CENDI/2001-2

Evaluating our Web Presence:
Challenges, Metrics, Results

A Summary of the Symposium
Sponsored by CENDI
And Co-Sponsored by
The National Library of Medicine

Lister Hill Center
National Library of Medicine
April 17, 2001

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

December 2001

CENDI WEB METRICS AND EVALUATION TASK GROUP

Terry Hendrix (Defense Technical Information Center)
Gail Hodge (CENDI Secretariat)
Ed Lehmann (National Technical Information Service)
Linda Tague (National Library of Education)
Mike Thompson (National Agricultural Library)
Dr. Fred Wood (National Library of Medicine) -- Chair
Lisa Zolly (U.S. Geological Survey)

CENDI is an interagency cooperative organization composed of the scientific and technical information (STI) managers from the Departments of Agriculture, Commerce, Energy, Education, Defense, the Environmental Protection Agency, 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.

DISCLAIMER

This document has been prepared as a report of the symposium proceedings and should not be construed as an endorsement of any particular organization, company, or product.

Symposium Speakers

Donald Lindberg, Director, National Library of Medicine

Kent Smith, Deputy Director, National Library of Medicine, and CENDI Chairman

Elliot Siegel, Associate Director, National Library of Medicine, and Symposium Moderator

Fred Wood, Special Expert, National Library of Medicine, and Chair, CENDI Task Group on Web Metrics and Evaluation

Kevin MableyVice President, Strategic & Analytical Services, CyberDialogue, Inc.

Keven Garton, Vice President, Marketing, comScore Networks, Inc.

Bill SilbergVice President and Executive Editor, Medscape, Inc.

Carlynn Thompson, Director, R&D and Acquisition Information Support, Defense Technical Information Center, Department of Defense

Lisa Zolly, NBII Knowledge Manager, Center for Biological Informatics, US Geological Service

Susan Elster, Consultant, University of Pittsburgh and Jewish Healthcare Foundation

Full Symposium Program is included in the Appendix

TABLE OF CONTENTS

1.0 INTRODUCTION AND SYMPOSIUM OVERVIEW

The Internet and the World Wide Web offer all of us tremendous opportunities and challenges in providing access to scientific, technical and medical information. We want to provide content that is timely, accurate and understandable, delivered by systems that are friendly and relatively easy to use; and in the final analysis, capable of empowering our users to make informed decisions — especially when health and wellness are at stake.

Upper most in the minds of many is the question, "How good a job are we doing?" This evaluation challenge can be especially daunting because the Web environment, itself, imposes a whole new set of constraints – technical, legal, and ethical -- on getting useful feedback. Our users in many instances are anonymous. Privacy protection is essential, especially for certain kinds of content applications, such as patient health information. Just measuring Web traffic and site usage can be complex and controversial, and an evolving state of the art when it comes to making reliable counts and projections.

The Symposium, summarized in this report, was intended to raise awareness of these issues – for those not already facing them head-on. Those with firsthand experience shared what they knew and raised the bar on what will become the best practices of tomorrow. From a technical standpoint, the symposium looked at audience measurement and drill downs; user surveys – online and telephone; focus groups; comparative Web site analyses; usability testing; and end-to-end Internet connectivity testing. A mix of perspectives was represented -- from information providers in both the public and private sectors who are getting the job done, and from members of the consulting community whose business it is to help organizations do the kinds of evaluations that are needed.

2.0 THE NEED FOR METRICS AND EVALUATION

The reasons for metrics and evaluation differ by organization and even within an organization, depending on the products and services being evaluated. However, there are benefits both for the private and public sectors.

For commercial companies, metrics are required as part of the assessment of the viability of the company in delivering Web-based information. Commercial companies are often asked for information by stockholders, the SEC, or the media. Metrics allow companies to improve their business models. The more that can be measured, the more can be predicted. Despite the complexity of the Internet environment, it may be possible to learn more about how users use information than in the print environment (Silberg). In the print environment, there is no way to really know whether people are answering consistently and accurately. Have they really read what they say they've read? On the Web it is possible, within privacy and ethical constraints, to actually gauge this.

Organizations in the public sector need metrics in order to provide measures of their performance in conformance with legislation, regulations and directives, such as the Government Performance Reform Act (GPRA) for federal agencies. Metrics and evaluation help to measure productivity

in a time of tight budgets. As e-government initiatives progress, these approaches allow agencies to determine how they might contribute most appropriately to these efforts. In addition, metrics and evaluation allow the public sector to conduct research and development into information systems in support of their missions.

