable" were being directed to our site, and not users who were using the more descriptive phrases.

Nielsen[31] argues that online advertising does not work because of low clickthrough rates for ads, and even though our experience buying banner ads was successful from an advertising perspective (1.6 per cent clickthrough rate overall), it is tempered by the fact that the campaign resulted in a total of only1,656 clicks on the ad, at a cost of over US$2 per click.

While we know that traffic increased over the term of the campaign from 22,463 page requests in September to 36,505 in February, this volume cannot be directly attributed to the ad campaign. We can speculate that visitors requested additional pages once they came to the site, but we can only be sure that the ads generated 1,656 page requests over the six-month period. We have no way of determining return visits from users originally brought to the site by the ad banner.

[31] Jakob Nielsen," Why Advertising Doesn't Work on the Web," *Alertbox*, September 1, 1997, <http://www.useit.com/altertbox/9709a.html>.

6.0 Conclusions

Over the past three years, the Sustainable Development Communications Network has experimented with a number of tools to evaluate use of its web communications. As a result of this work, we would like to caution site managers against putting an unreasonable amount of credence into web site statistics, and to emphasize that forethought and awareness of how Internet technologies work can help prevent faulty assumptions and inaccurate interpretation of statistics. Nevertheless, site statistics can provide some useful baseline information, although these statistics should be complemented by other information collection methods, including user testing, user surveys and user feedback.

Statistical measures that can provide a general baseline include: page requests, visits and number of products downloaded or subscribed to (PDF files and mailing lists). We have stressed that the nature of Internet technologies means that these measures are inherently *underestimates* of site traffic.

1. **Analyze log files regularly and act on findings.** Run an analysis of server log files monthly, with quarterly and annual totals. Once several months of data are in hand, it is possible to look for signals related to access and navigation, trends and spikes resulting from marketing efforts, and indicators of system performance. The most useful variables to monitor for internal purposes are:

   - page requests (as a rough equivalent to amount of information used);
   - visits (as a rough equivalent to number of users); and
   - PDF file downloads and subscriptions to mail lists with web archives (as an equivalent to interest in specific content).

   More in-depth analysis of log files can help to focus attention on:

   - heavy traffic areas;
   - user access to the site via search engines and other links;
   - additional or different meta tags required for improved retrieval by search engines;
   - maintenance of broken links; and
   - server load and processing times.

2. **Site statistics can be very misleading**. While useful to provide signals and baseline trends, one should not infer too much from site statistics, including statistics captured by cookies if these are used. Site statistics should not be used for corporate promotional purposes. It is simply not possible to claim levels of traffic, geographic and sectoral reach with any definitive degree of accuracy. Moreover, site statistics should never be used to compare one institution's web performance against another's.

3. **Focus on the user**. Remember that engaging and influencing users is your ultimate goal. You can determine their interests, motivations and opinions through log-on forms, feedback forms and user surveys. User testing, though, is the only reliable way to find out how they actually use the site. Only through user testing will you get the information

required to design and maintain a site that will allow users to access content as efficiently as possible.

4. **Combine methods**. A single data collection method will not provide an accurate representation of users' needs and whether those needs were met with the content on the site. Log file analysis has its place when combined with other user assessment methodologies.

5. **Remember the big picture** within your organization and for sustainable development. Web site usage information can be valuable for people with different priorities—for systems administrators, researchers, content managers, designers and marketing managers. Web site analysis will help to improve the communication of your sustainable development knowledge, with improved impact on target audiences, and increased global awareness of actions needed and solutions available.

Appendix 1: Key Definitions

Cache (see also Proxy server)
Caches come in many types, but all work in the same way—by storing information for fast access and retrieval. A web browser cache stores the HTML page's code as well as any graphics and multimedia elements embedded in it, to avoid downloading graphics and text from a server each time you access a page. Since hard disk access is much faster than Internet access, this speeds things up. Hard disk access however is slower than RAM, which is why there is disk caching, which stores information you might need from your hard disk.

Clickstream
"The virtual path or trail a visitor makes while surfing a site or the web itself." [32]

Clickthrough
The clickthrough rate is the rate at which ads were clicked on compared to the number of times the ad appeared. For example, a two per cent click through rate means that out of the total number of times an ad was viewed, users clicked on the ad two per cent of the time (or conversely that 98 per cent of the time the ad showed up, users did not click).

