The Invisible Web: Uncovering Sources Search Engines Can't See

Chris Sherman and Gary Price

Abstract

The paradox of the Invisible Web is that it's easy to understand why it exists, but it's very hard to actually define in concrete, specific terms. In a nutshell, the Invisible Web consists of content that's been excluded from general-purpose search engines and Web directories such as Lycos and LookSmart—and yes, even Google. There's nothing inherently "invisible" about this content. But since this content is not easily located with the 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.

In this paper, we define the Invisible Web and delve into the reasons search engines can't "see" its content. We also discuss the four different "types" of invisibility, ranging from the "opaque" Web which is relatively accessible to the searcher, to the truly invisible Web, which requires specialized finding aids to access effectively.

The visible Web is easy to define. It's made up of HTML Web pages that the search engines have chosen to include in their indices. It's no more complicated than that. The Invisible Web is much harder to define and classify for several reasons.

First, many Invisible Web sites are made up of straightforward Web pages that search engines could easily crawl and add to their indices but do not, simply because the engines have decided against including them. This is a crucial point—much of the Invisible Web is hidden *because search engines*

Chris Sherman, President, Searchwise, 898 Rockway Place, Boulder, CO 80303; Gary Price, Librarian, Gary Price Library Research and Internet Consulting, 107 Kinsman View Circle, Silver Spring, MD 20901.
LIBRARY TRENDS, Vol. 52, No. 2, Fall 2003, pp. 282–298

Partially excerpted from *The Invisible Web: Uncovering Information Sources Search Engines Can't See* by Chris Sherman and Gary Price (CyberAge Books, 0–910965-51-X).

have deliberately chosen to exclude some types of Web content. We're not talking about unsavory "adult" sites or blatant spam sites—quite the contrary! Many Invisible Web sites are first-rate content sources. These exceptional resources simply cannot be found using general-purpose search engines because they have been effectively locked out.

There are a number of reasons for these exclusionary policies, many of which we'll discuss. But keep in mind that, should the engines change their policies in the future, sites that today are part of the Invisible Web will suddenly join the mainstream as part of the visible Web. In fact, since the publication of our book *The Invisible Web: Uncovering Information Sources Search Engines Can't See* (Medford, NJ: CyberAge Books, 2001, 0–910965-51-X/softbound), most major search engines are now including content that was previously hidden—we'll discuss these developments below.

Second, it's relatively easy to classify some sites as either visible or invisible based on the technology they employ. Some sites using database technology, for example, are genuinely difficult for current generation search engines to access and index. These are "true" Invisible Web sites. Other sites, however, use a variety of media and file types, some of which are easily indexed and others that are incomprehensible to search engine crawlers. Web sites that use a mixture of these media and file types aren't easily classified as either visible or invisible. Rather, they make up what we call the "opaque" Web.

Finally, search engines could theoretically index some parts of the Invisible Web, but doing so would simply be impractical, either from a cost standpoint, or because data on some sites is ephemeral and not worthy of indexing—for example, current weather information, moment-by-moment stock quotes, airline flight arrival times, and so on. However, it's important to note that, even if all Web engines "crawled" everything, an unintended consequence could be that, with the vast increase in information to process, finding the right "needle" in a larger "haystack" might become more difficult. Invisible Web tools offer limiting features for a specific data set, potentially increasing precision. General engines don't have these options. So the database will increase but precision could suffer.

Invisible Web Defined

The Invisible Web: Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages. Sometimes also referred to as the "deep Web" or "dark matter."

This definition is deliberately very general, because the general-purpose search engines are constantly adding features and improvements to their services. What may be invisible today may suddenly become visible tomorrow, should the engines decide to add the capability to index things that they cannot or will not currently index.

Let's examine the two parts of this definition in more detail. First, we'll look at the technical reasons search engines can't index certain types of material on the Web. Then we'll talk about some of the other nontechnical but very important factors that influence the policies that guide search engine operations.

At their most basic level, search engines are designed to index Web pages. Search engines use programs called crawlers (a.k.a., "spiders" and "robots") to find and retrieve Web pages stored on servers all over the world. From a Web server's standpoint, it doesn't make any difference if a request for a page comes from a person using a Web browser or from an automated search engine crawler. In either case, the server returns the desired Web page to the computer that requested it.

