Sampling the Web: The Development of a Custom Search Tool for Research

Chareen Snelson, Ed.D.

Boise State University

December 2005

Abstract

Research designed to study the Internet is beset with challenges. One of these challenges involves obtaining samples of Web pages. Methodologies used in previous studies may be categorized into random, purposeful, and purposeful random types of sampling. This paper contains an outline of these methodologies and information about the development of a custom sampling tool that may be used to obtain purposeful random samples of Web page links. The custom search application called Web Sampler works through the Google Web APIs service to collect a random sample of pages from search results returned from the Google index. Web Sampler is inexpensive to develop and may be easily customized for specialized search needs required by researchers who are investigating Web page content.

The Internet is a vast network of interconnected computers that supports rapid access to and exchange of digitally encoded information including e-mail and Web pages. Working on top of the Internet is the World Wide Web, which is a hypertext system that enables retrieval and display of Web pages through the use of browser software. During the years since the idea for the Web was initially conceptualized by Berners Lee (1989, 1990), it has grown to include several billion publicly accessible pages that may be indexed by search engines (Gulli & Signorini, 2005). The Web has become a huge gateway to information that may be accessed by anyone with a computer, Web browser software, and Internet access. The quality and scope of information available on the Web is not fully documented through research due to the continual growth and dynamic nature of Web-based content. Web research comes with its own set of unique methodological challenges such as access issues, population definitions, sampling procedures, and selection or development of appropriate software applications for individual research needs. The purpose of this paper is to articulate the challenges associated with Web research, summarize several strategies that have been used to obtain samples of Web pages in previous studies, and describe the development of a custom search application designed to obtain purposeful random samples of Web pages.

The Challenges of Web Research

A variety of challenges need to be addressed in any research study designed to explore the Web. The process of finding solutions to these challenges may involve an evolution of traditional research methodologies. Consider the basic research procedure of defining a population and selecting a suitable sample to study. In Web-based research, complex problems emerge when defining populations or selecting samples. For example, if the population is defined to be all existing Web pages, then access problems are encountered. A portion of the total population of all existing Web pages is private, and public access is

restricted. If the private pages are eliminated from consideration for research, then the population under study may be redefined as the set of Web pages open to the general public. Samples may be collected from the set of public Web pages without access problems since they are available to everyone. Unfortunately, removing the access barrier does not solve the entire sampling problem. Web page sampling has not proven to be simple or straightforward. Several variations of sampling procedures have been used or proposed in order to extract information from the Web (Henzinger & Lawrence, 2004). As shown in Table 1, an examination of sampling techniques that have been used in Web research reveals that they tend to fall within the broad categories of random sampling, purposive sampling, and purposive random sampling.

Table 1: <i>List of Sampling Strategies for Web-Based Research</i>	
Sampling Method	Examples
Random Sampling	Random IP Addresses Random Walk Random Sampling From a List
Purposeful Sampling	Hand-Selected Links
Purposeful Random Sampling	Random Sampling From Search Results

Random Sampling in Web Research

The premise underlying random sampling in research is that every member of the defined population has an equal chance of being selected. When applied to the population of public Web pages, this would suggest that every page has equal likelihood of being chosen for a research sample. A true random sample of Web pages, like any other random sample, would be representative of the larger

population and would allow findings to be generalized (Gall, Borg, & Gall, 1996). Difficulties arise, however, when actually attempting to collect a simple random sample of Web pages. This is partially due to the dynamic nature of the Web. Web pages are continually being changed, posted, and removed, so the population is in constant flux. It seems uncertain whether or not a random sample collected on one day is still representative a day or a week later. Therefore, studies of content existing on the Web have a time-sensitive aspect to them. Findings from Web research studies provide information about the state of the Web at certain time checkpoints. This can be a useful line of inquiry for longitudinal studies. For example, a series of studies conducted at regular intervals can be combined to build a picture of how the Web changes over time.

