Where Is the Information Located?
Navigating the Invisible Web

by

Jane Long

School of Library and Information Studies
University of Oklahoma
January 27, 2006
Where Is the Information Located? Navigating the Invisible Web

A major issue regarding the use of the Internet involves the importance seekers place upon finding resources on the Web and accomplishing their searches as quickly as possible. For these reasons, users the world over want to see “more integrated, homogeneous, and thus simplified” material being made available through on-line resources (Webster, 2004, 30). Ease of use and links to a variety of materials are both essential factors for information seekers. As a result, tremendous efforts are being made to mine the material that is not available through public search engines. This information is located in an area that is called the invisible, hidden, or deep Web. These terms have been coined to describe the resources that are not accessible to users of the Internet through public search engines. In order to locate this data, a searcher must understand some of the facts about the hidden Web such as what it is and why it can be difficult to access. With an understanding of what is located in the deep Web, how the resources are indexed, and how to retrieve data, users will benefit from the wealth of information that is located in this mysterious part of the World Wide Web.

It was in 1995 when Internet use became popular, but in its early days, there were few search tools that could be used to explore what was available on the World Wide Web. The two tools that were most frequently used then were Yahoo! and Magellan. Interestingly enough, Magellan is no longer in existence. These tools were actually subject directories that could take an information seeker to an appropriate location when the individual typed a keyword into a search box. Because of the way they were used, the tools were renamed search engines. A short time later, WebCrawler and Excite created indexing programs that became known as spiders. A spider’s task was to search for an information seeker’s request. Typing key words in the search box produced hits that took users to information related to their search requests (Schrock, 2003,
Although the process has been streamlined over time by the introduction of new and improved search engines such as Google, AltaVista, MSN, and Lycos, the desired outcome is still the same. Peter Webster, a systems librarian at Saint Mary’s University in Halifax, Nova Scotia, reports that “Information seekers . . . are seeking the simplest and most direct method for searching the widest possible body of material to find relevant, good quality content” (2004, 30). Preeminent authorities on the subject of the invisible Web, Chris Sherman and Gary Price, define it as “. . . content [that] is not easily located with information-seeking tools used by most Web users, it’s effectively invisible because it’s so difficult to find unless you know exactly where to look” (2003, 282). Sherman and Price explain the definition more specifically by dividing it into two parts. First, they say, there are technical reasons why search engines are unable to find information on the Web. The search engine spiders can only link to information that has been indexed, so if the data is not included in a list of resources for the Web site, it cannot be accessed by a spider. Most of this material is located in databases that are not available to the general public. The second reason Sherman and Price give involves indexing of nontext files. They say, “Pages that consist primarily of images, audio, or video, with little or no text, make up another type of Invisible Web content.” This means formats such as Postscript, Flash, Shockwave, gif, jpeg, or other graphics are not identifiable by search engines (2003, 284-85). It is easy to see from this definition that a great deal of information is hidden from users, so as a result, it becomes part of the deep Web. In fact, a leading company in deep Web research and development called BrightPlanet says that approximately 550 billion documents are not available to the public through standard search engines (Lin and Chen, 2002). Thomas Pack, in writing about BrightPlanet for Link-Up, says, “‘The deep Web is content that resides in searchable databases, the results from which can only be discovered by a direct query’” (2002, 18). A
document called “White Paper: The Deep Web: Surfing Hidden Value,” by Michael K. Bergman compares searching the Web to dragging a net across the surface of the ocean. The net is unable to capture what lies below the surface; therefore, most of the documents that are below the surface are not accessed (2001). Additionally, information on BrightPlanet’s Web site states that the deep Web is approximately 500 times larger than the surface Web. BrightPlanet also states that 95% of the material that is located in the deep Web does not require a subscription or fee in order to be used. This information indicates that users are losing out if they do not learn how to search the invisible Web in order to use these valuable resources effectively.