A key difference between a person using a browser and a search engine spider is that the person can manually type a URL into the browser window and retrieve the page the URL points to. Search engine crawlers lack this capability. Instead, they're forced to rely on links they find on Web pages to find other pages. If a Web page has no links pointing to it from any other page on the Web, a search engine crawler can't find it. These "disconnected" pages are the most basic part of the Invisible Web. There's nothing *preventing* a search engine from crawling and indexing disconnected pages—but without links pointing to the pages, there's simply no way for a crawler to discover and fetch them.

Disconnected pages can easily leave the realm of the invisible and join the visible Web in one of two ways. First, if a connected Web page links to the disconnected page, a crawler can discover the link and spider the page. Second, the page author can request that the page be crawled by submitting it to "search engine add URL" forms.

Technical problems begin to come into play when a search engine crawler encounters an object or file type that's not a simple text document. Search engines are designed to index text and are highly optimized to perform search and retrieval operations on text. But they don't do very well with nontextual data, at least in the current generation of tools.

Some engines, like AltaVista and Google, can do limited searching for certain kinds of nontext files, including images, audio, or video files. But the way they process requests for this type of material are reminiscent of early Archie searches, typically limited to a filename or the minimal alternative (ALT) text that's sometimes used by page authors in the HTML image tag. Text surrounding an image, sound, or video file can give additional clues about what the file contains. But keyword searching with images and sounds is a far cry from simply telling the search engine to "find me a picture that looks like Picasso's 'Guernica'" or "let me hum a few bars of this song and you tell me what it is." Pages that consist primarily of images, audio, or video, with little or no text, make up another type of Invisible Web content. While the pages may actually be included in a search engine index, they provide few textual clues as to their content, making it highly unlikely they will ever garner high relevance scores.

Researchers are working to overcome these limitations. Google, for example, has experimented with optical character recognition processes for extracting text from photographs and graphic images, in its experimental Google Catalogs project (*Google Catalogs,* n.d.). While not particularly useful to serious searchers, Google Catalogs illustrates one possibility for enhancing the capability of crawlers to find Invisible Web content.

Another company, Singingfish (owned by Thompson) indexes audio streaming media and makes use of metadata embedded in the files to enhance the search experience (*Singingfish,* n.d.). ShadowTV performs near real-time indexing of television audio and video, converting spoken audio to text to make it searchable (*Shadow TV,* n.d.).

While search engines have limited capabilities to index pages that are primarily made up of images, audio, and video, they have serious problems with other types of nontext material. Most of the major general-purpose search engines simply cannot handle certain types of formats. When our book was first written, PDF and Microsoft Office format documents were among those not indexed by search engines. Google pioneered the indexing of PDF and Office documents, and this type of search capability is widely available today.

However, a number of other file formats are still largely ignored by search engines. These formats include:

- Postscript,
- Flash,
- Shockwave,
- Executables (programs), and
- Compressed files (.zip, .tar, etc.).

The problem with indexing these files is that they aren't made up of HTML text. Technically, most of the formats in the list above can be indexed. AlltheWeb.com, for example, recently began indexing the text portions of Flash files, and Google can follow links embedded within Flash files.

The primary reason search engines choose not to index certain file types is a business judgment. For one thing, there's much less user demand for these types of files than for HTML text files. These formats are also "harder" to index, requiring more computing resources. For example, a single PDF file might consist of hundreds or even thousands of pages, so even those engines that do index PDF files typically ignore parts of a document that exceed 100K bytes or so. Indexing non-HTML text file formats tends to be costly. In other words, the major Web engines are not in business to meet every need of information professionals and researchers.

Pages consisting largely of these "difficult" file types currently make up a relatively small part of the Invisible Web. However, we're seeing a rapid expansion in the use of many of these file types, particularly for some kinds of high-quality, authoritative information. For example, to comply with federal paperwork reduction legislation, many U.S. government agencies are moving to put all of their official documents on the Web in PDF format. Most scholarly papers are posted to the Web in Postscript or compressed Postscript format. For the searcher, Invisible Web content made up of these file types poses a serious problem. We discuss a partial solution to this problem later in this article.

The biggest technical hurdle search engines face lies in accessing information stored in databases. This is a huge problem, because there are thousands—perhaps millions—of databases containing high-quality information that are accessible via the Web. Web content creators favor databases because they offer flexible, easily maintained development environments. And increasingly, content-rich databases from universities, libraries, associations, businesses, and government agencies are being made available online, using Web interfaces as front-ends to what were once closed, proprietary information systems.

