INTRODUCTION

The Invisible Web presents many problems for the information world, but the essential one for educators and librarians is how to convey adequately its substance and importance. The Invisible Web represents the largest sector of online information resources on the Internet, and yet the first obstacle in discussing it is one of nomenclature. How substantive, after all, is anything labeled “invisible”? The very name implies obscurity and marginalization, which is why some prefer to use the term “deep Web” or “hidden Web” to refer to these resources. No matter how one refers to it, however, the Invisible Web is a phenomenon to be reckoned with.

Simply put, the Invisible Web is the term used to describe all of the information available on the World Wide Web that cannot be found by using general-purpose search engines (Devine and Egger-Sider, 2001). The Invisible Web has various characteristics that can help people understand it. This chapter demonstrates the importance of the Invisible Web by analyzing these characteristics.

Discussion of the Invisible Web should be part of any information literacy program. Knowledge about the Invisible Web should be part of everyone’s information and research skills education. Study of the Invisible Web promotes discussions about resource evaluation and how to apply critical thinking to research. It is possible that, if educators present the Invisible Web with the right approach and a clear definition, the very mystique that the name implies will interest and intrigue students. Perhaps educators can even involve young minds in solving some of the technical problems that create the Invisible Web.

Prevailing notions about Web searching, known as Web myths, are worth exploring toward opening a discussion on how to maximize Web research by
incorporating Invisible Web resources into the search process. The following are false impressions about the Web and the more complex realities that exist:

- **Everything worth finding is already on the Web, or if it can’t be found on the Web, it is not worth finding.** Who says so? Why does Web information seem better than information from other resources? The only certainty is that the former can be easier to access. However, the world of information still includes print sources, many other formats, (AV, CDs), etc., and people with expertise.

- **Google searches the whole Web.** The overwhelming number of results that users call up with their queries probably creates this mistaken impression. Studies have shown the Invisible Web to be about 500 times the size of the visible Web accessed by general-purpose search engines such as Google (Bergman, 2001).

- **The best information is found in the first ten results.** Search engines use formulas or algorithms to rank search results and present them in the order that the user sees them, a process known as relevancy ranking. This ranking may be based on how often a keyword or phrase appears on a Web page or on more sophisticated methods of evaluating linking, as is the case for Google. However the ranking is done, there is no guarantee that the first ten results are the best for the researcher’s purpose; only the user can judge. To further complicate the issue, Web designers can create Web sites that will ensure placement in the first ten results. Companies known as search engine optimizers offer this service to businesses seeking good placement.

- **Searching is easy.** A saying in the library world acknowledges that “searching is easy; finding is more difficult.” The Web has certainly made searching for information more accessible, but “easy” may depend on the expertise of the researcher and how far he or she is willing to take a search. Many use “easy” to imply speed as well, but successful research does require time for evaluation and reflection.

- **Everything important is free.** Only in a noncommercial world could one hope that everything important might be free. Information is a commodity and, as such, is expected to generate some profit. People are best served by being good information consumers, and when it is necessary to pay for the information, they should know what they are paying for.

- **Everything is truthful, authoritative, and accurate.** Expecting everything on the Web to be authoritative, accurate, and truthful is another ideal scenario that does not exist in the real world. Even with the best intentions of the providers, some information may be tainted with bias, opinions, and inaccuracies. Students and other researchers need to develop evaluation skills to cope with the Web environment.

Similar lists of Web myths have been compiled by Mark Jordan, “Ten Myths of the Internet” (Jordan, 1997), and Mona Kratzert and Debora Richey (1997), “Ten Internet Myths.”
RELATIONSHIP BETWEEN SEARCH ENGINES AND THE INVISIBLE WEB

If the Invisible Web comprises all information available on the World Wide Web that cannot be found by using general-purpose search engines, then clearly the existence of the Invisible Web bears a direct relationship to general-purpose search engines. We can even postulate that every search engine creates its own Invisible Web, composed of all the resources that it fails to find or “decides” that it will exclude from its indexing. Understanding the Invisible Web therefore begins with a better understanding of how general-purpose search engines work.