For both the public and private sectors, metrics and evaluation are key to continuous improvement because they provide mileposts. Without a gauge or benchmark, it is impossible to determine the effectiveness of any modifications that are made to information systems (Zolly). Along with mission and focus, learning from metrics is a major factor in why certain Web sites work (Thompson).

Given the fact that metrics and evaluation are beneficial, what is it that we want to know? Based on CyberDiaglogue's experience, there are several questions that most clients want answered. What is the demographic composition of my audience? What contents and features are they using? How do I measure up against similar sites? What is the value of what the organization is doing? And one of the hardest questions -- which visitors are more valuable and why? Rather than having everyone coming to your site, it is important to have those who are valuable to your business or those you are destined to serve. (Mabley)

Ultimately, organizations want to know if the services and products that are being provided are making a difference (Silberg). What value do they have for the user? The people who come to the services are not coming with information storage and retrieval models, but, rather, people with a need driven by events in their lives or organizations. Ultimately, organizations would like to be able to determine what the outcome has been (Lindberg). Has the user's behavior, whether health or purchasing behavior, been changed by what has been provided? Has the information provided actually added to the user's understanding?

However, despite the benefits and an identification of the questions that organizations would like answered, there are technical, legal and ethical constraints in the Internet environment that make it more difficult to obtain the information needed to answer these questions. Measuring even the usage of web sites can be time consuming and controversial. The Web and Internet environment are more complex, in part because of the speed and decentralized nature of the Internet environment. Internet-based products and services are often of a distributed nature, being created from content and services from different companies in different locations. New services, such as distance learning, may require new metrics.

For government agencies, a key difference between the Internet environment and the previous print and online environments is the anonymity of users. Many services that previously required registration (including those at DTIC and NLM) now can be accessed anonymously. While the provision of anonymous services is particularly important to government agencies, the lack of user information presents a challenge. In terms of metrics, organizations don't have much information on what some user groups, such as medical professionals, are doing with online information. It is also hard to extrapolate, because there is variability across specialties and multiple audiences require different, specific metrics.

In some cases, the types of users are so broad that it is difficult to evaluate their needs. For example, users of the NBII include natural resource and land managers, the general public, K-12 students, researchers, etc. The audiences have different information needs, require different levels of detail, and use different terminology and techniques when searching for information.

The bottom line is that it is more difficult to assess outcomes in the Internet world, it is harder to get meaningful direct feedback, and the system is much more complex, with more information sources, options and players than in the pre-Internet environment. This more complex environment requires more complex systems for analysis. (According to Ashby's Law of Requisite Variety, the complexity of the evaluation system must meet or exceed the complexity of the real world system that it is trying to evaluate.[1]) "Traditional evaluation approaches are less applicable or feasible; clearly new or reinvented approaches are needed." (Wood)

3.0 WEB METRICS AND EVALUATION METHODOLOGIES

In order to determine if the metrics and evaluation methodologies are gaining in complexity to match the systems they are meant to measure and evaluate, it is interesting to look at the state-of-the-practice only three years ago. The CENDI Metrics and Evaluation Task Group conducted a baseline study in 1998 [http://www.dtic.mil/cendi/publications/98-1webmt.html] and then updated the study in 2000 [http://www.dtic.mil/cendi/publications/00-2web_met_eval.html]. The work group included members from seven of the ten CENDI agencies. When the 1998 results are compared to the 2000 study, it is obvious that metrics and evaluation are becoming more important to the development and management of Web services. The methodologies are becoming more complex in response to the increasingly complex web environment, in accordance with Ashby's Law.

The recent study found that all participating agencies are using Web log analysis software. However, while there is still variety in the software that is used, there is increasing commonality in the simple metrics such as page counts, unique users, etc. First generation Web user surveys are being used, but with mixed results. Usability studies are now typically part of Web design. Surprisingly, most of the agencies are monitoring Internet connectivity and performance, but the details vary widely.

There was a great deal of progress in those two years. However, the current situation could be characterized as first generation, or generation 1.5 in terms of Ashby's Law (Wood). While quantitative information from Web log analysis is important, evaluative and qualitative information based on usability testing, satisfaction surveys, customer feedback and other evaluative techniques are used increasingly.

3.1 Web Log Analysis

The most common metrics approach is the collection and analysis of Web log data. Web logs commonly provide metrics such as the number of "page views", the number of unique visitors,

[1] Ashby, W. Ross, *Introduction to Cybernetics*, Routledge, 1964, and Stafford Beer, *Decision and Control*, John Wiley, 1966.

the domains of the visitors, and referral and exit information. By analyzing the logs with special software tools and some customized code, organizations can identify various aspects of Web usage. In addition to the analysis of a given set of Web log data, the analysis of the data over time provides other important information.