This approach was used in The Web Characterization Project (O'Neill, Lavoie, & McClain, 1999), which involved a series of studies initiated by the OCLC (Online Computer Library Center) Office of Research. This project was designed to answer fundamental questions about the size, structure, and content of the Web. Analysis of the data collected during the five years of the Web Characterization Project indicated that the number of public Web pages doubled between 1998 and 2002. By June 2002, the estimated size of the public Web had grown to approximately 3 million sites containing roughly 1.4 billion individual Web pages. Analysis of the use of metadata in the code of Web pages was also conducted during the project to learn more about content descriptions used by many search engines to index pages (O'Neill et al., 1999).

At the beginning of the Web Characterization Project, a methodology was developed for sampling the Web (O'Neill, McClain, & Lavoie, 1997). This methodology involved generating a random set of Internet protocol (IP) addresses to sample. An IP address is actually a set of four numbers from 0 to 255 that is separated by dots (i.e., 216.239.51.100). The IP address system is a standard method for identifying the location of computers on the Internet. A random set of IP

address numbers drawn from the entire set of possible numbers should theoretically provide a way of selecting a representative sample. This is the basic strategy that was used to sample in the Web Characterization Project. Although promising as a random sampling methodology, this approach has a few problems. For example, some Web sites are duplicated and posted under more than one IP address, which means that these sites have a greater chance of being selected due to overrepresentation in the IP address pool. It is also difficult to sample individual Web pages since multiple pages can exist under the same IP address. With the use of random IP address sampling methodology, Web pages can only be sampled indirectly through cluster sampling of entire Web sites containing multiple pages.

The random walk is another approach that has been discussed theoretically (Deo & Gupta, 2001) and used in research involving Web pages (Henzinger, Heydon, Mitzenmacher, & Najork, 1999). In a random walk, a beginning point is selected at random and then successive steps are taken in random directions. To sample the Web, this could be applied by sampling one page at random and then additional pages would be sampled through a random walk strategy. Software similar to a search engine crawler could walk from link to link in this random walk to sample pages. Eventually, it should be possible to randomly walk all around the Web to generate a random sample. This approach allows sampling of individual Web pages unlike the random IP address sampling procedure.

There are some problems with random walks on the Web, however. One of the problems is the difficulty in selecting the first page entirely at random. There is no comprehensive and up-to-date list of all available Web pages, so it is not feasible to randomly select the starting page to begin the random walk. Consequently, there will be some unavoidable bias introduced in the selection of a starting point. There is also the potential problem of becoming trapped in a random walk around a single Web site that is either huge or has few if any links leading out of the site.

This can potentially lead to the selection of a biased sample of multiple pages from the same site.

Random IP address sampling and random walk strategies both yield samples of sites from all over the Internet. These strategies are valuable when the goal of the research study is to reveal characteristics of Web pages from the full spectrum of possible sites on the Web. In this type of study the content can be on any topic. Unfortunately, random sampling across the entire landscape of the World Wide Web cannot be effectively used to draw samples of Web pages that are all on the same topic. For example, this sampling strategy would be problematic for a researcher who is interested in learning more about the quality of instructional content for mathematics that K-12 students have access to on the Web. This line of research is particularly relevant because more and more students attend schools with Internet access. At the present time, nearly 100 percent of all public schools in the United States are connected to the Internet (Parsad & Jones, 2005). If teachers and students are using the Web to locate information for classroom use, then the question of quality naturally arises. The researcher could conduct a study using random sampling methods that cover the entire Web, but the pages or sites located this way will be a collection of every possible type of content. The sample would need to be filtered to pull out pages meeting research criteria while discarding the remaining unwanted pages. A sampling methodology that is restricted in order to target the specific type of content most applicable to this research would be a valuable, time-saving alternative.