In truth, people do not really need to understand the technical aspects of how search engines work to find them useful; most of us are just looking for a friendly interface and a few appropriate results. However, a closer look at how search engines approach the Internet may help broaden one’s information horizons.

Nature of the Relationship

General-purpose search engine developers begin their work by making decisions about what kinds of information resources and Web sites they will include in their product. None of the search engines currently in existence are able to encompass the whole world of Web information. The reasons are many: To begin with, the resources needed to cover the entire Web are massive and prohibitively expensive. Searching so many resources may also be an unrealistic goal because it would require more time than most users would find acceptable. Search engine creators make decisions about what content will serve most of their users most of the time. To achieve their ends, they create computer programs that “travel” around the Web and index sites according to preset guidelines. These Web-searching programs are called Web spiders or Web crawlers or robots.

In the simplest nontechnical terms, a Web spider starts with a Web site to index and then searches that Web site for links to other sites. It follows those links and indexes the sites they lead to and then begins again by looking for more links to follow. The act of indexing makes those Web sites retrievable as search results. The spider program follows the parameters that have been set for it. These limits may involve decisions about what formats to include, how deeply a Web site is searched, and how often it is revisited. Any limitation means that some information sources will be excluded. These exclusions form the Invisible Web.

Search engines are created by three essential programs. The spider locates material to be “indexed.” An indexing program captures and stores the material in a way that makes possible fast and easy retrieval. The final program facilitates what we as users see on the search screen. This program allows users to enter
queries into a box and it supplies the search results. In effect, the search engine creates its own abridged version of the World Wide Web, which it then offers to its users.

**More Information on Web Spiders, Crawlers, Robots**

Here are some basic facts:

- Unless a Web site is linked to from another site, Web spiders may not find it. Web site developers can submit such sites to search engines directly for inclusion.
- A Web site can establish a nonindexing protocol to ensure that the site is not crawled. Existence of such a protocol means that a general-purpose search engine will not index this Web site. Some of the World Wide Web is composed of private sites.
- As Web crawlers are designed to follow links from one site to another, they cannot retrieve information from databases, whose dynamic content does not have permanent URLs that can be linked.
- Each search engine has its own crawler, and many crawlers operate at all times.

For more information about Web crawlers in general see the following:

- Web Robot Pages at www.robotstxt.org

For examples of specific crawlers, see the following:


The information resources found by general-purpose search engines are referred to as the “visible” Web or “surface” Web. The term “invisible” represents those resources that, because of their exclusion by general-purpose search engines, are not so easily found. A popular image that shows the relationship between the visible and Invisible Web is one of a fishing trawler with its nets out in the middle of the ocean (Bergman, 2001). The ocean stands for the world of information available on the World Wide Web. The depth of the ocean reached
by the nets and the content that they can capture represent the realm of the general-purpose search engine and the “surface” Web (see Figures 1-1 and 1-2). The ocean beyond the nets represents the Invisible Web and all of its possibilities. Of course, the fishing boat can go home with its catch and have fulfilled all it hoped to accomplish without even needing the rest of the ocean. Likewise, the information world as captured by general-purpose search engines is often functional enough for many researchers and their needs. However, as mentioned earlier, studies of the Invisible Web have calculated that it is more than 500 times larger than the visible Web (Bergman, 2001). Other studies find different numbers, but the results still show a tremendous difference in size. The Invisible Web represents
enormous information that might be important to the researcher. The researcher needs to know about the existence of this material in order to make informed decisions.

How useful the Invisible Web may be to a researcher will depend on what that researcher hopes to find and how willing he or she may be to continue searching beyond the comfort zone of the general-purpose search engine. What is certain is that when librarians and educators discuss information and research, they need to present the whole picture, not just the surface view.