The general term “invisible Web” sounds mysterious and almost frightening; however, it is just an identifying name for a portion of the World Wide Web. This part of the Web can only be accessed in a slightly more complicated way than can the traditional Web. Information found on the deep Web is either not indexed by standard search engines, or it is not created in a standard format that can be indexed. Price and Sherman explain that the invisible Web is comprised of four distinct types of materials. Examining these four types provides some understanding of the larger deep Web. The four areas include the categories of opaque, private, proprietary, and truly invisible. What is interesting about the first area, the opaque Web, is the fact that it includes files that could be crawled, but not all of them are indexed. Reasons such as the expense of indexing, constant change of what is on the Web, the continual addition of new sites, restriction of the number of results accessed, and disconnected urls are just a few explanations for why some of the materials located on the Web are considered to be opaque (2003, 293-95). As the term indicates, the resources are there, but they are not in plain view, and spiders cannot find them. A second area discussed by Sherman and Price is the private Web. As the title implies, these pages “have deliberately been excluded from inclusion in search engines.”
They are password protected, and material indexed is available for use by members of the owner organisation. It is used by those who need the information located there to accomplish specific objectives for the organization (2000, 296). A third area of consideration is the proprietary Web. Although somewhat like the private Web, these pages can only be accessed by those who have agreed to special terms for use. Those who access these resources must register, and oftentimes they must pay a fee in order to use the pages. The proprietary Web is similar to databases that are available to university students or organizations such as medical or law libraries, but according to Sherman and Price, these types of databases provided information to users before databases were available on the Internet. As a result, they fit into a different category (2000, 296-97). The final area includes the portion of the Web that is truly invisible. It cannot be spidered or indexed for a number of reasons including file format, basic search engine omission, problems with the search engine, and storage in relational database management systems that require a very specific query in order to access the information (2000, 297-98). Awareness of these complexities helps a user understand what is available and what is not. Although exploration of the hidden Web may be somewhat confusing at first, the quest to obtain quality information from new sources will be worth any difficulties that are encountered. Once the process of accessing the hidden Web is learned, the rewards of the experience will be invaluable.

As was pointed out by Sherman and Price, indexing of the invisible Web has not been adequate. Currently, more work is being done to provide indexing that will aid users in finding resources that should be available on the hidden Web. According to Yanbo Ru and Ellis Horowitz, indexing of the invisible Web is necessary in order for the information to become more easily available for a greater number of users. The necessary methods for effective use of this material are complicated, so many challenges are present in making the material available.
Ru and Horowitz state that in order to index the hidden Web, research must be done to determine the interfaces in addition to classifying the sites that are invisible (2005, 250-251). Work with methods of indexing will aid users in circumventing some of the complexities of accessing the sites. Through the findings of their study, Ru and Horowitz write, “there may be content within the Web site that users are unable to discover, even though the directory points them to the interface.” The indexing directories that have been created are a great aid for users, but more work must be completed that will continue to improve these indexing resources. Ru and Horowitz end their findings by saying, “A technique that can more comprehensively index the data in an invisible Web site, and that will not get swamped by the size of the data, is required” (2005, 261-262).

For information users and knowledge managers, finding information is one of the greatest pleasures as well as one of the greatest challenges that can be encountered. Over the past ten years, improvements have occurred regarding the structure and availability of materials in today’s current environment; however, the deep Web continues to be unavailable to users through general search engines. Individuals such as Chris Sherman and Gary Price have worked diligently to make the resources of the invisible Web available to more users. Both of these men have written extensively about the subject and have developed directories for available resources. Another individual who has inspired others to mine this valuable resource is Diana Botluk, a law librarian at Catholic University of America. Botluk has written an excellent article in which she explains how to use the hidden Web with efficiency. A primary purpose of Web access to resources, she says, is to make searching for information and finding the information easy. She also reminds users that search engines do not understand a request; they only attempt to match up the words that are typed into the search box. Botluk makes a valuable comment
when she says, “responsibility for the search is on the researcher herself. If she understands the research process, then unlocking the information found in deep Web databases becomes easy.” With this idea in mind, Botluk adds that it is essential for a searcher to consider the process that must be developed in order to locate information on the deep Web. She explains that the process is not difficult, for it involves only two steps. The first of the actions that a searcher must take is to determine where to look for the needed information. Botluk believes that a general search engine can often be used to find an appropriate hidden Web database (2000). Although this sounds simple, a searcher must then learn how to find the databases that are available on the deep Web before being able to use them. As part of the search process, the use of a directory helps a searcher find an appropriate database.

Fortunately, a number of directories to the invisible Web have been created that simplify searching. Some of the directory sources that are currently available according to an index compiled by Yanbo Ru and Ellis Horowitz include Beaucoup.com that accesses a variety of information databases; BrightPlanet’s CompletePlanet that includes links to over 70,000 searchable databases; Gary Price’s Direct Search that provides news for information professionals; ProFusion that arranges information by topic; Science.gov that links to First Gov’s science resources; Search Engine Colossus that is an international directory; and Search Bug that is a people and company finder. This list of directories is just a sample of the resources that are currently available for locating deep Web information (2005, 265). As the invisible Web continues to grow, a number of the original directories have been changed or eliminated while a variety of new resources have been added (Smith, 2001, 17).

Diana Botluk says the second part of the search should continue by narrowing the terms to focus upon the needed information. This part of the search should parallel a database search
conducted from an open Web location. At this point, the user should be successful in finding the information for which the search has taken place. In discussing the completion of a search, Botluk writes, “The point is that often the key to the answer is not locating the answer itself as the first step, but locating the right database in which to search for it” (2000). Creativity and willingness to search for the best place to find the information are keys to successful use of the deep Web. The steps suggested by Botluk apply to all categories of users, both information specialists and library patrons. When those who are seeking information follow these simple steps, searchers should be able to find quality information and feel confident about using these “invisible” resources.