In some cases a comprehensive list of Web sites for a related topic may be available. Random samples can be drawn from such a list to provide a selection that is constrained to a desired topic and also generalizes to the larger population of similar Web pages. This approach was used in a study of FM radio home pages (Potter, 2002). The sample was selected from a comprehensive list of radio station sites through a systematic process that began with a random starting point. A

variation of this list sampling strategy was used in a study designed to explore the accessibility of Web sites over time (Hackett, Parmanto, & Xiaoming, 2004). In this study, a random sample was drawn from a set of lists of top-ranked Web sites in various categories that were compiled by several online directory services. Archived copies of these sites were then studied to reveal trends in accessibility over time. The list of sites used for this study was not a complete list of related sites as in the aforementioned FM radio study. Unfortunately, the majority of content found on the Web is not organized into comprehensive and up-to-date lists. The possibility for randomly sampling from a published list is not realistic for many studies of Web content geared toward exploring specific types of content.

Purposeful Sampling in Web Research

Purposeful sampling is used in qualitative research to select a small sample based on specific research criteria. The purposeful selection of a small sample is undertaken to obtain a set of information rich cases for in-depth analysis (Patton, 1990). Purposeful sampling may be used in studies of Web page content for similar reasons. A set of Web pages may be purposefully selected based on content constraints defined in a study. The sample size may also be limited to keep the study manageable and feasible. The actual procedure that has been used for sampling in many studies of Web-based content involved the use of one or more online search tools. For example, a combination of online portals and search engines was used in an evaluation study of Web pages on Canadian history (Bowler, Nasset, Large, & Beheshti, 2004). The portals were searched to quickly locate links to relevant Web sites that had already been identified and categorized within the portals. The Internet search engines were used to verify search terms and explore the types of sites and information that were revealed through this alternative process. Sites that were selected for the study were evaluated on several factors related to information quality and reliability. The purposeful selection of Canadian history sites allowed the researchers to obtain a

sample of pages containing information that was highly relevant for the study. It would have been unfeasible to use random sampling techniques in this particular study due to the absence of a comprehensive list and the need to constrain the content to a single topic.

The strategy of purposeful sampling through the use of search engines has been used in several studies of health information found on the Web. For example, Murphy, Frost, Webster, and Schmidt (2004) conducted an evaluation of information on the treatment of eating disorders. The sample for this study was selected through the use of several search engines in order to locate a set of 15 sites that would represent what the typical user would find online. In another study on the quality of online health information about sore throats, samples of 150 sites were selected through the use of a meta-search engine (Curro et al., 2004). With a meta-search engine it is possible to conduct a parallel search on several search engines simultaneously and generate a set of combined results. Some researchers choose to use a single search engine to obtain a sample as was done in a study of cystic fibrosis information found on the Web (Anselmo, Lash, Stieb, & Haver, 2004). In this study the top 100 sites from the search engine Google were selected as the sample. It seems reasonable to do this since people have reported using search engines to locate health information on the Web. Results from survey research have indicated this is often the case for college students seeking health information (Escoffery et al., 2005). This finding is supported by information found in the literature about how people locate and use health information online (Morahan-Martin, 2004).

When considering the use of search engines such as Google for Web page sampling, one must remember that search engines do not search the Web directly. Instead, they search an index of sites created by robot software that crawls or spiders the Web. Over time, the index used by a search engine is expanded and updated, but it is never completely up to date or comprehensive. It takes time for

search engine crawlers or spiders to cover the vast number of sites on the Web; so when new content is added, it may not be added to the searchable index right away. Similarly, when content is removed from the Web, it may take awhile to update the index accordingly. Other sites may not be indexed because of access restrictions or technical constraints. Sites that cannot be indexed have been referred to as belonging to the *invisible Web* since they are invisible to search engine spiders (Descy, 2004; Sherman & Price, 2001). The *indexable Web* is that part of the Web containing sites that may be added to the index of at least one of the major search engines. None of the major search engines contains a full listing of the entire indexable Web, and the amount of coverage varies between them. Estimates of coverage for several search engines have been reported as part of a study of the overall size of the indexable Web (Gulli & Signorini, 2005). The results of this research indicate that the search engine having the largest index with the greatest coverage of the indexable Web is Google (76.16%), followed in descending order by Yahoo! (69.32%), Ask/Teoma (57.62%), and MSN Beta (61.90%). Despite the limitation of incomplete indexes, search engines provide a relatively simple way to obtain a purposeful selection of Web sites matching specific content criteria.