A Constantly Changing Relationship

As the visible and invisible parts of the Web information world are intrinsically linked, we can deduce more qualities about their relationship. General-purpose search engines vary in which parts of the Web they index and in the proportion of information resources that they offer. Therefore, not only does each search engine create its own Invisible Web of excluded items, but also the size of that content varies from one search engine to another (Sullivan, 2008).

Many search engines, of which Google is a good example, are always adding improvements that enable them to make more inroads into the Invisible Web. At the same time, new information formats that are not necessarily accessible to general-purpose search engines are added to the Web on a regular basis. Such changes add content to the Invisible Web; thus, the nature and size of the Invisible Web are constantly in flux as the Invisible Web adjusts to changes in the search engine world.

Implications of the Relationship

The illustration of the fishing boat on the ocean shows that there is one ocean and that the “visible” and “Invisible” Web are actually parts of the same world of information. As search engines shape the Invisible Web, any discussion of the Invisible Web must be based on knowledge of general-purpose search engines. When librarians and other educators discuss search engines without referencing the Invisible Web, that omission misrepresents the Web information world and may seem to support the Web myth that if a search engine cannot find something, then it is not worth finding. Dissatisfied users of general-purpose search engines may mistakenly conclude that, because they could not find what they needed, it does not exist in the Web information world. The general-purpose search engine is a powerful tool; it does not diminish its importance to present it in the context of the whole information picture.
Characteristics That Make the Invisible Web Important

Size and Quality

The size of the Invisible Web dwarfs the visible Web and that fact alone warrants its consideration; the Invisible Web is simply too large a source of valuable information to ignore. Chapter 7 provides a more detailed discussion of the size of the Invisible Web. The Invisible Web consists of resources not indexed by general-purpose search engines, but that does not mean that it is composed of mere leftovers and unimportant items. Search engines exclude many sources of information because of practical considerations of size, format, and ease of indexing: the limits they set on inclusion are not about quality of information. Hence, the information resources located in the Invisible Web have as much claim to quality as anything found by search engines. In fact, estimates show that as much as 95 percent of Invisible Web content is located in publicly accessible Web sites (Bergman, 2001). That means that anyone can use the material if he or she can find it. Databases and Web sites with extensive information contribute the most material to the Invisible Web, and both kinds of resources offer important content.

Fastest Growing Part of the Web

The Invisible Web also grows at a faster rate than the visible Web. The growth rate is easily explained when considering the amount of new material added to the Web every day. Much of that material does not appear as new Web sites but rather appears as content added to existing sites. Databases may appear as only one result for a search engine query but they represent numerous resources and they grow constantly. They also represent quality information. Search engines also short-change especially rich, deep Web sites, and they also represent large growth areas. Examining the types of materials found in the Invisible Web will help explain its size and growth.

Nature of the Material Included

Databases

The claim that a major source of Invisible Web content is databases can be confusing to users of search engines who can affirm that search engines do in fact find databases. However, search engines cannot enter databases or retrieve their content because databases are often dynamically generated. This is a technical point but an important one. When a user enters a search query, the database
programming searches and assembles an answer. There is no pre-set answer that can be identified quickly, and the files assembled do not necessarily have fixed URLs. Once the answer results are given and the user no longer needs them, the results are disassembled. They do not remain a fixed entity that can be readily identified or linked to again, as do the results screens of general-purpose search engines. The user will need to reconstruct the search query to get the same results. A search engine spider as a program cannot capture such dynamic information. It functions by looking for fixed answers with URLs. The Invisible Web is created in large measure by technical issues of this kind. Many databases, especially those offered by libraries, are subscription or fee-based. Users must have a subscription or belong to an organization such as a library that sponsors the cost for its members. Most libraries offer a selection of proprietary databases; the databases are accessible with a free library card—one of the best deals in the information world.