For example, a variety of metrics are provided on a weekly basis to DoD organizations for which DTIC hosts Web sites, based on analysis of log data, including the number of accesses. They provide the unique hosts that are coming in, the most frequently requested page, and the number of megabytes that are transferred. When DTIC's customers originally requested information about the most frequent hosts, there was initial concern at DTIC about privacy issues. However, the most common are from network hosts, so there isn't any concern about privacy issues. While most customers are concerned about the specific web-site statistics, senior managers at DTIC are provided with a roll-up summary. By analyzing the log data for the types of Web browsers that are being used, DTIC can make decisions about moving from one set of services to another.

MedScape found that the traditional advertising metrics model, which includes page views and unique visitors, is still valuable in the consumer business side (Silberg). However, in the professional side of the business these metrics are not as valuable. Session time (stickiness of the site) is important for sites such as MedScape that deal with professionals, since companies are competing for "mind-share" – the user's time. As a corporation, if you can say that a certain amount of time is spent on your site, knowing that professionals have a limited amount of time to spend in information activities to begin with, then this "stickiness" is a competitive advantage over other sites. "Stickiness" is an important metric as information companies try to insert themselves into the professional's daily routine.

"Click Through" metrics identify the number of times a user clicks on an available link. For example, if the user sees an advertisement on a page, does the user click on that ad? This metric is particularly valuable for commercial sites that depend on advertising. "Click through" is a valuable metric for MedScape's weekly customized e-mail product that includes links to specialty sites. For this product, the click through rate is approximately 7 percent compared to a general Internet-wide rate of 2-3 percent.

"Click Paths" capture the chronology of where the user came from, where he went and in what order while on the site, and where the user went when he left. In a site rich in links this is a valuable metric. It may also provide an indication of other companies with which it is important to collaborate to enter into partnership agreements.

3.2 Online Intercept Surveys

Intercept survey methodology intercepts the user based on a counter (every 10^{th}, every 20^{th}, etc.). An invitation in a pop-up window invites users to participate in the survey. The user clicks through to the survey. In the case of CyberDialogue's service, a referrer tag, which is captured in the log file, is used to return the user to the page from which he or she came when the survey began.

The surveys are provided through HTML forms and generally capture data that is similar to computer assisted telephone surveys. However, Web surveys have several advantages, including the fact that answers to open-ended questions can be easily obtained. It is possible to dynamically check for errors. Different types of media can be provided to the survey, such as full motion advertisements. HTML surveys are "in the right time and in the right medium." The survey is conducted during the activity that is being evaluated; i.e., the Web session, so the user's response to the experience is very immediate (Mabley).

An online user survey was recently completed on MedLinePlus, which is geared primarily toward the public health consumer. For this survey, NLM contracted with CyberDialogue, Inc. The survey was developed collaboratively and was cleared through the blanket OMB approval process. Approximately 3,000 users were selected randomly through a pop-up screen that appeared on the user's screen.

Intercept surveys provide information about a user that can only be self-provided and that cannot be revealed through log data. Examples of the types of data collected during the MedLinePlus survey include the type of person (patients, families, medical practitioners, etc.), the number of visits to the site in the past, and the access location (home, office, etc.). This information can then be used in concert with Web logs or traditional survey methods to paint a more complete picture of user demographics.

An online survey can capture information about compliance or understanding. In the case of the NLM survey, more than half of the respondents indicated that the use of the site increased their understanding of a particular condition or treatment.

Satisfaction questions can also be asked via an online survey. The responses to these questions can then be compared to normative measures that CyberDialogue has aggregated from several hundred clients in specific industries. As in traditional surveys, the people who respond tend to be those who like the organization or the product. However, the negative responses received will be more extreme (Mabley).

Intercept surveys also allow the client to build a panel for use in future surveys. As the last question of most surveys, CyberDialogue suggests that the client ask respondents if they are willing to be contacted about participation in future studies. This is a powerful tool, because it provides a database of people who have agreed to participate already and who are interested and knowledgeable. This decreases the recruiting costs for future research initiatives. In the NLM study, about one-third of the respondents indicated interest in participating in a future study.

However, online surveys must be developed with care. The user's reaction to survey length appears to change in five-minute increments. There is a significant drop off rate after 5 minutes and then again after 10 minutes. A survey more than 20 minutes in length may need to be split into multiple surveys or the use of another survey approach should be considered. The download rate for the survey should also be tested, since that will impact the time that people spend. High bandwidth audiences may appreciate having the whole survey download at once, while only one or two questions per page would be more appropriate with normal bandwidth.

3.3 Online User Panels

The online user panel technique uses Internet users who have agreed to allow their usage to be monitored. Other vendors are expanding the audience to the office, educational, and international sectors.