Databases pose a problem for search engines because every database is unique in both the design of its data structures and its search and retrieval tools and capabilities. Unlike simple HTML files, which search engine crawlers can simply fetch and index, content stored in databases is trickier to access, for a number of reasons that we'll describe in detail below.

Search engine crawlers generally have no difficulty finding the interface or gateway pages to databases because these are typically pages made up of input fields and other controls. These pages are *formatted* with HTML and look like any other Web page that uses interactive forms. Behind the scenes, however, are the knobs, dials, and switches that provide access to the actual contents of the database, which are literally incomprehensible to a search engine crawler.

Although these interfaces provide powerful tools for a human searcher, they act as roadblocks for a search engine spider. Essentially, when an indexing spider comes across a database, it's as if it has run smack into the entrance of a massive library with securely bolted doors. A crawler can locate and index the library's address, but because the crawler cannot penetrate the gateway it can't tell you anything about the books, magazines, or other documents it contains.

These Web-accessible databases make up the lion's share of the Invisible Web. They are accessible *via* the Web but may or may not actually be *on* the Web. To search a database you must use the powerful search and retrieval tools offered by the database itself. The advantage to this direct approach is that you can use search tools that were specifically designed to retrieve the best results from the database. The disadvantage is that you need to find the database in the first place, a task the search engines may or may not be able to help you with.

There are several different kinds of databases used for Web content, and it's important to distinguish between them. Just because Web content is stored in a database doesn't automatically make it part of the Invisible Web. Indeed, some Web sites use databases not so much for their sophisticated query tools, but rather because database architecture is more robust and makes it easier to maintain a site than if it were simply a collection of HTML pages.

One type of database is designed to deliver tailored content to individual users. Examples include My Yahoo!, Personal Excite, Quicken.com's personal portfolios, and so on. These sites use databases that generate "on the fly" HTML pages customized for a specific user. Since this content is tailored for each user there's little need to index it in a general-purpose search engine.

A second type of database is designed to deliver streaming or real-time data—stock quotes, weather information, airline flight arrival information, and so on. This information isn't necessarily customized, but it is stored in a database due to the huge, rapidly changing quantities of information involved. Technically, much of this kind of data is indexable because the information is retrieved from the database and published in a consistent, straight HTML file format. But because it changes so frequently, and has value for such a limited duration (other than to scholars or archivists), there's no point in indexing it. It's also problematic for crawlers to keep up with this kind of information. Even the fastest crawlers revisit most sites monthly or even less frequently (other than news crawlers, which are designed to track rapidly changing news sites). Staying current with real-time information would consume so many resources it is effectively impossible for a crawler.

The third type of Web-accessible database is optimized for the data it contains, with specialized query tools designed to retrieve the information using the fastest or most effective means possible. These are often "relational" databases that allow sophisticated querying to find data that are "related" based on criteria specified by the user. The only way of accessing content in these types of databases is by directly interacting with the database. It is this content that forms the core of the Invisible Web.

Let's take a closer look at these elements of the Invisible Web and demonstrate exactly why search engines can't or won't index them.

Why Search Engines Can't See the Invisible Web

Text—more specifically *hyper*text—is the fundamental medium of the Web. The primary function of search engines is to help users locate hypertext documents of interest. Search engines are highly tuned and optimized to deal with text pages and, even more specifically, text pages that have been encoded with the HyperText Markup Language (HTML). As the Web evolves and additional media become commonplace, search engines will undoubtedly offer new ways of searching for this information. But for now, the core function of most Web search engines is to help users locate text documents.

HTML documents are simple. Each page has two parts: a "head" and a "body" which are clearly separated in the source code of an HTML page. The head portion contains a title, which is displayed (logically enough) in the title bar at the very top of a browser's window. The head portion may also contain some additional metadata describing the document, which can be used by a search engine to help classify the document. For the most part, other than the title, the head of a document contains information and data that helps the Web browser display the page but is irrelevant to a search engine. The body portion contains the actual document itself. This is the meat that the search engine wants to digest.

The simplicity of this format makes it easy for search engines to retrieve HTML documents, index every word on every page, and store them in huge databases that can be searched on demand. Problems arise when content doesn't conform to this simple Web page model. To understand why, it's helpful to consider the process of crawling and the factors that influence whether a page either can or will be successfully crawled and indexed.

The first thing a crawler attempts to determine is whether access to pages on a server it is attempting to crawl is restricted. Webmasters can use three methods to prevent a search engine from indexing a page. Two methods use blocking techniques specified in the *Robots Exclusion Protocol* that most crawlers voluntarily honor and one creates a technical roadblock that cannot be circumvented (*Robots Exclusion Protocol,* n.d.).