Cookie
A "cookie" is a command sent by a web site to a user's computer to capture and store user information on the user computer. Information captured by the cookie can later be read back by the site from that computer. Cookies are useful for storing passwords and preferences of start pages. Microsoft and Netscape rely on cookies to allow personalized versions of their home pages. Cookies can be used to personalize shopping baskets in an e-commerce setting. They can also be used obtain information about a computer for targeted marketing purposes.

Crawler (see Spider)

Firewall
A firewall is a device that protects a private network from the public. A computer set up to monitor traffic between an Internet site and the Internet. It is designed to increase a server's security[33] and keep unauthorized outsiders from tampering with a computer system.

Hit
A hit is any request sent to a web server, and is recorded as an entry in the log file. A hit is generated each time a web server is asked to provide a page, graphic, or other object to a user's computer. The total number of hits is generally considered a rather meaningless indicator of site traffic since pages that use more graphics will generate more hits. Note that some log analysis software packages use the term "request" in place of "hit."

Host
A host is defined as a unique IP (Internet Protocol) address. Practically speaking, it is a computer that requests information from a server. It might represent an individual, a spider/crawler, a cache, a

[32] Client Help Desk, <http://www.clienthelpdesk.com/dictionary/clickstream.html>.
[33] "Firewall," *NetLingo: The Internet Language Dictionary,* <http://www.netlingo.com/>.

proxy server or an Internet Service Provider. An IP address can identify one user, but more often it is shared by many people. Therefore there might be a higher number of individuals visiting the site than the number of "hosts" counted. We do not consider it interchangeable with the term "user."

IP address (Internet Protocol Address)
An IP Address is a numeric address that is given to servers and computers connected to the Internet. It may represent a unique person, but often many people share one IP address. Therefore there may be a higher number of individuals visiting the site than the number of "users" counted. When you connect to the Internet, your Internet Service Provider (ISP) assigns you an IP address. This IP address may be the same every time you log on (i.e., a static IP) or it can change and be assigned each time you connect based on what's available (i.e., a dynamic IP).

Page request
A page request is defined as any collection of hits that successfully retrieve content (i.e., a single web page viewed). Since the text and graphics are not counted separately, the number of page requests is a more useful piece of information. Note that some log analysis software packages use the term "request" to mean "page request" as defined here.

Proxy server[34]
A server that resides between a client application (e.g., a Web browser) and a real server, intercepting all requests to the real server to see if it can fulfill the requests itself. If not, it forwards the request to the real server. Proxy servers have two main purposes:

- **Improve Performance:** Proxy servers can dramatically improve performance for groups of users. This is because it saves the results of all requests for a certain amount of time. Consider the case where both user X and user Y access the World Wide Web through a proxy server. First user X requests a certain web page, which we'll call Page 1. Sometime later, user Y requests the same page. Instead of forwarding the request to the Web server where Page 1 resides, which can be a time-consuming operation, the proxy server simply returns the Page 1 that it already fetched for user X. Since the proxy server is often on the same network as the user, this is a much faster operation. Real proxy servers support hundreds or thousands of users. The major online services such as Compuserve and America Online, for example, employ an array of proxy servers.
- **Filter Requests:** Proxy servers can also be used to filter requests. For example, a company might use a proxy server to prevent its employees from accessing a specific set of web sites.

Robot (see Spider)

Single page visit
A user leaves the site after viewing only one page.
Spider (or Crawler, Robot, or bot)

[34] "Proxy Server," *Webopedia: Online Computer Dictionary for Internet Terms and Technical Support*, 2000, intenet.com, <http://www.pcwebopedia.com/>.

A spider is an automated program that explores the web, looking for information. The most common kinds of spiders are the ones that index sites for search engines, collect e-mail addresses or check links.

User (also see Host)
A user is any person accessing your web pages. Often log analysis software will define a "user" as a unique IP address, however we use "host" or "visitor" to denote a unique IP address, and do not use this to refer to an individual.

User sessions (See Visit)

Visit
A visit is defined as a series of consecutive page requests from a host to a site (e.g., in one "visit" a host may view several different pages of a site).