NLM contracted with PCData Online for Internet audience measurement and online panels. PCData provided access to data from 120,000 home users who allowed their usage to be monitored. The results were extrapolated to a nationwide audience. The audience is now being expanded to office and educational sectors.

Dr. Wood presented results from one to two years' worth of data collected by PCData. The metric to which PCData gave highest priority was the number of unique users using the Web site in a given period of time. The three top-level government domain sites, when adjusted for income tax and student loan seasons, were the U.S. Postal Service, NASA and the National Institutes of Health. PCData also provided drill down data beneath the top-level domain. Within the NIH domain, the NLM domain usage ranged from 43 to50 percent of total NIH usage during this period.

The margin for error on the results increases with the degree to which you drill down into the sites. If you drill down too low within a site, PCData would not even provide a confidence number.

PCData's client base was bought by comScore Networks, Inc. The data collected from PCData is being compared with that from comScore. Even though the absolute numbers are somewhat different, the relative ranking of Web sites is similar. It may be that there is validity in a relative sense, more than in the raw numbers.

comScore's approach to user panels was founded on two ideas. The first is the concept of the panel size, which has traditionally been used in market research. There is often discrepancy between what companies report and what syndicators report. A recent study concluded that the sample size was the primary problem. The Internet is "so broad and so deep and [there are] so many users and so many transactions, you have to step out of the traditional way of doing research." (Garton) Therefore, comScore set out to build an extremely large panel to statistically and accurately reflect the Internet.

In order to have the same level of precision with the Internet that you have with traditional audiences you would need to have a user panel of 1M people. There are a variety of panel recruitment methodologies. The oldest is perhaps random digit dialing (RDD). Other methods include random direct mail and television advertising. Often a multi-channel approach is used which combines these techniques. New approaches to recruiting panels include partnerships with Internet Service Providers and e-commerce companies.

comScore has built a panel of 1.5M people primarily through online recruiting methods and the use of incentives. A calibration sample through random methods is used to do projections, weighting to eliminate bias. A weekly enumeration is done in order to compare the total panel to

the actual population. The goal is to have a panel of 1 percent of the U.S. Internet population. It is also important to look at different audience segments – home, work, school, international, etc. The at-work sampling tends to be very small because of company prohibitions.

The second concept addressed by comScore related to the appropriate metric. The initial metric for evaluating Web sites was the number of hits or visitors to the site, but being visited by a lot of people does not necessarily mean that the organization is serving its constituency effectively. Therefore, comScore developed technology to capture information at the browser or machine level rather than at the user level. A network of 300 servers track visitors from 246 countries, monitoring across backbones and hosting organizations. comScore collects billions of pages per month from over a million Web sites, and it actually serves the Web pages to the users who request them.

Data syndicators use three basic techniques for collecting data – web log file analysis, client site metering software, and browser proxy configuration. There are many companies that will analyze web logs. Of the remaining techniques, client site metering is most widely used. It requires the user to install software on his or her personal computer. The data is collected on a real time basis, stored in a file on the user's computer, and then transferred periodically to the syndicator.

The browser proxy configuration approach has been pioneered by comScore. When a person agrees to become a member of a comScore panel, he is sent to a Web site where he fills out basic demographic information. comScore then resets all the accesses from the user's machine through its proxy server, allowing the comScore software to capture pages and transactional information. All the information is sent to a secure server center, where the information is extracted, scrubbed and put into a data warehouse. Personal or household information is either discarded or scrubbed. Proprietary analytical software is used to analyze the information. comScore also purchased a decision support tool that allows customers to do ad-hoc queries against the customer's dataset in the repository. Information can be sent back to the customer in a variety of formats (Excel, Word, etc.), including an online Web-based report.

Companies that use browser proxy configurations differ in the type of Internet traffic they can capture – HTTP, HTTP-S, private network (such as AOL), or client-based application information. They also differ in the types of projection methods that are used. comScore uses a combination of approaches to minimize bias, including a weekly random digital dial survey and a calibration sample, which is a randomly recruited sample, used to balance their overall panel.

A variety of traditional metrics are reported by most syndicators, including unique visitors, popularity versus another site, visits per visitor, minutes per visitor, etc. Some newer metrics are being developed. comScore has developed the "Buying Power Index" which is the value of the site versus the amount spent by the visitor across the whole Internet. Work is progressing by comScore and others to measure and assess transactions and banner impressions. More customized methods are also available including where traffic is coming from and where the traffic goes next. Cross-visiting analysis can also be performed.

