CENDI -- 1998-1

CENDI

WEB METRICS AMONG THE CENDI AGENCIES: Collection and Analysis

Submitted by
The Web Metrics Task Group
CENDI User Education Working Group

Prepared by
Gail Hodge
Information International Associates, Inc.
Oak Ridge, Tennessee

April 1998

CENDI WEB METRICS TASK GROUP

Cheryl Hunter (DTIC), Chair
Bruce Ansley (NASA/CASI)
Barbara Bauldock (DOE OSTI)
Tammy Borkowski (DTIC)
Robert Bunge (NTIS)
Simon Chung (NASA STI Program)
Dave Duran (USGS/BRD)
John Howard (NASA/CASI)
Paul Schnake (NLM)

Gail Hodge (CENDI Secretariat)

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Commerce, Energy, Defense, Health and Human Services, Interior, and the National Aeronautics and Space Administration (NASA).

CENDI's mission is to help improve the productivity of Federal science- and technology-based programs through the development and management of effective scientific and technical information support systems. In fulfilling its mission, CENDI member agencies play an important role in helping to strengthen U.S. competitiveness and address science- and technology-based national priorities.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In August 1997, the CENDI members established the Web Metrics Task Group to review the types of metrics being collected by the CENDI agencies, to review the analysis and uses to which these metrics are put, and to share lessons learned.

The task group met on November 17, 1997, at the DOE Forrestal Building, Washington DC, moderated by chair person Cheryl Hunter (DTIC). The team was comprised of representatives from the National Aeronautics and Space Administration (NASA), the National Library of Medicine (NLM), the US Geological Survey/Biological Resources Division (USGS/BRD), the National Technical Information Service (NTIS), and the Defense Technical Information Center (DTIC).

The group found that there was little overlap in the actual web utilities used to collect and analyze the data. The most commonly used product was Access Watch. Much of the software was customized shareware/freeware or internally developed scripts and programs. Many of the problems expressed in relation to web statistics and analysis is because of the inherent nature of the Internet — the inability to resolve IP addresses and the volume of log information. Most Webmasters felt that their customers were pleased with the type of statistics provided. In many cases, the concern is for the presentation and aesthetics of the graphics, not for the raw data.

The web statistics are used for a variety of purposes—to allocate resources, for configuration management, to provide customers with "marketing" information regarding the use of their web sites, for understanding how sites are used, and making changes in the way the sites are presented or navigated.

Archiving is a concern because of the implications of Freedom of Information Act (FOIA) and because of the storage that will be required. Each agency has a slightly different policy on the archiving of the log data. They also differ in the degree to which the information is provided to internal customers, external customers, and the public.

Privacy and "cookies" are other issues. "Cookies" can be attached to the transactions of an individual user and can provide additional information about the user while using the site. Though the use of "cookies" is a technique for providing more customized products and interactive government/public communications, agencies are hesitant to use them because of the privacy implications.

Recommendations:

- Host a joint meeting with one or more members of the Federal Web Consortium team to review the current guidelines

- Develop a white paper related to the issues of privacy and the use of "cookies" and other mechanisms to enhance the ability of federal agencies to provide interactive services to the public

- Develop a registry of Web utilities (particularly those developed by the agencies and those customized shareware/freeware products) that would promote re-use rather than redevelopment of tools

- Host a CENDI-only Webmasters listserv to promote networking among the group

1.0 INTRODUCTION

In August 1997, the CENDI members established the Web Metrics Task Group to review the types of metrics being collected by the CENDI agencies, to review the analysis and uses to which these metrics are put, and to share lessons learned.

The task group met on November 17, 1997, at the DOE Forrestal Building, Washington DC, moderated by chair person Cheryl Hunter (DTIC). The team was comprised of representatives from the National Aeronautics and Space Administration (NASA), the National Library of Medicine (NLM), the US Geological Survey/Biological Resources Division (USGS/BRD), the National Technical Information Service (NTIS), and the Defense Technical Information Center (DTIC).

Prior to the meeting, the CENDI Secretariat and the User Education Working Group Chair prepared a list of discussion question as an outline for the presentation from each agency representative. The questions are provided as Appendix A.