The Robots Exclusion Protocol is a set of rules that enable a Webmaster to specify which parts of a server are open to search engine crawlers, and which parts are off-limits. The Webmaster simply creates a list of files or directories that should not be crawled or indexed and saves this list on the server in a file named robots.txt. This optional file, stored by convention at the top level of a Web site, is nothing more than a polite request to the crawler to keep out, but most major search engines respect the protocol and will not index files specified in robots.txt.

The second means of preventing a page from being indexed works in the same way as the robots.txt file, but it is page-specific. Webmasters can prevent a page from being crawled by including a "noindex" metatag instruction in the "head" portion of the document. Either robots.txt or the noindex metatag can be used to block crawlers. The only difference between the two is that the noindex metatag is page specific, while the robots.txt file can be used to prevent indexing of individual pages, groups of files, or even entire Web sites.

Password protecting a page is the third means of preventing it from being crawled and indexed by a search engine. This technique is much stronger than the first two since it uses a technical barrier rather than a voluntary standard.

Why would a Webmaster block crawlers from a page using the Robots Exclusion Protocol rather than simply password protecting the pages? Password protected pages can be accessed only by the select few users that know the password. Pages excluded from engines using the Robots Exclusion Protocol, on the other hand, can be accessed by anyone *except* a search engine crawler. The most common reason Webmasters block content from indexing is that a page changes far more frequently than the engines can keep up with.

Pages using any of the three methods described above are part of the Invisible Web. In many cases, they contain no technical roadblocks that prevent crawlers from spidering and indexing the page. They are part of the Invisible Web because the Webmaster has opted to keep them out of the search engines.

Once a crawler has determined whether it is permitted access to a page, the next step is to attempt to fetch it and hand it off to the search engine's indexer component. This crucial step determines to a large degree whether a page is visible or invisible. Let's examine some variations crawlers encounter as they discover pages on the Web, using the same logic they do to determine whether a page is indexable or not.

Case 1

The crawler encounters a page that is straightforward HTML text, possibly including basic Web graphics. This is the most common type of Web page. It is visible and can be indexed, assuming the crawler can discover it.

Case 2

The crawler encounters a page made up of HTML, but it's a form, consisting of text fields, check boxes, or other components requiring user input. It might be a sign-in page, requiring a user name and password. It might be a form requiring the selection of one or more options. The form itself, since it's made up of simple HTML, can be fetched and indexed. But the content *behind* the form (what the user sees after clicking the submit button) may be invisible to a search engine. There are two possibilities here:

- The form is used simply to select user preferences. Other pages on the site consist of straightforward HTML that can be crawled and indexed (presuming there are links from other pages elsewhere on the Web pointing to the pages). In this case, the form and the content behind it are visible and can be included in a search engine index. Quite often, sites like this are specialized search sites for specific types of content. A good example is Hoover's Business Profiles, which provides a form to search for a company, but presents company profiles in straightforward HTML that can be indexed (*Hoover's Online,* n.d.).

- The form is used to collect user-specified information that will generate dynamic pages when the information is submitted. In this case, although the form is visible, the content "behind" it is invisible. Since the only way to access the content is by using the form, how can a crawler, which is simply designed to request and fetch pages, possibly know what to enter into the form? Since forms can literally have infinite variations, if they function to access dynamic content they are essentially roadblocks for crawlers. A good example of this type of Invisible Web site is the World Bank Group Economics of Tobacco Control Country Data Report Database, which allows you to select any country and choose a wide range of reports for that country (*Economics of Tobacco-Country Data Report,* n.d.). It's interesting to note here that this database is just one part of a much larger site, the bulk of which is fully visible. So even if the search engines do a comprehensive job of indexing the visible part of the site, this valuable information still remains hidden to all but those searchers who visit the site and discover the database on their own.

In the future, forms will pose less of a challenge to search engines. Several projects are underway aimed at creating more intelligent crawlers that can fill out forms and retrieve information. One approach uses preprogrammed "brokers" designed to interact with the forms of specific databases. Other approaches combine brute force with artificial intelligence to "guess" what to enter into forms, allowing the crawler to "punch through" the form and retrieve information. It's not a trivial problem: In a conversation with Google's Chief Technology Officer, Craig Silverstein, he estimated that it may take as long as fifty years before Google has the capability to index all Invisible Web content. And even if general-purpose search engines do acquire the ability to crawl content in databases, it's likely that the native search tools provided by each database will remain the best way to interact with most databases.