In addition to the intercept survey methodology, CyberDialogue provides "competitive analysis". From its database of approximately 100,000 online users, users are screened to be in the client's category of interest. The users are asked to randomly visit several competitive sites and to take a survey for each one. The analysis of the data is similar to a SWAT (Strengths, Weaknesses, and Threats) analysis. Attributes of importance to the user are plotted against the client's strengths and weaknesses in satisfying those attributes. The same is done for competitor's sites and the results are compared.

3.4 Usability Testing

Usability analysis includes heuristic evaluation and actual testing. Heuristics is the application of accepted standards. The industry is beginning to develop a set of heuristics, best practices and guidelines that can be used to review existing Web sites or plan initial designs. Guidelines for this type of analysis are available in books, from conferences and on the Web. While consultants can be hired to perform this type of analysis, it is also relatively easy to conduct this analysis using internal staff. The negative aspect of using heuristics is that the results are general and not specific to a particular Web site, its content or audience. Much of the guidance that is available is for e-commerce, and government sites are not well addressed.

DTIC has used heuristics to evaluate its Web site. Under contract, NCSA applied the Nielsen ratings to DTIC's web sites. The results were "rather shocking," including comments about the "kludgyness" and lack of flow of the site (Thompson). Color schemes, inconsistent design, etc. were mentioned. A major redesign of the site was completed and DTIC got a much better rating two years later.

Usability tests involve bringing actual users into a controlled environment and watching as they use the specific site. This approach is more expensive, requiring significant resources and planning. However, the results are specific to the particular Web site and to the particular content and audience.

Similar usability test results were obtained by DTIC. Ten common tasks were identified. A representative sample of users was recruited. A variety of computer systems and browser types, including a Braille reader, were included in the study. The subjects were observed and asked to think out-loud about their experience. Once again, this was an eye-opener. Users found the search engine frustrating. They did not understand Boolean logic or the search results.

The National Biological Information Infrastructure Project Team used a combination of heuristics and usability testing to analyze the NBII Web design (Zolly). Since the team knew that there were major problems with the existing NBII Web design (18 months ago), the team decided to first redesign the site based on a heuristic analysis that identified the major problems. These problems were resolved in the redesign, and then the usability testing was conducted on the new site.

The usability test was conducted at the Usability Laboratory of the Bureau of Labor Statistics (BLS). The lab was built for BLS' own needs, but when it is not in use, it can be scheduled for up to a week by any other federal agency free of charge. The lab is equipped with audio and

video recording devices at various angles (including the user's face and hands) and a one way mirror from which observers can watch the users as they use the system.

A consultant was hired to conduct the test. The use of a consultant proved to be very beneficial, because she simplified the questions originally developed by the project team and the testing approach. The consultant also conducted the actual tests and was able to assure the participants that she was an unbiased observer, since she had nothing to do with the Web design or with the NBII Program.

The only aspect of the usability test that was not performed by the consultant was the recruitment of participants, since this would have added significantly to the cost of the test. Instead, the NBII advertised on various listservs, at local universities and within the agency itself. A good representative sample of types of users included librarians, researchers, an environmental lawyer, Congressional staffers, and students. Of the12 people invited, 10 people actually participated.

Terminology proved to be a major issue. The labels for functions and for the categorization of resources, which the Team had labored over extensively, were identified as major problems. Terms such as Education, which seemed transparent to the developers, caused confusion for the testers.

There is a need for increased attention to usability testing, because this can avoid a lot of problems and issues before they occur. A combination of heuristics and usability testing is likely to provide the best results. Heuristics can correct obvious flaws, while testing can focus on specific tasks.

3.5 Environmental Scanning

Environmental scanning techniques integrate various metrics and evaluation techniques. Environmental scanning explores the information seeking behavior of a particular constituency (Elster). An environmental scan is driven less by the need to answer hypotheses than to collect information. It is more descriptive, searching widely and looking for patterns. An environmental scan may involve advisory panels that are able to act on what is discovered. The goals of the scan can be to identify consumer information patterns, the types of technologies used, and the problems encountered. It can answer questions about what precipitates a search, how the users went about the searching, how they evaluated search results, and what role information technologies play.

In this scan of health information seeking behavior in Allegheny County, Pennsylvania, pre-survey focus groups were used prior to the main study. These pre-survey focus groups are extremely beneficial when you know little about the topic and you need to pare down to the core issues and the audience to be included in the primary survey.

The pre-survey was followed by a telephone survey and a library patron survey. The telephone survey included a random digital dial survey of 1000 people. Follow up methodologies included meetings with support groups in some health areas. A short survey of patrons and librarians was

also conducted to determine the impact of public libraries on the constituents' information seeking behavior.