Purposeful Random Sampling Through Search Engines

Although search engines are useful tools for extracting specific information from the Web, the number of sites returned from a search often exceeds reasonable feasibility for a research study. For example, a search on Google for the word *algebra* yields an estimated result count of over 37 million pages. This number of results is far too large for many research studies. This is particularly true for any study involving qualitative analysis. A more specific search phrase such as *linear algebra* will typically reduce the total number of results. This works reasonably well if the desired content is really linear algebra rather than any type of algebra. Either way, since Google (n.d.-a) automatically caps the maximum

number of search results it returns to the top 1,000, a limit is set anyway. Consequently, search results from Google can yield anywhere from zero to 1,000 links to sites on the Web. These results become the accessible set of links to sites for any given search phrase.

Random sampling from within search engine results is a type of purposeful random sampling. Search results are selected purposefully through a query to a search engine. Results are returned as a list ranked for relevance through the application of a mathematical algorithm, which makes them a type of ordinal data. A random sample may be chosen from the set of ranked results to obtain a small set of sites that match research criteria for content. This approach was used in a study of Web pages designed for online mathematics instruction (Snelson, 2002). A mixed-method approach involving quantitative and qualitative methodologies was used in this study. A purposeful random sample was selected to make the study manageable during the process of in-depth qualitative analysis.

Purposeful random sampling provides a way to obtain rich information about content quality from the larger group of search engine results. Instead of simply skimming off the top 100 hits, it is possible to sample randomly from the first 1,000 to draw from a wider range of relevant sites. A random sample from the entire set of search results will allow exploration of all sites made available through a specific search query. The suitability of this approach will be dependent on the nature of the research question. If the researcher is really interested in the content a typical user will most likely access, then the top 100 or fewer results may be more appropriate. The random sample could be drawn from the top 100 sites if desired. Alternatively, the question may be geared toward analysis of any content made available through a search for Web pages related to a specific topic. For example, the focus of a study may be to compare the quality of content that has been assigned high or low ranking in a search results list. The issue of ranking has increased in importance due to the massive amount of information on the Web

(Zhao, 2004). The sites that rise to the top of the list are the most likely to be viewed. Web optimization strategies are now marketed as a way to provide advantages to sites so that they rank higher in the search results list. A random sample drawn from the entire set of search results yields a range of optimized and non-optimized content for comparison. In this case the purposeful random sample from the entire set of results may be more beneficial.

Web Services

One of the drawbacks of using purposeful random sampling is that search engines are not designed specifically for the task. Search results are printed in linear order on a Web page. Typically, they are divided over multiple pages. Google search results are printed ten per page by default although references can be set to display up to 100 per page if desired. Even with the 100-results-per-page option, the results are divided across as many as ten pages. For the typical user, the results are output in a useful form. It is not as useful for the researcher who may be interested in quickly downloading a set of search results to load into a spreadsheet or database. The links and other site information yielded by a standard search engine must either be manually copied into a database or somehow transformed into a structure such as delimited text that can be imported. If a purposeful random sample is desired, then all of the links can be copied into a spreadsheet or database, and a random number generator can be used to select the sites for the study. Those sites can then be copied from the spreadsheet or database into a new file to be used in the research study. Unfortunately, this is a tedious process that can consume hours of a researcher's valuable time. One way to solve the problem is to build a custom application to automate the sampling task.