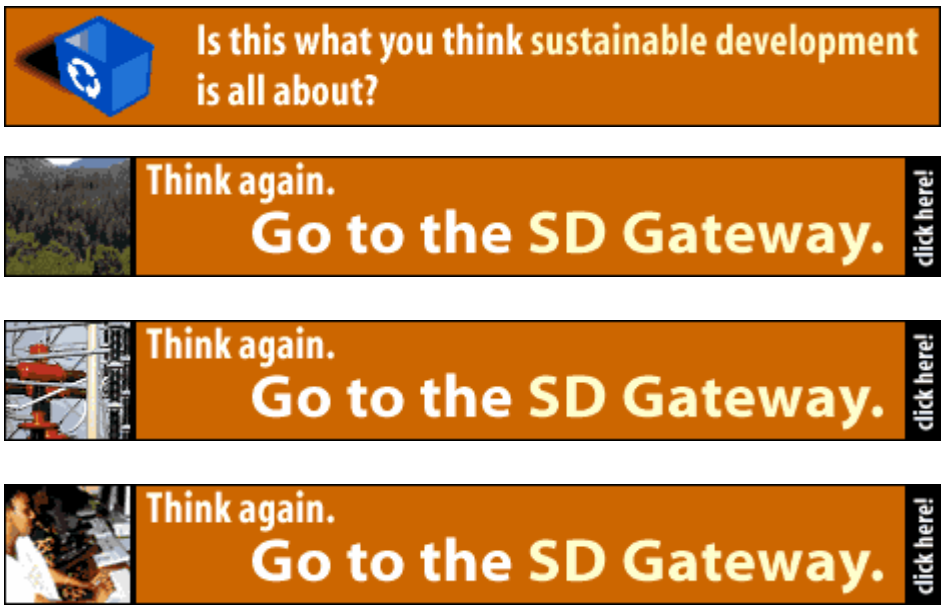
*Visitor (*See Host).

Appendix 2: Web Site Analysis Services and Software

COMPANY	PRODUCT
The Counter (http://www.thecounter.com)	Provides daily free in-depth traffic reports. Tracks number of visitors, referrers, browser popularity, operating system. Requires a few codes of HTML code be added to each page.
ExtremeTracking (http://www.extreme-dm.com/tracking/)	Offers web site traffic analysis service, requiring users to put a small logo on the pages being analyzed.
Hit-o-meter (http://hitometer.netscape.com/)	Offers daily, weekly and monthly hits, e-mail reports, visitor browser/platform, screen resolution, java script usage, referral information, search engine utilization and visitor loyalty/frequency. The paid service offers more detailed analysis of visitors, including type of browser, country from which the visitor comes from, who links to your page and how many hits are generated from those links. Free service provides limited variables. For cost, service is complete.
NedStat (http://usa.nedstat.net/)	Three levels of software, (Basic, for private users, Pro, for professional web sites, and Site Stat for large, busy web sites).Basic is a free service for non-commercial homepages and web sites. NedStat Pro is not free (US$495 per year for analysis of 20 pages), but is available for a freefour week trial. Software is available in seven languages.
Hitbox (http://www.hitbox.com)	The user installs a banner or button on their web page (the button is invisible with the paid service). Each time a visitor accesses that page, the image is served from the WebSideStory (http://www.websidestory.com) servers. This method allows for the collection of hundreds of statistics in real time, 24 hours a day.
Accrue Software (http://www.accrue.com)	Accrue Insight and Accrue Hit List analyze clickstreams to help refine content, sales efforts, pricing and merchandise placement.
WebTrends (http://www.webtrends.com)	Offers downloadable software and “real time” live site traffic analysis. For this last option, a personal 14-day trial is free, requiring the placement of a small button on the tracked pages. E-commerce version tracks total page views, number of daily unique visitors, total visits, first-time visitors, returning visitors, average page views per visitors and more. Executive and e-business versions must be bought.
Web site Traffic Report (http://www.web sitetrafficreport.com/)	Offers web site traffic reports, supported by advertising revenues from ads placed in the traffic reports.

Appendix 3: Sample banner ad for Yahoo™ campaign

- http://us.yimg.com/a/in/intl_sustainable/newabout.gif



The other ad banners can be viewed at:

- http://us.yimg.com/a/in/intl_sustainable/newad.gif
- http://us.yimg.com/a/in/intl_sustainable/newlanguage.gif
- http://us.yimg.com/a/in/intl_sustainable/energyeff2.gif
- http://us.yimg.com/a/in/intl_sustainable/newestloop.gif