An environmental scan can also identify some specific cultural issues that would not otherwise be identified. For example, Ms. Elster indicated that her research found that there is a culture of distrust of government information sources particularly among lower-income constituencies. This can have an impact on the results of more quantitative metrics and evaluation approaches.

4.0 INTEGRATING EVALUATION APPROACHES

One of the major differences between first and second-generation initiatives is the degree to which the various evaluation approaches -- web log analysis, online surveys, online user panels, usability testing, and environmental scans -- are integrated. Second generation initiatives not only include more sophisticated uses of specific evaluation approaches, but more integrated strategic combinations of the approaches. The integration can provide cross comparisons, identify bias among respondents, and result in the collection of a much richer set of data. The various metrics and evaluation approaches complement each other.

With a variety of evaluation approaches, it is possible to do cross comparisons. Dr. Wood presented some cross comparisons between the survey results and other resources. Forty-two percent of the random survey respondents self-reported that they visit MedLinePlus at least monthly. This compares positively with what CyberDialogue found in terms of high site loyalty. There was high home use, a result that was also found by other types of surveys conducted. The survey also supported the fact that users are highly focused on specific conditions, diseases or diagnoses.

A combination of approaches can be used to identify bias in the respondents, particularly in intercept surveys where the non-response rate can be substantial. CyberDialogue compared the "time of day/day of week" metric from intercept survey responses to the general usage patterns in the Web log data. There was a high correlation, which indicates that the variation by time of day/day of week was not a point of non-respondent bias. The number of U.S. versus non-U.S. users varied somewhat between the intercept survey and the Web logs, but approximately ¾ of the log data cannot be resolved in terms of geographic origin because these users are accessing via Internet Service Providers. The third factor evaluated was repeat visitors (two or more times per month). Here, there was a moderate difference between the number of those who self-responded as multiple users and the results obtained from analyzing the Web logs. The bottom line is that the non-response bias, which was judged to be relatively minimal for the NLM survey, can be identified by looking at the results of a variety of metrics and evaluation approaches.

Offline surveys in the same industry or market sector can be compared with online surveys. This allows the client to understand the differences between its Web users and those of other products. For some commercial clients, CyberDialogue uses cookies to collect information that is then compared with or added to the information gathered through the intercept survey responses. The user has the option to be "cookied" and tracked for a 30-day period. This technique can also be

used on a random group that provides benchmark information as a general Web user "control group". The information collected by these accepted cookies can be used to determine if the client is measuring the users in the intercept survey that it wants to measure. In addition, the actual area on a Web site visited by the user can be plotted against the intercept survey data providing a much richer set of data. NLM opted out of using these persistent cookies.

The results of the different survey techniques can be used to map audience demographics against their responses on content interests and Web usage/ interface behavior. This is a classic audience segmentation technique. A few quick questions answered by the user at the beginning of each visit can be used with the database and rules identified through the metrics and evaluation to customize the content and behavior of the site based on slotting the user into a predefined audience segment (Mabley). Looking at audience segments across the various types of survey methodologies is also very insightful.

5.0 QUALITY OF SERVICE AND INTERNET CONNECTIVITY

Ultimately, organizations are concerned about how their systems are viewed by users. A critical part of this view is "quality of service". This involves issues such as reliability and response time. For example, those who work in telemedicine are interested in high-speed telecommunications for short periods of time only when it is needed. The metrics in this case are connected to the patient's wait time during diagnostic screenings. Insights from metrics and evaluation studies can contribute greatly to identifying what quality of service means.

Internet connectivity is closely related to quality of service. Internet connectivity has been studied at NLM for approximately three years. NLM identified the key metrics as the time required for downloads and other transactions to be conducted over the Internet. The Internet pathways between NLM and six different target sites, four U.S. and two foreign, were studied. The bulk transfer capacity (effective throughput in megabits per second) and mean round trip time (measured in milliseconds) were calculated in 1998 and 2001. This comparison shows improvement in both throughput and round trip time, particularly for the international links, which were much slower than domestic in 1998. Some of the domestic sites have now switched to the Abilene high bandwidth network, rather than the commercial Internet, but despite these changes there were no dramatic improvements.

One might question whether the problem is at NLM's end of the connection or with the recipient's. In order to answer this, NLM compared the results between NLM and two recipients in the same geographic area. One recipient had a much lower baseline throughput than the other recipient, but not much degradation over the weekends. There was less variability, and the pattern was cleaner. The difference in the connectivity patterns of the two recipients indicates that the problem is less at NLM than at the recipient's site.