2.0 AGENCY DESCRIPTIONS

2.1 Defense Technical Information Center (DTIC)

DTIC's web environment is fairly complicated with over 100 DoD public sites housed on 12 SUN servers running Solaris OS and NCSA Netscape Enterprise web servers. DTIC hosts several restricted sites, including *Early Bird*, that is restricted to .mil and .gov accesses. STINET, DTIC's Scientific and Technical Information Network, has both public and secure sites running SSL technology. Internally, there is a mirror development environment which is IP restricted. External customers can FTP their site changes to DTIC. The revised sites are then moved to the production environment by DTIC staff, either manually or by automated scripts. DTIC web-based Intranet is available only to the DTIC subnet and certain external customers based on IP.

DTIC uses Access Watch software developed by David Maier, a local consultant, to gather web statistics. (Access Watch is free for academic and government use.) Glue scripts are used to customize Access Watch. The statistical package is run once per week for each root level directory. Graphical displays of the results are available on the Web, but restricted by .mil and IP addresses for their external customers.

Statistics presented include hourly server load, accesses by the top pages within a given directory, and accesses by domain and host. Accesses by domain provides information at the country domain and domain name levels. The summary data breaks down the accesses by internal and external users, allowing project developers to distinguish the accesses within DTIC for development and monitoring purposes from the outside user accesses.

The number of accesses are counted based on a server request for an HTML page. A hit is any server request. Therefore, the opening of ".gif" files are counted as additional hits, but only one access for the page. Hourly statistics (maximum and minimum) can be gathered. Hourly statistics are also presented by percentile. The number of days that fall into the 90th percentile are

displayed in color in the graphical display. Accesses per day are determined by the average hourly accesses times 24 hours.

Information on page demand can also be presented. The demand can be configured as low as ten. When configured at ten, Access Watch gives you the top ten hits for that data root level. If the site has a very sophisticated structure, the number would likely be configured higher. The page demand can be configured for any subdirectory level, not just the root.

DTIC web customers have been very pleased with the information provided by Access Watch. According to the content developers, the statistics are used to: 1) identify the site audience in order to focus the content, 2) determine the target audience for the site content in order to hone the marketing strategies, and 3) justify the budget (perhaps less from an economic point of view than to justify staffing for Web-based information dissemination). The statistics are helpful to determine the success of a site and its contents. It is especially interesting to look for historical trends.

DTIC does not use the statistics to determine capacity planning either for server or network bandwidth. For DTIC's own site, they use the statistics to hone the content.

Since DTIC hosts so many sites, web metrics help to answer the question of what are the high profile sites and where additional manpower may be needed. When politics and the number of hits are escalating, sites such as Gulf Link will make changes quickly in order to be most responsive.

Raw statistical data is compressed and archived for two months on the server. DTIC uses the Legato backup system clones for disaster recovery. Eight weeks of statistical data is therefore easily available without having to retrieve the clones from offsite.

The graphical output is kept for the life of the project with a weekly roll up. Some combined trend information would be helpful.

DTIC had previous experience with Web Usage software that did not handle large log files. Some combination of the size of the log files and the number of directories caused the software to run unsuccessfully.

What statistics are important and to whom? Everyone wants their own kinds of stats. *Technology Navigator* wanted statistics by e-mail address. An online survey was the answer but it isn't possible to satisfy all the needs this way.

Access to the DTIC statistics are limited to DTIC's external customers and the ".mil" domain. Statistics are restricted from public access at this time.

The issue of the FOIA as applied to web statistics came to a head within the last year. The request was for the last year of the web logs. The request went to the head Administrator who ruled that web logs were excepted from FOIA, because of privacy issues. However, FOIA will continue to be an issue. A DoD-wide policy has been discussed. FOIA responsibilities also effect the length of time that the logs can be archived. Archival time is also a server space issue.

Future plans include the addition of referrer information to the output; i.e., whether the user is coming from another site or directly to the DTIC-managed site. This information will improve site marketing. The statistics will also be presented in tabular format. Version 2.0 of the Access Watch software is currently in beta testing. However, DTIC is planning to customize it internally. Information about the current browser will help the site developer decide whether the text-only site must be maintained as well as the Java version.

Users want an improved graphical display. It is currently just a tally of the raw data. Pie charts or bar charts would be more visually comprehensible.

DTIC would also like to add trend analysis to the package so that monthly and yearly roll ups could be supported. Weekly linked statistics are now used to do trend analysis. (The linked weeks are all available online at this point. DTIC will eventually have to make a decision about how long to retain these linked sets, but the HTML files are relatively small.) The hope is to have this capability within Access Watch by the beginning of the year, rather than exporting the data to another package by parsing the HTML files to fill a spreadsheet.