Subscription databases are usually designed for target audiences such as students and may be the ideal tool for them to use. They offer authority in their selection of materials, aggregating standard journals and publications for inclusion or utilizing editorial boards of experts to make inclusion decisions. By doing so, databases alleviate some of the user’s need to be concerned about the reliability of the information source. By contrast, general-purpose search engines offer no assurance of quality and require that users be concerned with authority, accuracy, and source.

The need for specialization leads to the creation of databases, which enable researchers to tap products specially designed for their area of expertise and need. Databases rely on their own search functions to query content, and these search procedures can be very individual. They pose a problem for the search engine spider that can identify the database homepage but cannot, as yet, fill out search forms and retrieve focused answers. They also pose a problem for the researcher who must “learn” to use the query function. In fact, the researcher may have to keep learning, as he or she moves from one database to another and uses each individualized search form. It is easy to understand the appeal of the simple search box offered by the popular search engines.

Google and other general-purpose search engines are making inroads into this area of proprietary information by establishing agreements with vendors and making listings available through Google Scholar and other such portals. Full-text access to information is still limited, however, to those with subscriptions. In 2008, Google announced that it is experimenting with ways to enable its programming to fill out HTML search forms and thus enable it to capture database content (Sullivan, 2008). If successful, this technology breakthrough will dramatically change Google’s relationship to the Invisible Web.
More about Databases versus Search Engines

Databases and search engines have some distinct differences, although they may not be clear to student researchers. In fact, the distinctions may be blurring as the information world keeps changing.

<table>
<thead>
<tr>
<th>Database</th>
<th>General-Purpose Search Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Collects material:</strong> Utilizes submission process based on peer review, editorial boards, and other review processes.</td>
<td><strong>Collects material:</strong> Uses a Web crawler or by submission.</td>
</tr>
<tr>
<td><strong>Indexing:</strong> Information and data are stored in a uniform way utilizing “fields.” Fields allow for the retrieval of specific pieces of information.</td>
<td><strong>Indexing:</strong> Retrieval is facilitated across several formats.</td>
</tr>
<tr>
<td><strong>Retrieval of information:</strong> Finds results by matching information listed in fields. All fields are searchable.</td>
<td><strong>Retrieval of information:</strong> Finds results by searching for keywords found in text. (Relevancy determined by algorithm or similar such criteria.)</td>
</tr>
<tr>
<td><strong>Scope:</strong> Databases are organized with a special purpose or subject in mind or are intended for a specific audience or membership. All entries in the database relate to its subject, purpose, or some kind of specialization.</td>
<td><strong>Scope:</strong> General-purpose search engines provide across-the-board coverage.</td>
</tr>
<tr>
<td><strong>Authority:</strong> A submission process is in place to ensure authority.</td>
<td><strong>Authority:</strong> Search engine does not evaluate information or Web sites.</td>
</tr>
<tr>
<td><strong>Updating:</strong> Every day new information can be added to a database easily.</td>
<td><strong>Updating:</strong> New Web sites and Web content created every day. Search engines are always crawling for new content and revisit Web sites on a regular basis, although frequency varies from one search engine to another.</td>
</tr>
<tr>
<td><strong>Cost:</strong> Free or fee-based access may depend on membership in a professional association.</td>
<td><strong>Cost:</strong> Free access.</td>
</tr>
<tr>
<td><strong>Examples:</strong> ERIC (free); PubMed (free); EBSCOhost (fee-based)</td>
<td><strong>Examples:</strong> Google; Yahoo!, Live Search, Ask</td>
</tr>
</tbody>
</table>