In another approach for measuring connectivity, both NLM and DTIC have experimented with Keynote, Inc., which has a proxy network of 25 U.S. and up to 25 foreign sites in major metropolitan areas, all using T1 or better lines. KeyNote measures how long it takes every hour to download the client's Web page. They compare the client's download speed against a

baseline of 40 user-oriented company Web sites. This type of data can be evaluated over time to identify long-term trends or to look for anomalies during certain periods. By comparing the DTIC response times against KeyNote's benchmark data, DTIC is able to diagnose performance problems. In the early days of the Web DTIC took advantage of the information provided by KeyNote to resolve some of the routing problems that occurred.

DTIC's services require a great deal of stability and security. For issues such as Internet performance, DTIC has indicators that are constantly evaluated automatically. If any of the indicators drop below established thresholds, notification is provided automatically to the network administrator.

DTIC also monitors bandwidth utilization. This information is used on a daily basis to identify problems and to track the need for customer support. Monitoring software runs across Web servers looking for different types of activity. It indicates where there is trouble or where attention is needed. System administrators are alerted through automatic e-mail or pager notification. In addition to monitoring the current situation, DTIC uses Internet connectivity metrics for capacity planning.

6.0 FEDERAL GOVERNMENT INFORMATION POLICIES THAT IMPACT M&E

While the public sector has challenges related to successful M&E implementation, the challenges of the government are unique in some regards. There are several federal government policies that make certain aspects of Metrics and Evaluation difficult to implement. These include privacy, security, incentives and survey practices. The government has to prove its case, but often without the means to collect the data that is needed. This requires more ingenuity.

6.1 Privacy and Cookies

Standard log file characteristics include date and time, ID, browser, and an optional registration file or cookie. Cookies can be used in two different ways. Session cookies last only for the extent of the session and are not permanently saved on the user's machine. These cookies allow log file data, which is linear by date and time, to be rolled up to recreate a particular visit. Permanent cookies are saved on a user's machine for a period longer than the session in order to track user behavior. Cookies can also improve the usefulness of web sites by remembering what the user has done before or by storing user profile characteristics.

The limitation of cookie use constrains the government's ability to analyze Web usage. According to current OMB memos, a session expiring cookie can be used. A persistent cookie requires approval from the agency's department Secretary. DTIC has one persistent cookie for which it is seeking approval. This cookie will be used in its portal service to provide profiling information required to customize the content provided to the user.

Increasing customer and governmental concern about privacy in certain areas such as medicine and financial services are affecting the private sector as well (Silberg). Even for commercial organizations, particularly when dealing with consumers, there are legal and regulatory issues

related to the tracking of what users do. The questions include what information can be collected, what information can be provided to others, and what can the third-party collectors of the information do and not do with the information collected on behalf of the organization?

6.2 Security

Privacy and Security are a balancing act. It is necessary to secure an organization's Web servers, while balancing privacy issues. Privacy and security policies should be posted on the Web and in the case of the government, also published in the *Federal Register*. They should spell out how, what and why information is being collected.

6.3 Paperwork Reduction Act and the OMB Survey Clearance Process

The OMB approval process for surveys is based on the Paperwork Reduction Act. It requires that an agency must get approval from OMB for an information collection request, which asks the same questions of more than nine individuals. This process can hinder collection of data for Web analysis, since the life of some Web sites may be shorter than the time required to get approval for the survey. Alternatively, the OMB blanket approval process addresses agency requests for multiple surveys of a similar type. Once the blanket approval has been granted, individual questionnaires must still be sent to OMB, but the approval is an exception process. If OMB does not respond within 10 days, the agency can assume that it is okay to use the survey. Most agencies also have expedited intra-agency approval processes.

6.4 Incentives

Often Web publishers offer incentives such as cash, gifts, online currency spenders, etc., to users, who participate in online panels or usability tests. More than half of CyberDialogue's online survey clients use incentives, and, generally, the response rates are higher in an incentive survey as opposed to a non-incentive survey. While Internet surveys have enjoyed higher response rates than more traditional methods such as phone surveys over the last few years, this edge is beginning to decline. While incentives may become more important, it is difficult for government agencies to offer incentives.

Despite the better response rate generally attributed to incentive surveys, NLM's survey, which did not use incentives, had a better response rate than the average for non-incentive surveys. Whether the presence or absence of incentives impacts the survey's response rate or quality may depend on how important the survey is to the respondent (Mabley).

There may be some incentives that could be used in the government environment. For example, EPA distributed screen savers, since there is no concern about the transfer of money, and the screen saver can be made available for download immediately upon responding to the survey.

7.0 COST AND FUNDING OF M&E EFFORTS

The cost of developing metrics and evaluation studies should not be minimized. There are contractor costs, but also considerable internal costs, based on an initial learning curve. It is

expected that there would be some economies of scale if more than one survey were done (Wood).