Questions.

Have you ever had to use Legato to restore any files? DTIC has used it to restore access logs. There was an early problem with Legato due to the UNIX file system limit on files over two gigabytes. Legato was producing index files over this limit and bringing down the server. Solaris 2.6 should fix this problem.

When changing from NCSA to Netscape, did you have to do any changes to the log files? There is no problem as long as the common log file format is used.

Why did DTIC move from NCSA to Netscape? DTIC was using Commerce Netscape already, so they wanted to use the Netscape solution rather than another vendor. DTIC has had excellent support from Netscape. There was a problem getting GulfLINK set up because of the traffic, and Netscape mirrored the site on the West Coast in order to solve the problem.

What kind of volume has the DTIC site seen? There have been hundreds of thousands of accesses for the top projects. It is hard to determine the total number because, until recently, DTIC had all the big projects on the main production server. They have now moved them off and done an internal segmentation to accommodate the large traffic loads. They do not have total statistics across these two environments.

2.2 National Aeronautics and Space Administration Center for AeroSpace Information (NASA CASI)

Three sets of statistics are currently gathered—WAIS database access logs, Web access logs, and CASI technical report server (a WAIS interface to RECONplus, a BASISplus database) logs. In addition, there are statistics from the access logs to BASISplus (RECONplus) using the AMIABLE character-based interface.

Cron jobs process the statistics on a daily and monthly basis. Two freeware packages are used wwwstat 2.01 and a companion piece called wwwsplit, which is used for the monthly process. On the eighth day of the month, wwwsplit splits the file and takes the preceding information as the monthly level. This is prone to problems.

Gwstat takes the wwwstat HTML output and creates bar graphs from it. Currently, only internal CASI staff and certain outside IP addresses have access to the access.comf file.

Trend analysis is desired. The results of wwwstat are manually entered into a 1-2-3 spreadsheet. Bar graphs are provided to the users.

RECONplus runs on an IBM RISC 6000. The GILS, WAIS Server and WWW server are all housed on the Sun SPARC 10 server running SOLARIS 2.4. Under WWW server there are two special requests. The Publications Group wanted to know how many times the electronic publications were being downloaded.

RECONplus usage summary breaks out minimum, maximum and average search times for daily, hourly and monthly. The usage summary also breaks down the number of searches performed, the users who perform them, and the number of sessions for each. There is also a breakdown by database. Usage can be identified by NASA Center and individual user ID. The users must be registered, so there is a name and password retained.

In the case of the CASI TRS and GILS systems where there is no user registration, the usage is tracked by the IP name. With the AMIABLE interface, the response time for certain search functions is also monitored.

Gwstat is used to produce graphical bar charts. The technical staff received a special request from the NASA monitor, who wanted to see the high level domain grouped by country and then in order by the most frequent user within each country. To do this, it was necessary to copy the section of the statistics to a text file, import them to Dbase, and then perform a report on that data which was resorted as required.

One of the main ways that users use the statistics is to compare mainframe access with the client/ server access to the database. Some statistics such as the publications request are supplied to NASA by the contractor as part of monthly reports.

Raw data is tarred and removed after a certain length of time. The resulting HTML files are backed up as normal. Only the past year is linked on the online view. The online view is an image map so that you get the full size version if you click on it. The directory level is archived as

the past twenty-four hour period, the last week, the last 30 days, and the last 12 months.

Daily usage statistics are produced for RECONplus, where average search time and the number of searches are tracked.

The GILS connections do not distinguish searches from regular accesses, but it looks like about 90 percent of them are the result of a search.

The freeware Web Usage package didn't cost a lot of money up front, but it has a maintenance cost and liability. Many customers are interested in where people are coming from, but NASA CASI has problems with IP addresses that can't be resolved. There are a fair percentage coming from foreign sites that cannot be resolved. This impacts the ability to provide accurate, meaningful statistics.

NASA decisions regarding archiving have been influenced by NARA and FOIA requests. A NASA CIO Executive Notice at LARC stated that access logs will not be kept beyond 30 days. No backup or historical archive is made.

Another concern is server traffic. With multiple services on one server (NetServ has the WAIS interface, GILS and Listserv and domain name server), there are times of the day when it is really hit hard. This is a concern not so much for keeping up the statistics, but being able to use the statistics better to document this effect.