Custom automated sampling tools for searching the Web can be developed through the use of one of the Web services currently available. Web services are

designed to be accessed across the Web through the use of software or some other type of program developed for this purpose (Cerami, 2002). The Google Web APIs service, for example, has been established to allow programmers to develop search tools to meet their individual needs (Calishain & Dornfest, 2005; Google, n.d.-b; Mueller, 2004). Through this service, the user may access Google directly without going through the standard search engine page established by the company. It is possible through this service to develop highly specialized search tools that save time and tap into such existing resources as the massive Google index.

The Google Web APIs service is free for use in strictly non-commercial applications (Google, n.d.-b). To use the service, an account simply needs to be created to obtain a license key. Despite the designation as an experimental or beta service, it is valuable as a tool for experimental development of new search and sampling tools for academic research studies of Web page content. It is designed to work through the use of Simple Object Access Protocol (SOAP) and the Web Service Description Language (WSDL), which are both Extensible Markup Language (XML) based. XML was designed to serve as a universal language for data sharing between diverse computer systems and programs. It is possible, therefore, to write applications for the Google Web APIs service using a variety of programming languages and computer operating systems. This allows great flexibility in development of custom search tools for research. The major limitation is due to search restrictions set in the terms of service. Currently, there is a limit of 1,000 queries per day, and only ten search results (links to sites) may be obtained in a single query. In addition, access is restricted to the top 1,000 results just as it is for standard searches using the Google search engine. Customized search tools using the Google Web APIs must function within these constraints.

Development of Web Sampler: A Custom Search Tool Using the Google APIs Service

Web Sampler is a custom search tool that was developed using the Google Web APIs service. A screen capture of the interface is shown in Figure 1.

Web Sampler
Generate and Download a Random Sample of URLs

1 Limit sampling range to somewhere between top 100 and top 1,000 Google search results.
☒ 100 ☐ 200 ☐ 300 ☐ 400 ☐ 500 ☐ 600 ☐ 700 ☐ 800 ☐ 900 ☐ 1,000

2 Select a sample size.
☒ 1 ☐ 2 ☐ 5 ☐ 10 ☐ 15 ☐ 20 ☐ 25 ☐ 50

3 Type a search phrase.

4 Wait for the results.

5 Download the delimited text file at the bottom.

Figure 1: The Web Sampler Interface

When a search is conducted using Web Sampler, a random sample of URLs (links to sites on the Web) may be obtained. Sample size may be selected by clicking the radio button next to the desired number. The user also has the option of limiting the range of sampled search results from the top 100 all the way up to the maximum of 1,000. This feature was added to provide sampling flexibility and also to handle situations where searches yield fewer than the 1,000 results. Once a range restriction and sample size have been selected, a search query may be typed into the text box and the form submitted.

Web Sampler was written using Hypertext Pre Processor (PHP) and Hypertext Markup Language (HTML) code. Both types of code may be written using a plain text editor or Web page authoring software. PHP is active script that runs on a Web server and provides the basic engine that runs the Web Sampler search tool.

HTML was used to create the form elements for the Web Sampler interface. PHP code processes the form data after submission and makes the call to the Google

Web APIs service. A random number generator was added to the PHP code in order to select and return a random sample from the search results.

Queries to the Google Web APIs service are sent as SOAP messages across the Web using the WSDL that has been developed specifically for the service. PHP can be used to write SOAP clients if the appropriate extension is turned on during installation on the server. This becomes problematic if the developer does not have direct access to the PHP installation on the server where the application pages are stored. A viable alternative is to use the NuSOAP Tool Kit for PHP, which is a set of classes developed to process SOAP messages (SourceForge.net, 2005). NuSOAP code can be copied onto the server along with the pages containing the associated PHP SOAP application. It is independent of the PHP installation and provides greater access and flexibility when developing code. Because of this, NuSOAP was used as part of the Web Sampler project.

A diagram showing the process used by Web Sampler to send search queries and receive results from the Google Web APIs service is displayed in Figure 2. First, user input is entered into the form. After the form has been submitted, the PHP processor is activated. A set of random numbers is generated and a series of calls are made to Google as the PHP processor loops through the random numbers. Search results corresponding to the random numbers are returned and printed onto the search page. They are also printed into a delimited text file for download.