Deep Web Sites

Similar to databases, Web sites that are very deep and rich in content form a substantial part of the Invisible Web. Search engines set a limit on how much material they index from a site. The “depth of crawl” limitation means that not every page of a Web site will be indexed (Sherman and Price, 2001: 70–71). Information about
the exact “depth of crawl” is hard to find, as search engines do not advertise their limitations. At one time, it was possible to find on the Google site the fact that they limited their crawl to 110 K per site, but that information is no longer available. Although Google may have increased its crawl depth, it does not post the information, and it is unlikely that the increase is without limits (Calishain, 2005). This limit means that very rich, extensive, deep Web sites contain regularly overlooked material, and as the site grows, the proportion of material excluded also grows. Examples of rich, complex sites that receive only partial indexing from search engines include government Web sites such as the Library of Congress (www.loc.gov) and the Census Bureau (www.census.gov). BrightPlanet, a company that provides “deep Web” research for the business world, has produced a list of 60 of the “Largest Deep Web Sites” (www.brightplanet.com/infocenter/largest_deepweb_sites.asp). A review of this list shows open access and fee-based Web sites, the overall preponderance being databases. BrightPlanet reports that the content of these 60 Web sites alone represents 40 times the information found on the surface Web.

Other Invisible Web Resources

Formats

Other varieties of information that form part of the Invisible Web include formats usually untapped by general-purpose search engines. Every format newly available on the Internet requires that search engine producers make decisions to readjust their spider programming to include the new format, develop a special search function for the new format, or simply omit the material. For instance, many search engines offer image searching, but it takes place as a special search option. The basic search may also give results that include images but in a less specific way. The search results will not be images but Web sites to review. These kinds of decisions help create and shape the Invisible Web.

Older formats may also be overlooked, and, again, that material enriches the Invisible Web. A review of several of the search engines can give users a sense of favored formats and excluded formats. Usually, included formats appear on the advanced search page where users may isolate searches according to a particular format. The review should also consider formats that are accessible by special search options, such as images, audio files, and video files.

Forms to Be Completed

Some Web sites, while not presenting themselves as databases, generate dynamic information and present the search engine spider with problems similar to those of databases. For example, sites offering travel directions or job descriptions require information from users. To supply the needed information, they must learn from users where they are and where they need to go or what kind of job they are
seeking. Once users supply the information, the site can generate an answer to their query. Again, the answer is created dynamically for the user and disappears when the user is finished with it. The information that these sites create also falls into the “Invisible Web.”

Current Material

Very current information and new Web sites represent another elusive area for search engines. Search engine spiders not only look for new content and revisit indexed material to locate changes and revisions but the frequency of crawling and revisiting sites varies from one search engine to another (Notess, 2004). A lapse in time can easily occur during which new sites and new material may go unfound. This type of material may eventually be found and cross over from Invisible Web to surface Web content.

SEARCHABILITY OF THE INVISIBLE WEB

Complexity

A search of the surface Web and the Invisible Web requires searching in two distinct types of environments. The surface Web is created and searched by all-purpose tools, general-purpose search engines such as Google, Yahoo!, and Live Search. No corresponding tool accesses the diverse resources of the Invisible Web. It may be discouraging to the researcher who must learn to utilize many search tools and give more time to Invisible Web searching. Any in-depth research of the Invisible Web will be time-consuming and challenging. Experience can help researchers choose useful tools, but certain elusive items may never be found, even though one suspects they exist. The Invisible Web defies easy solutions, although the initiatives described in Chapter 7 are trying to make it more accessible. Individuals will need their navigation skills in place to mine effectively the Invisible Web’s content.

Tools

Chapter 6 covers some of the tools that seem most helpful, but a quick summary of the best resources may be useful here. No one tool dominates but all may be helpful. Some Web tools call themselves search engines for the Invisible Web. CompletePlanet (www.completeplanet.com) and IncyWincy (www.incywincy.com) are examples. Directories such as the Librarian’s Internet Index (www.lii.org) can also guide users to Invisible Web content. Most directories are hybrid tools by nature; they list quality information, but they cannot give the quantity expected by search engine users. Guides to Invisible Web content, especially those designed
Going Beyond Google: The Invisible Web in Learning and Teaching

by librarians and appearing on library Web sites, may also be helpful but limited in scope. Using several and a mixture of these types of tools may be the best way to find appropriate Invisible Web content.