Case 3

The crawler encounters a dynamically generated page assembled and displayed on demand. The telltale sign of a dynamically generated page is the "?" symbol appearing in its URL. Technically, these pages are part of the visible Web. Crawlers can fetch any page that can be displayed in a Web browser, regardless of whether it's a static page stored on a server or generated dynamically. A good example of this type of Invisible Web site is

Compaq's experimental SpeechBot search engine, which indexes audio and video content using speech recognition and converts the streaming media files to viewable text (*SpeechBot,* n.d.). Somewhat ironically, one could make a good argument that *most* search engine result pages are *themselves* Invisible Web content, since they generate dynamic pages on the fly in response to user search terms.

Dynamically generated pages pose a challenge for crawlers. Dynamic pages are created by a *script,* a computer program that selects from various options to assemble a customized page. Until the script is actually run, a crawler has no way of knowing what it will actually do. The script *should* simply assemble a customized Web page. Unfortunately, unethical Webmasters have created scripts to generate literally millions of similar but not quite identical pages in an effort to "spamdex" the search engine with bogus pages. Sloppy programming can also result in a script that puts a spider into an endless loop, repeatedly retrieving the same page.

These "spider traps" can be a real drag on the engines, so most have simply made the decision not to crawl or index URLs that generate dynamic content. They're "apartheid" pages on the Web—separate but equal, making up a big portion of the "opaque" Web that potentially can be indexed but is not. Inktomi's FAQ about its crawler, named "Slurp," offers this explanation:

> Slurp now has the ability to crawl dynamic links or dynamically generated documents. It will not, however, crawl them by default. There are a number of good reasons for this. A couple of reasons are that dynamically generated documents can make up infinite URL spaces, and that dynamically generated links and documents can be different for every retrieval so there is no use in indexing them. (*Slurp,* n.d.)

As crawler technology improves, it's likely that one type of dynamically generated content will increasingly be crawled and indexed. This is content that essentially consists of static pages that are stored in databases for production efficiency reasons. As search engines learn which sites providing dynamically generated content can be trusted not to subject crawlers to spider traps, content from these sites will begin to appear in search engine indices. It's important to note that even as search engines learn which content is acceptable, they still may not index everything, as evidenced by this statement from Google's Webmaster tips page: "We are able to index dynamically generated pages. However, because our web crawler can easily overwhelm and crash sites serving dynamic content, we limit the amount of dynamic pages we index" (*Google Information for Webmasters,* n.d.).

Another development that has reduced the barriers for dynamic content is the increasing adoption of *paid inclusion* programs by the major search engines. These programs are designed to allow Webmasters to specify specific pages for crawling and guaranteed indexing, in exchange for an annual fee. The search engines give no preferential treatment to these pages beyond guaranteed indexing, and spam rules still apply. Any pages that violate search engine spam policies, whether crawled or submitted via paid exclusion, are subject to removal from the index. Paid inclusion is a means for search engines to trust dynamic content, on the theory that nobody would willingly pay just to have their content removed anyway.

Case 4

The crawler encounters an HTML page with nothing to index. There are thousands, if not millions, of pages that have a basic HTML framework, but which contain only Flash; images in the .gif, .jpeg, or other Web graphics format; streaming media; or other nontext content in the body of the page. These types of pages are truly parts of the Invisible Web because there's nothing for the search engine to index. Specialized multimedia search engines are able to recognize some of these nontext file types and index minimal information about them, such as file name and size, but these are far from keyword searchable solutions.

Case 5

The crawler encounters a site offering dynamic, real-time data. There are a wide variety of sites providing this kind of information, ranging from real-time stock quotes to airline flight arrival information. These sites are also part of the Invisible Web, because these data streams are, from a practical standpoint, unindexable. While it's technically possible to index many kinds of real-time data streams, the value would only be for historical purposes, and the enormous amount of data captured would quickly strain a search engine's storage capacity, so it's a futile exercise. A good example of this type of Invisible Web site is Cheap Ticket's FlightTracker, which provides real-time flight arrival information taken directly from the cockpit of in-flight airplanes (*FlightTracker*, n.d.).

Case 6

The crawler encounters a PDF or Postscript file. PDF and Postscript are text formats that preserve the look of a document and display it identically regardless of the type of computer used to view it. While many search engines index PDF files, most do not index the full text of the documents. Google stops indexing after 120KB; AlltheWeb stops indexing after 110KB.