When asked what the users would like, they have indicated that they would like to know how many results were retrieved from each search. This is currently not tracked. Was it a successful search? This raises the question of how to determine the quality of what the user receives.

NASA plans to survey commercial-off-the-shelf solutions that might present the data more graphically and in a more automated, real time mode.

NASA CASI is using the commercial WAIS server for GILS. However, this has not been supported since October. Blue Angel is now being promoted instead. (DTIC is looking at the beta version.)

2.3 National Library of Medicine (NLM)

Analog, developed by Stephen Turner from the Statistical Laboratory at Cambridge, is being used after some slight modifications. The product was chosen because it is written in C not in PERL. He also found it to be faster than wwwstat.

At NLM the same log file is used for the Intranet and Internet. In hindsight this makes parsing more difficult. Analog assumes that the access log is one file and so it parses the whole file even if you only want part of it. Because it was important to be able to do only part of the file, this modification was made to the original program.

HTML output is produced on Requests per page. It really means a page. It doesn't count a hit as a graphic or a bit map. However, the software allows the user to defined a page as necessary,

including a CGI script or an HTML page. All hits or external only.

Analog is used for ad hoc reporting. Reports are generated on the fly. There is an HTML form so that the users can request the reports themselves, freeing up the time of the Web staff.

NLM also uses http-analyze from Rent-a-Guru. Customizable reports can be created but they are not as good as Analog's standard reports. However, the graphics are "prettier". This approach is good for the historical trends. The trend on the chart shows the increase in usage based on Free MEDLINE.

The total number of 304s is also provided. The number that have resulted since the last time the content was changed is analyzed.

It is possible to see the peaks and valleys using Analog and http-analyze. The maximum, average accesses, top hosts, top domains, etc. are collected. The least active URLs on the site can be reported. These normally turn out to be typos in the name of the URL or lack of active links. The tool has a host of reports, referrer logs, browser logs, etc.

The web statistics are used for capacity planning. Is the hardware sufficient? Do we have proper CPU and bandwidth necessary? Since March 1996, the server has been upgraded at least twice. SUN Ultra 1 is now used. The machine not only supports the Web but public ftp.

Statistics are also used to analyze the organization of the web site. What is being accessed and what isn't? This includes Free MEDLINE and Visible Human under the Hot Topics. What is not hot? Should there really be a link on the front page or is indirect navigation adequate? When NLM reorganized the site, long directory lines were shortened. The URLs were changed, and in doing so, old logs were reviewed for frequent referrers to those URLs. NLM notified some of the top referrers about these changes so that broken links would not result from the URL changes. Domain information is not as helpful, except to determine the foreign audience and Intranet vs. Internet accesses.

NLM is also interested in how often searching of the web site is performed. If the users search this may mean that they could not find what they were looking for directly from the homepage. This leads to thoughts of site reorganization.

NIH has published web guidelines to which NLM adheres. Any external page must work with 85 percent of the browser market back to one year. Internal pages must work with browsers back to 6 months. The Webmaster keeps track of this to try to gear the site accordingly.

Web statistics are used for justification and management decisions. For example, they meter the access to the NLM Fact Sheets in electronic form. The regional medical libraries can now print them directly from the web site. The same is true with the NLM Technical Bulletin from January 1998 on. NLM's annual report includes the web statistics. Resource allocation questions are being supported by these statistics. Log files are processed daily. Resolve IP addresses to names, compress vis gzip and move to a new partition for statistics only. All log files back to March 1996 are available via NLM's intranet; the statistics are not made public. There are no logs prior to this time. The log files are retained online so that the users can do customizable reporting,

freeing up staff for other things. This may have to be questioned as the disk space gets used up.

Privacy is an issue. Host names and IP addresses are collected. This could be a problem. There is the possibility of using cookies to track users. This would be helpful in understanding the audience, but it raises privacy issues.

Storage is also an issue. The log files grow at 3.5 megabytes per day, compressed. The statistics could be archived or they could allocate more disk space. The desire to run reports on an ad hoc basis requires that the raw data be maintained. It is possible that a process could be developed to restore archived data on demand.

There are ongoing questions about what we really get from the web statistics. What does a hit really tell you? There is the need for more information in the statistics which might be aided by cookies. Cookies could provide the number of pages accessed per visit. It would be possible to find out what the user did and how they got to the site. This could result in reorganization of the site to make access easier.