Budgets in the government environment are an issue. Therefore, collaboration should be encouraged. It was noted that not only would there be cost savings, but knowledge savings as well. More opportunities for sharing information across and within sectors, such as this symposium, were recognized as being invaluable in moving forward.

Organizations could collaborate on the development of metrics that are meaningful across organizations. Based on these share metrics, an interagency group could approach a vendor as a group. The results could also be used to jointly develop decision models that consistently use metrics to support the decisions made by scientific and technical information programs. These models would add meaning to the data that is being collected.

8.0 CONCLUSIONS

The urgency for metrics and evaluation has been accelerated by the marketplace and by the government's need to take a market-centered approach to provision of services. The old rules apply. Understanding your business is key to understanding the metrics that are most important. Business issues that the organization are trying to solve should be identified, because collecting has a cost and it is important to ensure that the organization is getting something for the metrics that are being collected and the evaluation that is being done.

A corollary to "know your business" is "know your audience." Business models are changing, and it is important to understand the audience for Web-based products. It is important to communicate with customers, with others in the same sector who face similar metrics and evaluation issues, and between public and private sectors.

Evaluation can increase awareness, accessibility and usage of information; help evaluate positive impacts of the information, advance efforts to reduce disparities and the digital divide, improve the quality of information and of information professionals, and encourage innovations in a particular subject area and the application of those innovations. Why does this really matter? It matters because evaluation helps us to effectively perform our missions – whether government, commercial or non-profit -- and to understand how information can improve people's lives at the individual, national and global levels.

Do we have evolution in Web evaluation? Yes, we have evolution, and we may have some aspects of revolution. However, what kind of revolution are we going to have? We hope this discussion has helped to chart the way and serves to measure the progress as Web metrics and evaluation methodologies meet the challenge, produce better, reliable metrics, and enhance the results of our Web presence.

Appendix

Symposium Program
"Evaluating Our Web Presence: Challenges, Metrics, Results"

April 17, 2001

EVALUATING OUR WEB PRESENCE:
CHALLENGES, METRICS, RESULTS

CENDI

A CENDI-Sponsored Symposium
Co-Sponsored by the
National Library of Medicine
Auditorium--Lister Hill Center (Bldg. 38A)
National Library of Medicine/National Institutes of Health
Bethesda, MD 20894

Tuesday, April 17, 2001

UNITED STATES National Library of Medicine

Morning Program

8:00 am	**Registration**
9:00 am	**Introduction and Welcoming Remarks** Kent Smith, NLM Deputy Director and CENDI Chairman **Symposium Overview** Elliot R. Siegel, NLM Associate Director and Symposium Moderator
9:30 am	**Keynote: Donald A.B. Lindberg, NLM Director**
10:00 am	**A Strategic Approach to Web Evaluation** Fred Wood, Special Expert, NLM, and Chair, CENDI Task Group on Web Metrics and Evaluation
10:30 am	**BREAK**
10:45 am	**Online Surveys for Web Evaluation: Research Methods for Understanding Your Audience** Kevin Mabley, Vice President—Strategic & Analytical Services, CyberDialogue, Inc.
11:15 am	**Collecting and Analyzing Web Usage Data from User Panels** Kevin Garton, Vice President—Marketing, comScore Networks, Inc.
11:45 am	**Web Evaluation from a Corporate Perspective** Bill Silberg, Vice President and Executive Editor, Medscape, Inc.
12:15 pm	**LUNCH**

CENDI

EVALUATING OUR WEB PRESENCE: CHALLENGES, METRICS, RESULTS

Afternoon Program

1:30 pm	**Web Sites That Work: The Role of Evaluation at DTIC** Carlynn Thompson, Director, R&D and Acquisition Information Support, Defense Technical Information Center, U.S. Department of Defense
2:00 pm	**The Role of Usability Testing in Web Site Design: The National Biological Information Infrastructure's Experience** Lisa Zolly, NBII Knowledge Manager, Center for Biological Informatics, U.S. Geological Survey
2:30 pm	**Consumer Health Information in Allegheny County, Pittsburgh, PA: An Environmental Scan** Susan Elster, Consultant, University of Pittsburgh and Jewish Healthcare Foundation
3:00 pm	**BREAK**
3:30 pm	**Panel Discussion: Web Evaluation--Lessons Learned, Future Directions** Moderator: Fred Wood Panelists: Susan Elster, Kevin Garton, Kevin Mabley, Bill Silberg, Carlynn Thompson, and Lisa Zolly
4:15 pm	**Concluding Remarks and Wrap-up** Elliot R. Siegel, NLM Associate Director and Symposium Moderator
4:30 pm	**ADJOURN**

UNITED STATES National Library of Medicine