An experimental search engine called ResearchIndex, created by computer scientists at the NEC Research Institute, not only indexes the full text of PDF and Postscript files, it also takes advantage of the unique features that commonly appear in documents using the format to improve search results (*CiteSeer*, n.d.). For example, academic papers typically cite other documents and include lists of references to related material. In addition to indexing the full text of documents, ResearchIndex also creates a citation index that makes it easy to locate related documents. It also appears that citation searching has little overlap with keyword searching, so combining the two can greatly enhance the relevance of results.

Case 7

The crawler encounters a database offering a Web interface. There are tens of thousands of databases containing extremely valuable information available via the Web. But search engines cannot index the material in them. Although we present this as a unique case, Web-accessible databases are essentially a combination of cases 2 and 3. Databases generate Web pages dynamically, responding to commands issued through an HTML form. Though the interface to the database is an HTML form, the database itself may have been created before the development of HTML, and its legacy system is incompatible with protocols used by the engines, or they may require registration to access the data. Finally, they may be proprietary, accessible only to select users, or users who have paid a fee for access.

Ironically, the original HTTP specification developed by Web inventor Tim Berners-Lee included a feature called format negotiation that allowed a client to say what kinds of data it could handle and allow a server to return data in any acceptable format. Berners-Lee's vision encompassed the information in the Invisible Web, but this vision, at least from a search engine standpoint, has largely been unrealized.

These technical limitations give you an idea of the problems encountered by search engines when they attempt to crawl Web pages and compile indices. There are other, nontechnical reasons why information isn't included in search engines. We look at those next.

Four Types of Invisible

Technical reasons aside, there are other reasons that some kinds of material that can be accessed either on or via the Internet are not included in search engines. There are really four "types" of invisible Web content. We make these distinctions not so much to make hard and fast distinctions between the types, but rather to help illustrate the amorphous boundary of the Invisible Web that makes defining it in concrete terms so difficult.

The four types of invisible are:

- The "Opaque" Web,
- The Private Web,
- The Proprietary Web, and
- The Truly Invisible Web.

The "Opaque" Web

The "Opaque" Web consists of files that *can be*, but are not, included in search engine indices. The Opaque Web is quite large and presents a unique challenge to a searcher. Whereas the deep content in many truly

Invisible Web sites is accessible if you know how to find it, material on the Opaque Web is often much harder to find.

The biggest part of the Opaque Web consists of files that the search engines *can* crawl and index, but simply do not. There are a variety of reasons for this; let's look at them.

Depth of Crawl

Crawling a Web site is a resource-intensive operation. It costs money for a search engine to crawl and index every page on a site. In the past, most engines would merely sample a few pages from a site rather than performing a "deep crawl" that indexed every page, reasoning that a sample provided a "good enough" representation of a site that would satisfy the needs of most searchers. Limiting the depth of crawl also reduced the cost of indexing a particular Web site.

In general, search engines don't reveal how they set the depth of crawl for Web sites. Increasingly, there is a trend to crawl more deeply, to index as many pages as possible. As the cost of crawling and indexing goes down, and the size of search engine indices continues to be a competitive issue, the depth of crawl issue is becoming less of a concern for searchers. Nonetheless, simply because one, fifty, or five thousand pages from a site are crawled and made searchable, there is no guarantee that every page from a site will be crawled and indexed. This problem gets little attention and is one of the top reasons why useful material may be all but invisible to those who only use general-purpose search tools to find Web materials.

Frequency of Crawl

The Web is in a constant state of dynamic flux. New pages are added constantly, and existing pages are moved or taken off the Web. Even the most powerful crawlers typically visit only about 10 million pages per day, a fraction of the entire number of pages on the Web. This means that each search engine must decide how best to deploy its crawlers, creating a schedule that determines how frequently a particular page or site is visited.

Web Search researchers Steve Lawrence and Lee Giles, writing in the July 8, 1999, issue of *Nature,* state that "indexing of new or modified pages by just one of the major search engines can take months" (Lawrence and Giles, 1999). While the situation appears to have improved since their study, most engines only completely "refresh" their indices monthly or even less frequently.