It is very important for information professionals to understand the value of the hidden Web in order to access the extensive materials that are available through its use. King-Ip Lin and Hui Chen point out that much of the data that is part of the deep Web is located in databases with specific domains (2000). BrightPlanet, as well, explains that at least half of the material located on the hidden Web is found in “topic specific databases.” Additionally, BrightPlanet states that a major percentage of the information found on the hidden Web does not reside on the private Web. This means it is available to all users without subscriptions or fees. This is another positive factor for invisible Web users. Credible sources of information are available to users at no cost. The opportunity to utilize these vast resources makes it essential for librarians and other information specialists to navigate successfully this portion of the Web. With practice, it is not difficult to use databases, so by applying basic knowledge regarding database search techniques and use of the simple steps explained by Botluk, information professionals should be able to access this portion of the Web and share quality information with their patrons.
Of course, learning about the deep Web’s resources presents a number of challenges for the information professional. Use of the deep Web involves implementing knowledge regarding how to access this vast amount of information efficiently as well as learning how to share this knowledge with patrons. In an article published in *The Journal of Academic Librarianship*, Jane Devine and Francine Egger-Sider emphasize the importance of using the invisible Web as a tool for promoting the value of librarianship as well as providing new avenues for finding research materials. Devine and Egger-Sider advocate using the invisible Web as a way to promote the library profession. They write, “. . . introducing the concepts of the Invisible Web to library patrons can promote appreciation for librarians’ work and library services” (2004, 268). Suggestions made by the authors show effective ways to use the hidden Web’s resources positively in academic settings.

A priority for librarians is to understand how to use the invisible Web; without knowing how to access this information themselves, they cannot assist patrons in using this valuable resource. Additionally, the invisible Web should be used by librarians to conduct research for faculty members or administrators. Utilization of the hidden Web for research is a way to share its value with others. Also, the value of the hidden Web can be shared with students who are conducting research because it is a resource that makes available new ways to access quality information. An excellent point is made when Devine and Egger-Sider state, “Even the simple suggestion that a whole world of research exists beyond the reach of Google will broaden student horizons” (2004, 268). Use of the invisible Web by students will aid them in seeking out credible Web sites and getting away from overusing the dot coms that can be accessed too easily.

Some other ways to encourage use of the invisible Web, say Devine and Egger-Sider, are conducting workshops that explain how to gain access to the invisible Web by using tools
beyond the basic search engines. Advertising information about the deep Web through links and guides on the library’s Web site will increase communication provided to library patrons. This should generate interest in these resources that will aid in searching for information in the library catalog and finding materials needed in the classroom setting. Writing about the invisible Web in university newsletters is another way to share information about its value as a search tool. In addition, these articles will remind the academic community that librarians provide a valuable service (2004, 268-269). Through these various ways of sharing information, the librarians’ roles as information professionals who teach and assist patrons will be reinforced.

In addition to these ways to address the significance of the invisible Web to information users, it is worthwhile to know that along with other changes regarding this area of the Web, some states are establishing their own invisible Web programs that serve as a means to establish communication among libraries within their state and provide linking services that will be available statewide. A metasearch application called Library of Texas (LOT) has been developed in Texas. This statewide virtual library enables access by members to library catalogs from public and academic libraries across the state. Additionally, a core of databases is available to users through the state library system as well as the state’s archives commission (2004, 22). LOT is an excellent example of one of the valuable capabilities of the hidden Web. Kathleen R. Murray, president of KRM Consulting, writes, “Metasearch applications . . . provide new visibility to deep Web resources not available through general-purpose Web search engines. The technical infrastructure of the LOT provides a solid foundation for future growth” (2004, 24). This type of growth includes new and improved opportunities for information seekers and information professionals.
Peter Webster writes that many librarians as well as their patrons believe that everything is available on the Web. If so much information is available, the big question facing the library profession is how can this information be retrieved easily and quickly? By learning what is accessible through the resources of the hidden Web, Webster believes the search for outstanding information may become one single simple search operation (2004, 34). Karen R. Diaz who is the Education Librarian for Internet Instruction at Ohio State University writes, “The more we learn about the compartments on the Web . . ., the less invisible the resources become. . .” (2000, 134). This is excellent guidance for information professionals to apply, for with this idea in mind, satisfaction should be the ultimate result.

Many changes have taken place in just the past few years regarding how to discover new resources on the Web. As a result, information professionals are learning to eliminate the mystery surrounding the hidden, deep, or invisible Web, and they must demonstrate to users that the invisible Web is a valuable tool for finding credible information. Areas such as improving hidden Web indexing, learning successful search techniques, and advertising the search capabilities that are available by using the hidden Web’s immense body of information are ways that this tremendous resource can be used positively by information seekers. It is the responsibility of trained librarians to help others take advantage of the vast and valuable resource known as the invisible Web.
Reference List