The shareware versus commercial web statistics packages are an issue. The benefit to shareware is the ability to modify.

NLM discussed its use of Apache. Apache has a Usernet newsgroup and support is very good. NLM had very poor support from Netscape.

NLM has also been looking at Microsoft Usage Analyst from Microsoft's BackOffice suite of products. Trend analysis can be performed via cookies. Generally everything is still in UNIX, but as systems switch to NT there will be more issues related to the switch.

NLM indicated that there are people who misinterpret the search capabilities of the Web site, thinking that it is searching MEDLINE. Bruce mentioned that there may be a chicken and egg on the searching, since some people are browsers and others are searchers, and the current environment does not allow developers to distinguish well.

Free MEDLINE has moved from the fourth most accessed page to dwarf all others. The Webmaster's e-mail is increasing because of the public availability. The e-mail is increasingly forwarded to the library reference section. NLM has a contact us page which clearly delineates the reference from the general mail, but this doesn't seem to help. NLM is working to try to get the e-mail forwarded with some intelligent software.

2.4 National Technical Information Service (NTIS)

FedWorld operates Internet services for NTIS and other agencies. It includes about 20 different Web servers, two WAIS servers with about 30 different databases, and an ftp server with about 10,000 files. Other services include bulletin board services. The services are varied including normal web site like the US Customs Service and the IRS sites, and specialized servers such as databases and specialized applications such as the US Coast Guards distributed bibliographic collection system, and the Defense Acquisitions University's distance learning system.

Shopping Malls include the selling of information products. This type of service has led NTIS to use cookies and SSL support for credit cards. Subscription fee based services include the Bureau of Export Regulations and the World News Connection. Some of these services are IP controlled while others are password protected.

NTIS uses cookies to provide state in a stateless environment. Cookies are used to track the order process not for statistical analysis. The cookies content goes away when the order has been processed.

The degree and type of web statistics requested by the customers varies. Some agencies never ask about web statistics, and others want more. Most customers who want statistics are budget driven.

Most of the sites are in the open public environment, but there are some internal Intranet and also intranet virtually controlled by IP. There is a serious focus on systems security, because NTIS take credit card and also because there are customers concerned about the integrity of their data. The level of security effects how the logs are handled. In some cases the security folks use the logs to determine hacking attempts. The logs can also include some information which have raise FOIA and privacy issues.

UNIX servers and some NT systems are used. There are eight to nine systems on NT at this point. Most of the public sites are still on UNIX systems with Apache Web servers. NTIS has been very satisfied with this. Netscape is used for security. The IRS site runs Netscape Enterprise's Internet Information Server.

NCSA's common log format is used. The IRS site also uses the default format as well, which of course is different from the NCSA format. This format allows the log entries to be sent directly to an Oracle or SQL database with the Internet Information Server.

Two tools are used: Statbot is the default tool. It costs about $35. However, the downside is that it has not been updated in over a year. However, the version does what they need, and there are no real problems. NTIS has gained enough experience with it in the last two years, so there is little need for vendor support. The product is compiled C code, so NTIS is unable to change it. However, Statbot produces graphics that are particularly appealing to management.

Statbot collects the number of HTML page hits by hour and day. For daily traffic comparison, NTIS has the software configured to track up to the last 6 weeks, graphed by the number of hits by the number of days. The graphic output has a different line for each week. This presentation seems to be the most popular from the users standpoint.

Statbot also tracks the top 10 most active sites. However, many of the accesses are repeat traffic perhaps from a crawler and from AOL. Traffic can be broken down by domain. The data output in bytes also reported.

Statbot keeps an internal binary database. It could almost be run in real-time. It is run either every half hour or every hour. The output is also updated that often. As the site receives more

traffic, Statbot should be run more often rather than less often. The IRS site, which has a million hits, runs Statbot every 10 minutes. During tax season, it is run every minute.

There are many features that can be configured. NTIS typically reports the previous day's total back to 90 days. They also tracks domain information. Statbot can be tricked into counting specific pages to track to see how many times they come into the FedWorld site and search. It can also determine how often the CGI tools are being used.

Statbot can also be configured to show the IP hits. NTIS does not typically configure Statbot to track individual IP addresses by machine.

Access Watch is also used. This is used when customers want to have a count by individual page and by directory. The primary difference is that it is run every night. NTIS's use of Access watch is similar to that discussed by DTIC. Separate instances of Access Watch are run based on projects. Both Statbot and Access Watch can be configured to look at multiple logs that would combine everything. This would be very difficult and would take network drive space.