It's not enough for a search engine to simply visit a page once and then assume it's still available thereafter. Crawlers must periodically return to a page to not only verify its existence, but also to download the freshest copy of the page and perhaps fetch new pages that have been added to a site. According to one study, it appears that the half-life of a Web page is somewhat less than two years and the half-life of a Web site is somewhat more than two years. Put differently, this means that if a crawler returned to a site spidered two years ago it would contain the same number of URLs, but only half of the original pages would still exist, having been replaced by new ones ("Graph Structure in the Web," n.d.; "Altavista, Compaq, and IBM," n.d.).

New sites are the most susceptible to oversight by search engines because relatively few other sites on the Web will have linked to them compared to more established sites. Until search engines index these new sites, they remain part of the Invisible Web.

Maximum Number of Viewable Results

It's quite common for a search engine to report a very large number of results, sometimes into the millions of documents. However, most engines also restrict the total number of results they will display for a query, typically between 200 and 1,000 documents. For queries that return a huge number of results, this means that the majority of pages the search engine has determined might be relevant are inaccessible, since the result list is arbitrarily truncated. Those pages that don't make the cut are effectively invisible.

Good searchers are aware of this problem and will take steps to circumvent it by using a more precise search strategy and the advanced filtering and limiting controls offered by many engines. However, for many inexperienced searchers this limit on the total number of viewable hits can be a problem. What happens if the answer you need is available (with a more carefully crafted search) but cannot be viewed using your current search terms?

Disconnected URLs

For a search engine crawler to access a page, one of two things must take place. Either the Web page author uses the search engine's "Submit URL" feature to request that the crawler visit and index the page, or the crawler discovers the page on its own by finding a link to the page on some other page. Web pages that aren't submitted directly to the search engines, and that don't have links pointing to them from other Web pages, are called "disconnected" URLs and cannot be spidered or indexed simply because the crawler has no way to find them.

Quite often, these pages present no technical barrier for a search engine. But the authors of disconnected pages are clearly unaware of the requirements for having their pages indexed. A May 2000 study by IBM, AltaVista, and Compaq discovered that the total number of disconnected URLs makes up about 20 percent of the potentially indexable Web, so this isn't an insignificant problem ("Graph Structure in the Web," n.d.; "Altavista, Compaq, and IBM," n.d.).

In summary, the Opaque Web is large, but is not impenetrable. Determined searchers can often find material on the Opaque Web, and search engines are constantly improving their methods for locating and indexing Opaque Web material.

The three other types of invisible are more problematic, as we'll see.

The Private Web

The Private Web consists of technically indexable Web pages that have deliberately been excluded from inclusion in search engines. There are three ways Webmasters can exclude a page from a search engine:

- Password protect the page. A search engine spider cannot go past the form that requires a username and password;
- Use the robots.txt file to disallow a search spider from accessing the page;
- Use the "noindex" metatag to prevent the spider from reading past the head portion of the page and indexing the body.

For the most part, the Private Web is of little concern to most searchers. Private Web pages simply use the public Web as an efficient delivery and access medium, but in general are not intended for use beyond the people who have permission to access the pages.

There are other types of pages that have restricted access that may be of interest to searchers, yet they typically aren't included in search engine indices. These pages are part of the "Proprietary" Web, which we describe next.

The Proprietary Web

Search engines cannot for the most part access pages on the Proprietary Web, because these pages are only accessible to people who have agreed to special terms in exchange for viewing the content. Proprietary pages may simply be content that's only accessible to users willing to register to view them. Registration in many cases is free, but a search crawler clearly cannot satisfy the requirements of even the simplest registration process.

Other types of proprietary content are available only for a fee, whether on a per-page basis or via some sort of subscription mechanism. Examples of proprietary fee-based Web sites include Hoover's and the Wall Street Journal Interactive Edition.

Proprietary Web services are not the same as traditional online information providers, such as Dialog, Lexis-Nexis, and Dow Jones. These services offer Web access to proprietary information but use legacy database systems that existed long before the Web came into being. While the content offered by these services is exceptional, they are not considered to be Web or Internet providers.

The Truly Invisible Web

Some Web sites or pages are truly invisible, meaning that there are technical reasons that search engines can't spider or index the material they have to offer. A definition of what constitutes a truly invisible resource must necessarily be somewhat fluid, since the engines are constantly improving and adapting their methods to embrace new types of content. But at the time of writing truly invisible content consisted of several types of resources.

The simplest, and least likely to remain invisible over time, are Web pages that use file formats that current generation Web crawlers aren't programmed to handle. These file formats include PDF, postscript, Flash, Shockwave, executables (programs), and compressed files. There are two reasons search engines do not currently index these types of files. First, the files have little or no textual context, so it's difficult to categorize them, or compare them for relevance to other text documents. The addition of metadata to the HTML container carrying the file could solve this problem—but it would nonetheless be the metadata description that got indexed rather than the contents of the file itself.