*S = Send Query, R = Return Result, P = Print Result

Figure 2: Process Used by Web Sampler to Obtain Random Search Results

NOTE: Soap above should be changed to SOAP

The results that are returned from Web Sampler are printed on the screen in a manner similar to what is shown in Figure 3. Each search result is numbered so that the user can be certain that the selected sample size was retrieved. The Google result number shows the position of each randomly selected search result in the complete list of results that were returned for the query. Active links allow the pages to be viewed individually. At the end of the search results list, there is a link to download the results in a delimited text file. This file can be imported into spreadsheet or database software for further analysis.

Figure 3: Results Obtained from Web Sampler

The Web Sampler search application was tested on both a desktop computer and a university Web server. The desktop computer was a Windows XP machine with the IIS server component installed. PHP was also installed on the machine. A high speed Internet connection allowed search queries to be sent and results to be received by Web Sampler on the desktop machine. The university Web server used for testing was a Windows 2003 system with PHP installed. Web Sampler ran smoothly on both systems during all testing phases. The ability to run the search tool on either a desktop machine or a Web server allows flexibility in the selection of hosting options.

Summary and Conclusions

Research studies designed to investigate the characteristics, size, or quality of content on the Web have been associated with numerous challenges. The selection of a representative sample of pages or sites from the Web has proven to be one of the more daunting of these challenges. In Web research sampling, methodologies are typically chosen from among one of three broad categories. A representative sample with the highest degree of generalizability may be chosen from across the landscape of the public Web, but this is not suitable for studies focused on evaluation of a specific type of content. Comprehensive lists are

seldom available for sampling. As a result, many researchers have turned to purposeful sampling through the use of search engines or meta-search engines. This has led to the use of a sampling methodology in which the top hits from search results are selected. In some studies it may be useful to randomly sample from the full array of search results.

Web Sampler was developed as a custom search tool that enables purposeful random sampling from the Google index. The disadvantages associated with this are due to the constraints of using the service, including limitations on the number of search queries that may be obtained per day. Constraints on the depth of sampling to the top 1,000 results are also restrictive for sampling across the entire set of matching results. This may be a minimal limitation if the top 1,000 results are truly the most representative sites in the index. The advantages of developing and using a tool such as Web Sampler lie in the minimal cost of development, free access to an existing index of billions of pages, potential development of custom features for Web research, automation of time-consuming and tedious tasks, and the possibility of hosting the search tool on either a desktop machine or a Web server.

An ideal search service for academic Web research would provide access to an index of catalogued sites that is comprehensive and updated continually. The addition of advanced search and automated analysis tools would provide a way for researchers to answer questions about information quality on the Web. Unfortunately, such a service would require considerable resources to develop and may be cost prohibitive for both developers and users. At the present time, tools such as Web Sampler provide a way to develop inexpensive custom search applications for research. More research is needed to identify and test appropriate sampling methodologies and develop appropriate software applications for Web research. The importance of this line of research is directly proportional to the increasing role of the Web as an information resource.