For its NT servers, NTIS has gone to the WebTrends product which runs natively in the NT environment. It is a very complicated product, because it has many features. You can tell it how long to consider the length of a session. It will produce different graphics. It is very configurable. IIS Server logs lots of information out of the box and this fits well. WebTrends can supposedly be used with UNIX log files by ftp'ing, but this doesn't seem to work that well. The disconnect is that the log files are very large and this doesn't work well with the ftp'ing process. The other issue is that WebTrends outputs in HTML and gif images. This raises interesting issues the results are to be stored on a UNIX system. The cost is about $160. The documentation is adequate.

NTIS also has developed some homegrown software for specific systems. The shopping mall applications have software that scans the log files for specific CGI calls and pulls some information to determine how the systems are being used.

WUFTP from Washington University in St. Louis is used for the ftp server. It reports on the number of downloads by directory by day. A homegrown script takes the WU log and looks for individual outputs to a text file that has path/filename and number of downloads. Cornshell scripts are used to parse the data.

The WAIS logging feature is generally "turned off", because the log files get very large and they are not analyzed. Sometimes NTIS turns on the WAIS logging for a day or two to get a sense of how the databases are being accessed. WAIS Reporter runs slow and it produces a far amount of information that NTIS does not use. NTIS has written a few homegrown scripts for WAIS in order to pull out the search terms that are being used. NTIS is hoping that some of the CENDI representatives might have some WAIS-related scripts that they are willing to share.

By default, almost all tools are configured conservatively. Page impressions (HTML pages) are counted rather than hits.

NTIS has found that they cannot get the same statistics for Statbot and Access Watch. This

causes problems when both packages are used for the same site

NTIS supports a broad range of customers, so the usage of the metrics varies widely. Some user may be doing daily reports and others might never be reading them. Internally NTIS uses the statistics for load capacity planning. The web log is an additional component that give additional statistics for long term trends and a sense of where performance issues and problems may be coming from.

Currently, most logs are rotated daily or monthly. After a few months the logs are archived to tape. The archival requirements are set by the customer. Some of the logs are sent to the customer. There is no overall log archiving policy. The policies might differ, but this has not come up as a real problem.

Only in rare cases does NTIS have links back to previous data. The metrics pages are overwritten by Statbot, and NTIS has configured Access Watch to overwrite itself. Some customers save the HTML to their local machines, and they may be doing this linking.

Statbot or Access stops running sometimes because of database or network problems. Log file size is an issue at NTIS. The IRS logs get very large very quickly. The log is sent to a network file system and one machine is dedicated to moving the logs around. Moving the logs to tape requires a lot of space and manpower. In some cases, NTIS would like to run Access Watch more often than once per day, but there is always an issue of system load and computing power to do this.

For the future, more referral and browser logs would be helpful. NTIS would be interested in cross tracking the error logs and the usage tools. However, they have not seen many commercial tools for this purpose. The McClure team produced PERL scripts that did some of this for GILS. This would be CPU and IO intensive so would probably require a C program.

NTIS would like to know more about how to improve the search interfaces, navigation and help information. NTIS would like to spend more time observing user behavior. The Web logs could be fed to a product such as Astrosite to see how a user has navigated the site. However, this also has privacy implications.

2.5 USGS/Biological Resources Division (USGS/BRD)

Throughout the USGS, there are about 300 registered web servers and perhaps as many as 400 (or more) unregistered web servers. USGS/BRD consists of approximately 50+ web servers. The BRD/Center for Biological Informatics' two web servers house approximately 25+ sites/projects including Center homepages, the NBII (National Biological Information Infrastructure), and the BRD homepage. The BRD consists of virtually every server type and platform (NCSA, Apache, WebSite, IIS, Netscape, CERN, MicroSoft's Personal WebServer, etc., on UNIX, PC, and Macintoshes) and maintains casual oversight of the 50+ sites.

The physical server with which the BRD representative is most familiar (biology.usgs.gov) is a SUN Ultra running Solaris 2.6 and Apache 1.2.4. Until recently, the system was a Sun Sparc5 with the NCSA server and used common NCSA log files. Weekly, a CRON job executes that runs

Webusage 2.3. The BRD recently purchased an upgrade to Version 5.0, significantly more configurable and produces color graphics.