The second reason certain types of files don't appear in search indices is simply because the search engines have chosen to omit them. They *can* be indexed, but aren't. You can see a great example of this in action with the Research Index engine, which retrieves and indexes PDF, Postscript, and even compressed files in real time, creating a searchable database that's specific to your query. AltaVista's Search Engine product for creating local site search services is capable of indexing more than 250 file formats, but the flagship public search engine includes only a few of these formats. It's typically lack of willingness, not an ability issue with file formats.

More problematic are dynamically generated Web pages. Again, in some cases, it's not a technical problem but rather unwillingness on the part of the engines to index this type of content. This occurs specifically when a noninteractive script is used to generate a page. These are static pages, and generate static HTML that the engine could spider. The problem is that unscrupulous use of scripts can also lead crawlers into "spider traps" where the spider is literally trapped within a huge site of thousands, if not millions, of pages designed solely to spam the search engine. This is a major problem for the engines, so they've simply opted not to index URLs that contain script commands.

Finally, information stored in relational databases, which cannot be extracted without a specific query to the database, is truly invisible.

Crawlers aren't programmed to understand either the database structure, or the command language used to extract information.

Conclusion

The Invisible Web is a vast portion of cyberspace, and offers invaluable resources that should not be overlooked by serious searchers. Although search engine technology continues to improve, the Invisible Web is largely an intractable problem that will be with us for some time to come. Although it's a vast and useful resource, it's important not to get bogged down in the semantics. An information professional should treat these types of resources like traditional reference tools. Learn what's available and have them ready to go. The best way for searchers to access the Invisible Web is to build and bookmark a personal collection of resources, treating them as a personal "reference library," and using them when needed, rather than relying on search engines that in many cases simply cannot access the content residing on the Invisible Web.

References

Altavista, Compaq, and IBM researchers create world's largest, most accurate picture of the Web. (n.d.). [Summary of "Graph Structure in the Web," (n.d.)]. Retrieved August 27, 2003, from http://www.almaden.ibm.com/almaden/webmap_release.html.

CiteSeer: The NEC Research Institute Scientific Literature Digital Library. (n.d.). Retrieved April 17, 2003, from http://www.researchindex.com. Commonly referred to as ResearchIndex.

Economics of tobacco-country data report. (n.d.). Retrieved April 14, 2003, from http://www1.worldbank.org/tobacco/database.asp.

FlightTracker. (n.d.). Retrieved April 16, 2003, from CheapTickets Travel Web site: http://www.cheaptickets.com/trs/cheaptickets/flighttracker/flight_tracker_graphic.xsl.

Google catalogs. (n.d.). Retrieved April 17, 2003, from http://catalogs.google.com.

Google information for Webmasters. (n.d.). Retrieved April 17, 2003, from http://www.google.com/webmasters/2.html.

Graph structure in the Web. (n.d.). Retrieved August 27, 2003, from http://www9.org/w9cdrom/160/160.html.

Hoover's online. (n.d.). Retrieved April 15, 2003, from http://www.hoovers.com/.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the web. *Nature, 400,* 107–109. Additional material can be found at: http://wwwmetrics.com/.

Robots exclusion. (n.d.). Retrieved April 17, 2003, from http://www.robotstxt.org/wc/exclusion.html.

Shadow TV. (n.d.). Retrieved March 17, 2003, from http://www.shadowtv.com.

Singingfish. (n.d.). Retrieved April 9, 2003, from http://www.singingfish.com/.

Slurp. (n.d.). Retrieved April 17, 2003, from http://www.inktomi.com/slurp.html.

Speechbot. (n.d.). Retrieved April 15, 2003, from http://www.speechhot.com.

Additional Readings

Guernsey, L. (2001, January 25). Mining the "deep web" with sharper shovels. *New York Times,* p. G1.

Price, G. (2002). Specialized search engine FAQs: More questions, answers, and issues. *Searcher, 10*(9), 42–48.

Price, G., & Sherman, C. (2001). Exploring the invisible Web: Seven essential strategies. *Online, 25*(4), 32–35.

Sherman, C. (2001). Google unveils more of the invisible Web. *Search Day.* Retrieved April 17, 2003, from http://www.searchenginewatch.com/searchday/article.php/2158091.