References

- Anselmo, M. A., Lash, K. M., Stieb, E. S., & Haver, K. E. (2004). Cystic fibrosis on the Internet: A survey of site adherence to AMA guidelines. *Pediatrics*, 114(1), 100-103.
- Berners-Lee, T. (1989, 1990). *Information management: A proposal*. Retrieved May 18, 2005, from <http://www.w3.org/History/1989/proposal.html>
- Bowler, L., Nasset, V., Large, A., & Beheshti, J. (2004). Using the Web for Canadian history projects: What will children find? *Canadian Journal of Information & Library Sciences*, 28(3), 3-24.
- Calishain, T., & Dornfest, R. (2005). *Google hacks: Tips & tools for smarter searching* (2nd ed.). Sebastopol, CA: O'Reilly, Inc.
- Cerami, E. (2002). *Web services essentials*. Sebastopol, CA: O'Reilly.
- Curro, V., Buonomo, P. S., Onesimo, R., De Rose, P., Vituzzi, A., Di Tanna, G. L., et al. (2004). A quality evaluation methodology of health web-pages for non-professionals. *Medical Informatics & the Internet in Medicine*, 29(2), 95-107.
- Deo, N., & Gupta, P. (2001, February 26-March 2). *Sampling the Web graph with random walks*. Paper presented at the 32nd Southeastern International Conference on Combinatorics, Graph Theory and Computing, Baton Rouge, LA.
- Descy, D. E. (2004). Searching the Web: From the visible to the invisible. *TechTrends: Linking Research & Practice to Improve Learning*, 48(1), 5-6.
- Escoffery, C., Miner, K. R., Adame, D. D., Butler, S., McCormick, L., & Mendell, E. (2005). Internet use for health information among college students. *Journal of American College Health*, 53(4), 183-199.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

Google. (n.d.-a). *Google help center: How can I see more than the first 1000 results?* Retrieved November 20, 2005, from <http://www.google.com/support/bin/answer.py?answer=484&topic=359>

Google. (n.d.-b). *Google Web APIs (beta): Develop your own applications using Google*. Retrieved November 21, 2005, from <http://www.google.com/apis>

Gulli, A., & Signorini, A. (2005, May). *The indexable Web is more than 11.5 billion pages*. Paper presented at the 14th International World Wide Web Conference, Chiba, Japan.

Hackett, S., Parmanto, B., & Xiaoming, Z. (2004, October 18-20). *Accessibility of Internet websites through time*. Paper presented at the ACM SIGACCESS Conference on Assistive Technologies, Atlanta, GA.

Henzinger, M. R., Heydon, A., Mitzenmacher, M., & Najork, M. (1999, May). *Measuring index quality using random walks on the Web*. Paper presented at the 8th International World Wide Web Conference, Toronto, Canada.

Henzinger, M. R., & Lawrence, S. (2004). Extracting knowledge from the World Wide Web. *Proceedings of the National Academy of Sciences of the United States of America*, 101(1), 5186-5191.

Morahan-Martin, J. M. (2004). How Internet users find, evaluate, and use online health information: A cross-cultural review. *CyberPsychology & Behavior*, 7(5), 497-510.

Mueller, J. P. (2004). *Mining Google web services: Building applications with the Google API*. San Francisco: Sybex.

Murphy, R., Frost, S., Webster, P., & Schmidt, U. (2004). An evaluation of web-based information. *International Journal of Eating Disorders*, 35(2), 145-154.

O'Neill, E. T., Lavoie, B. F., & McClain, P. (1999). *Web characterization project: An analysis of metadata usage on the web*. Retrieved May 18, 2005, from <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003486>

O'Neill, E. T., McClain, P. D., & Lavoie, B. F. (1997). *A methodology for sampling the World Wide Web*. Retrieved May 18, 2005, from <http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003447>

Parsad, B., & Jones, J. (2005). *Internet access in U.S. public schools and classrooms: 1994-2003* (No. NCES 2005-015). Washington, D.C.: U.S. Department of Education, National Center for Education Statistics.

Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd ed.). Newbury Park, CA: SAGE Publications.

Potter, R. F. (2002). Give the people what they want: A content analysis of FM radio station home pages. *Journal of Broadcasting & Electronic Media*, 46(3), 369-384.

Sherman, C., & Price, G. (2001). *The invisible web: Uncovering information sources search engines can't see*. Medford, NJ: Information Today, Inc.

Snelson, C. (2002). Online mathematics instruction: An analysis of content. (ERIC No. ED470536), 1-12.

SourceForge.net. (2005). *NuSOAP - SOAP toolkit for PHP*. Retrieved November, 15, 2005, from <http://sourceforge.net/projects/nussoap/>

Zhao, L. (2004). Jump higher: Analyzing Web-site rank in Google. *Information Technology & Libraries*, 23(3), 108-118.