In addition, BRD uses several UNIX and PERL scripts: BrowserCount, WebStats, RefStats, etc. that post-process the log files. The staff also manually perform reviews of the error logs. In conjunction, BRD uses an Excite Search on a monthly basis that searches the Internet and mirrors each of the BRD sites (using PERL script WebCopy).

One product of WebCopy is a log file for all the interactions it has with the remote site, and is used to trace 404 errors. BRD periodically runs Webxref, which identifies internal and external dead links. BRD has modified the original script to ignore the internal links and identify only at the external links. To augment this, BRD has written home-grown UNIX shell scripts to evaluate the log files looking for activities. One script used to respond to Webmaster e-mail, looks at the last access log and grips the "click-path" from an individual visitor (visitor = IP address). The output includes the hostnam/IP, date stamp, and URL. Proxy servers such as AOL cause confusion, but in all this additional information is very helpful. This information helps the Webmaster to be better informed when answering the e-mail. This could also be of help to in reference helpdesk. An enhancement would be to have the script do an additional step which would locate the owner of the page and forward the message to that owner instead of requiring the Webmaster to do this manually.

BRD has reviewed other relevant software, but settled on this combination about two years ago as it meets present requirements and needs. BRD has minimized upgrades for hardware and software. Other USGS divisions and BRD Centers use LogProfile originally written by the HQ Operations group and upgraded by the Water Resources Division. Others are using http analyzer.

Early in the use of web statistics, BRD used programs that looked at each file and compared it to the referrer log. Information about how many times a particular URL came to the page was output for public perusal. However, there were two problems -- FOIA and privacy. BRD took the statistics offline because the Webmaster was able to use AltaVista and find any page that had ever been accessed from the server. The AltaVista report included information about BRD pages which were obsolete, because it was extracting information from the log file statistics pages. Primarily due to this the statistics files are now restricted access.

The best reference book on the topic of web statistics and analysis is Rick Stout's *Web Site Stats (Tracking Hits and Analyzing Traffic)* Osborne/McGraw Hill ISBN 0-07-882236-X which gives an overview of the major Web packages, what they can and can't do, and how they present their data.

BRD would like to utilize knowledge from Marketing and Statistician groups regarding collection, analysis, and interpretation of web log files.

3.0 GENERAL DISCUSSIONS

Cookies were discussed at some length. Cookies are used by few of the agencies because of privacy concerns. Cookies can be annoying from the users standpoint. If you don't accept a

cookie, you often don't see the whole site. Also, if you don't accept the cookie, it will often come back several times, asking if the user wants to accept the cookie or not. Online surveys also have some of this problem, and several agencies reported finding important information about their target audience from online surveys or guest books. However, there are instances where cookies would be helpful. A recent EDUCOM listserv article indicated that there were agencies that had been using cookies that had removed them because of privacy concerns. There was concern among the participants that a broad-based denouncement of the use of cookies would also create a problem when cookie technologies become more integrated with regular log analysis products.

The participants discussed web policies within their agencies. All have guidelines of some sort. Some are more detailed than others. Some address the issues of privacy and others do not. It does not appear that cookies are specifically addressed. The group is interested in better understanding the various web Guidelines, particularly those of the Federal WWW Consortium.

Education of users was discussed. Many of the representatives indicated that they had spent time educating their management as to the meaning of the statistics that are being gathered. The agencies hosting web sites of others doubted that the content developers of those sites were really aware of the types of statistics that they could have.

The group also briefly discussed other tools used with their sites. NLM and NTIS test their sites with Lynx. This is primitive enough that it avoids the need for special utilities. Americans with Disabilities Act compliance is checked by using Lynx for text and then running it through a speech generation program.

Validation software is also used. At NLM, HTML 3.2 is used. WebTechs.com provides an HTML validation service where snippets of code or a whole page can be submitted for validation and error reporting online.

The issue of the rating of web sites for PICS was discussed. OSTP sees the need to extend this to handle the government sites. Problems occur with sites like the Visible Human and History of Medicine where the content could be interpreted as equivalent to that of pornographic sites.

Several agencies noted problems with GSA restrictions on domain names. NTIS wanted trade.gov and BRD requested nbii.gov to get away from agency specific domain addresses. While GSA has approved both, it took a long time to get approval.

4.0 COMMON ISSUES AND CONCERNS

- There are numerous packages and combinations of packages in use depending on the agency's needs, the platform and the search engine. Most of the packages are shareware or freeware. Customization is important. The most common package is User Access Watch.