Navigation Skills

Finding the right content may also mean digging deeper into likely Web sites. Navigating a Web site requires more than just using its search features. It may require probing the material available and being able to recognize what links can lead to useful information. Also helpful is an understanding of the structure of the Web site. At the very least, users should be prepared to experiment. Site search functions differ. Directories ask the user to browse through successive menus to find material. They usually start with the broadest approach to a subject and then let the user make choices that lead to more specific material. The researcher makes the decisions and does not depend on a computer program algorithm to determine relevancy. Experience usually makes this process easier, but no less time-consuming.

Lack of Stability

Another aspect of the Invisible Web world of information is that it does change all the time and what worked several months ago may not be helpful the next time around. Smart search engine companies are always finding ways to include more “deep Web” content in their own indexes. Invisible Web tools also change and/or disappear. Directories consume many man-hours to develop, yet still can produce only limited access. If budgets decrease or sponsorship fails, they may disappear or become much less effective. All of the problems that affect general-purpose search engine tools also affect Invisible Web searching tools: many kinds of formats and many more materials being added every day.

The Invisible Web is too vast and problematic in nature to be readily accessed. In fact, it inspires a chicken-and-egg analogy: Because it is so vast and diverse, it lacks the desired organization and is difficult to navigate, and because it is difficult to organize and navigate, it remains and continues to grow even vaster and more diverse. Until further technological inroads are made, the Invisible Web will continue to present difficulties for the researcher.

PLACE IN THE RESEARCH WORLD

We can begin with the premise that most people will start any search for Web information by using a general-purpose search engine. Some people, however, may need to take their research further and go beyond those tools to the Invisible Web.
Characteristics of the Invisible Web

Teachers

These professionals are an important influence on students and the resources they will use for their class assignments and even after they graduate. As discussed in the next chapter, students follow certain patterns in their research and skilled guidance and effective class assignments are required to lead them out of their comfort zones. As is often the case, teachers must stay one step ahead. Educators will also be using the Invisible Web resources for their own research and professional development.

Librarians

Most library users will have already attempted a Web search for themselves before they seek assistance at the reference desk. They will not necessarily feel confident in a reference librarian who appears able to do only the same things that they can do for themselves. Librarians should consider the Invisible Web as the added value that they bring to user requests and be ready to use it effectively—no excuses! Librarians usually have the advantage of subscription databases at their command, selected to meet community needs and often ideally suited to student needs.

Researchers in Any Professional Field

Researchers in any professional field will need to tap Invisible Web content to take their research beyond the level of ordinary expectation. As students progress, their teachers/professors expect higher levels of research skills. That growing level of expertise will include the Invisible Web. Professionals in any field, especially medicine and other sciences, will need to know more than the layperson, and that will require using Invisible Web content.

Dissatisfied Web Searcher

This picture also has a place for the dissatisfied Web searcher. This searcher has tried to use general-purpose search engines to find material that he or she believes should be readily available on the Web but could not locate. Dissatisfaction occurs as a result of frustration and lack of knowledge of alternative ways to search. General-purpose search engines simply cannot always do the job that is needed. Those who have the skills to go beyond search engine resources may find exactly what they want.

Anyone Wanting to Excel

In fact, anyone wanting to excel will want to know about the Invisible Web. People who want to demonstrate that they can produce better and more complete
answers to complex questions, who want to be more creative than their peers, will clearly see the advantages of searching the Invisible Web. It becomes a basis for new ideas, thinking outside of the box, and pushing limits.

CONCLUSIONS

The Invisible Web offers a great array of materials that are important for research on many levels. It offers too much valuable content to be left out of the information picture. Only by understanding its potential, publicizing it, and teaching it can informational professionals hope to bring the Invisible Web to the prominence that it should enjoy. Greater demand for Invisible Web resources will hopefully stimulate development of new technology to solve its access issues. To what extent it is part of student and faculty research is examined in the next chapter.

REFERENCES


