



"The Invisible Web"

There's a big problem with search engines, and many people aren't aware of it. The problem is that vast expanses of the Web are completely invisible to general purpose search engines like Google. Even worse, this "Invisible Web" is in all likelihood growing significantly faster than the visible Web you're familiar with.

So what is this Invisible Web and why aren't search engines indexing it? To answer this question, it's important to first define the "visible" Web, and describe how search engines compile their indexes.

HTML documents are simple: they consist of a "head" portion, with a title and perhaps some additional meta data describing the document, and a "body" portion, the actual document itself. 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.

What's less easy is the task of actually finding all the pages on the Web. Search engines use automated programs called spiders or robots to "crawl" the Web and retrieve pages. Spiders function much like a hyper-caffeinated Web browser - they rely on links to take them from page to page.

Crawling is a resource-intensive operation. For this reason, search engines will often limit the number of pages they retrieve and index from any given Web site. It's tempting to think that these unretrieved pages are part of the Invisible Web, but they aren't. They are visible and indexable, but the search engines have made a conscious decision not to index them.

Why can't these pages be indexed? The most basic reason is that there are no links pointing to a page that a search engine spider can follow. Or, a page may be made up of data types that search engines don't index, such as some CGI scripts. Or the search engine will not index aspects of a particular web site - an example of this is that Google will not index a site past the 100Kb point.

Much has been made of these overlooked pages. Many of the major engines are making serious efforts to include them and make their indexes more comprehensive. Google claims to have indexed 40% of the indexable web pages. But previous search engines have needed to cull their indexes to compensate for spam and duplicates. This is almost certainly true of Google as well, but it has not recently given the size of its index. However, these numbers

don't include Web pages that can't be indexed, or information that's available via the Web but isn't accessible by the search engines. This is the stuff of the Invisible Web.

The biggest part of the Invisible Web is made up of information stored in databases. 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. Spiders can record the library's address, but can tell you nothing about the information it contains. Examples of this are the databases subscribed to by most university libraries.

There are thousands - perhaps millions - of databases containing high-quality information that are accessible via the Web. But in order to search them, you typically must visit the Web site that provides an interface to the database. 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 a search engine may not be able to help you with.

Another problem is that content in some databases isn't designed to be directly searchable. Instead, Web developers are taking advantage of database technology to offer customized content that's often assembled on the fly. Search engine results pages are an example of this type of dynamically generated content. As Web sites get more complex and users demand more personalization, this trend toward dynamically generated content will accelerate, making it even harder for search engines to create comprehensive Web indexes.

In a nutshell, the Invisible Web is made up of unindexable content that a search engine either can't or won't index. It's a huge part of the Web, and it's growing. Fortunately, there are several reasonably thorough guides to the Invisible Web.

Prior to joining Search Engine Watch and ResourceShelf, Gary Price worked as a reference librarian at George Washington University in Washington, D.C. He has compiled several well-known web research tools, including Price's List of Lists and Direct Search, a compilation of invisible Web databases. Gary currently works for Ask.com, where his role is to help make searches more effective. He has assembled a massive collection of links to Invisible Web resources at his Direct Search page <http://www.resourceshelf.com/category/uncategorized>.

What kinds of databases does Price consider to be essential Invisible Web search tools? He names four as examples:

- The many databases that make up GPO Access.
<http://www.gpoaccess.gov/databases.html>

- Any of the telephone directory databases, such as

www.canada411.com

And two that are crucial to the business searcher:

- Any of the many flavors of EDGAR, particularly the 10K Wizard.

<http://www.tenkwizard.com/>

PricewaterhouseCoopers Money Tree Survey of Venture Capital

<http://www.pwcmoneytree.com/moneytree/index.jsp>

Two other Invisible Web resources Price maintains are his NewsCenter <http://www.freepint.com/gary/newscenter.htm>, which focuses on sources providing up to the minute news stories on any subject imaginable, and his Web

Audio Current Awareness Resources page

<http://www.freepint.com/gary/audio.htm>, with links to hundreds of live and recorded audio/video news and public affairs programming on the Web. He also

points to one of the largest gateways to the Invisible Web - the Librarian's Internet Index at <http://lii.org> Being human edited and indexed, it is a high quality collection of highly targeted databases that contain specific answers to specific questions. Most modern search engines claim to operate on a similar basis to the Librarian's Internet Index. Both the ordinary and advanced searches of Google do not.

An excellent article on the Invisible Web can be found at <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>.

Other notable Invisible Web resources include:

Infomine Multiple Database Search

<http://infomine.ucr.edu>

Infomine might be called an "academic" search engine, focusing on scholarly resource collections, electronic journals and books, online library card catalogs, and directories of researchers. Unlike many Invisible Web search tools, Infomine allows simultaneous searching of multiple databases.

Do not mistake an interest in the Invisible Web as a slam on the general search engines. General search tools are still 100% essential for accessing material on the Internet. A very good site to analyse the capabilities and effectiveness of search engines is <http://searchenginewatch.com/showPage.html?page=2156161>. This site also has 3 excellent articles on how search engines work – and how they can be manipulated.

The major search engines have done a creditable job of scaling with the visible Web. For the foreseeable future, however, valuable resources that are part of the Invisible Web will be beyond their reach.