- The problems encountered when collecting statistics and identifying metrics are more a result of problems with the log files and the Internet in general, e.g., the inability to resolve domain names and the random addresses arriving from networks such as AOL, rather than the packages that are used.

- There are a number of scripts that have been written by the agency staff to supplement the other packages and to overcome some of the problems noted. There was interest in sharing some of these packages.

- Statistical software is used increasingly to aid in the reorganization and general maintenance of the web sites. This includes information on referrer pages and analysis of the pages for dead links.

- The majority of the requests from users regarding metrics are able to be met. In many cases, the users are more interested in the clarity of the graphical presentations than in the raw data. There is some use of historical trend analysis.

- While there is still a tendency on the part of some users to quote the highest numbers rather than the most meaningful numbers, the participants thought that most users had been sufficiently educated in the differences.

- Discussions with the Federal Web Consortium on the status of modifications to the guidelines and on their interest in statistics would be valuable. [Following the meeting, the Senior Analyst verified that the December 1996 Guidelines are the most recent. She also obtained a report from NCSA on sites of Web Consortium members to show the kinds of statistics that NCSA is collecting. This will be distributed to those participants who requested it.]

- There is also interest in better understanding of the implications of Freedom of Information Act (FOIA) and privacy. While the actual practices are dictated by individual agency interpretations, the group felt that this is an issue that should be included in their decision making more routinely.

- Cookies are a topic of interest particularly as they relate to the future ability to promote public/government interaction and electronic commerce.

- The group noted that education may important to ensure that the content develops know about the statistics that are available. A guide from the Webmasters would be helpful.

5.0 RECOMMENDATIONS

- Host a joint meeting with one or more members of the Federal Web Consortium team to review the current guidelines

  Discussions with Carlynn Thompson of the Web Consortium indicate that they have some issues that could be developed into another draft of the guidelines. Perhaps a joint discussion would result in issues of interest to the CENDI Webmasters being included in the discussions.

- Develop a white paper related to the issues of privacy and the use of "cookies" and other mechanisms to enhance the ability for federal agencies to provide interactive services to the public

  "Cookies" and other methods for learning about customers and their usage patterns can be helpful in developing a site that is responsive to users. This is particularly valuable when trying to develop interactive applications for document ordering online reference. Concern about the privacy implications may be hindering these developments. There may also be alternatives.

- Develop a registry of web utilities (particularly those developed by the agencies and those customized shareware/freeware products) that would promote re-use rather than redevelopment of tools

  Much of the information shared during the workshop involved customization to shareware and freeware products and information about internally developed software to "fill the gaps" of the commercial products. A registry of customized software shareware/freeware and of government developed software and scripts would allow this development to be reused, saving the resources of redundant development.

- Host a CENDI only Webmasters listserv to promote networking among the group

  News and views on usage statistics, web server software, problems, beta test results, etc., were also shared during this meeting. An ongoing need for this type of information could be filled by a Webmasters listserv hosted by one of the CENDI agencies.

APPENDIX A
QUESTIONS FOR AGENCY WEB METRICS PRESENTATIONS
CENDI Web Metrics Task Group
November 17, 1997

In an attempt to cover the topic and also provide information in a form that is easy to document, please address the following questions in your presentation on your agency's web metrics practices. The presentations should be no more than about ½ hour in length. Each person can decide if they would like to entertain questions during the presentation or have them held until the end. Other information can be added as time allows. Overheads and handouts, particularly of report layouts, are encouraged.

The Current Environment:

Describe the general environment in which the web statistics/metrics are being collected. Do you register users? If so, how is that connected to their web usage patterns? Do you have an open environment via the Internet, or is it password protected or otherwise fire walled? Are you running on an Intranet?

What statistics are collected, both usage and user? What software is used to collect and analyze these statistics?

How are the statistics analyzed? What metrics are developed from them?

How are the statistics/metrics used? How are they reported and to whom? What importance do they hold within the organization?

How long are the raw data and the analyzed results retained? How is it archived?

Problems/Issues:

What specific problem/issues have arisen with regard to the statistics that are being collected?

What broader issues do you see -- privacy, FOIA, log storage, etc.?

Future Plans:

What statistics or metrics would you like to use that you cannot now collect? Why can't they be collected?

What do you expect your practices to be in the future?

What software are you now investigating and why?