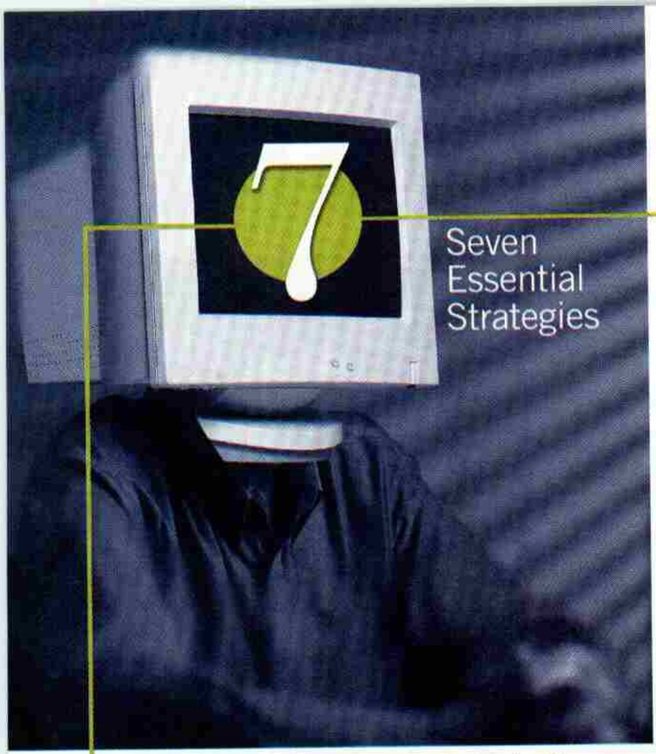

Exploring the Invisible Web

Just as the universe has two types of matter—the visible stuff of which stars are made and "dark matter" like black holes—the Web has both visible and invisible components. One of history's delicious ironies is that the man responsible for essential theories underlying search engine operations on the visible Web was also the first person to postulate the existence of celestial black holes.

By Gary Price and Chris Sherman

THIS IS LAPLACE

The French mathematician Pierre-Simon Laplace may not be the first person who comes to mind when you think about search engines and how they work. Nonetheless, much of Laplace's pioneering work in probability theory forms the core of fundamental processes search engines use to determine which documents are relevant to a query. Laplace was no one-idea savant, however. For a time, he led the Bureau des Longitudes and the Paris Observatory, where he was the first to propose the idea of black holes—collapsed stars that were so dense that even light could not escape from their gravity fields.

We searchers owe Laplace a debt of gratitude for thinking the big thoughts that help make search engines work. Unfortunately for us, however, Laplace offered no theories that could help search engines locate the Web's own form of "dark matter"—think dense repositories from which *information* cannot escape—known as the Invisible Web.

The Invisible Web is huge, and in all likelihood is growing more rapidly than the visible Web. It consists of material that general-purpose search engines either cannot, or perhaps more importantly, *will not* include in their collections of Web pages. Search engines simply cannot "see" the contents of the Invisible Web.

NOT SEEING IT ALL

Why? There are several reasons. One reason relates to the costs involved in operating a comprehensive search engine. It's expensive for search engines to locate Web resources and maintain up-to-date indices. Then there's the scourge of "spam"—Web pages that, like their unsavory email kin, are either junk or offer deceptive or misleading information. Most of the major engines have developed strict guidelines for dealing with spam that sometimes has the unfortunate effect of excluding legitimate content.

There are also technical reasons that search engines cannot index certain types of material. Search engine technology is actually quite limited in its capabilities, despite its tremendous usefulness in helping searchers locate text documents on the Web. It's difficult for search engines to process nontext information—sounds, images, streaming media, and, perhaps most importantly, information stored in Web-accessible databases.

The Invisible Web contains vast amounts of authoritative and current information that's accessible using your Web browser or add-on utility software—but you have to know where to find it, since you simply cannot access it using a search engine like Google or AltaVista.

FINDING THE INVISIBLE

So the question invariably arises: if all this great Invisible Web content is largely hidden from search engines, how do you go about finding it? Without benefit of your familiar information seeking tools, isn't the process sort of like stumbling around in a library with all the lights turned off?

Discovering Invisible Web resources can be a challenge, but it's not as hard as it may seem—in fact, once you get a feel for how to proceed, the process of exploring the Invisible Web is actually a lot of fun. In this article, we offer seven strategies we've found to be essential for successfully navigating in the Web's hidden territories. They won't help you find everything on the Invisible Web (an impossible task anyway), but will provide solid starting points and techniques for finding content that's all but impossible to find with general-purpose search engines.

1 STRATEGY 1: ADOPT THE MINDSET OF A HUNTER

The most important strategy is to stop thinking of yourself as a "searcher" and start thinking of yourself as a "hunter." Searchers are essentially passive users of information-seeking tools. Hunters use tools (weapons) but also take advantage of their environment, the weather, and knowledge of their quarry to act opportunistically whenever possible, using all manner of tactics when stalking their elusive prey. If the hunter metaphor makes you queasy, simply adopt the mindset of a passionate collector of rare books, artwork, or antiques who goes to great lengths to strengthen his or her library or collections.

What are the similarities between hunting and exploring the Invisible Web? You are always hunting for material. It never stops. You need an active mind ready to turn over every stone, or in the case of hunting for Invisible Web resources, every site you come across, ceaselessly looking for "possibilities." The most important similarity is realizing that you will never find "everything." There's always newer, bigger game out there on the Invisible Web.

What are the differences? You can do most of your "stalking" with your eyes and ears. No traveling is necessary, except for perhaps a trip to your public library to examine print publications. The importance of using your eyes is obvious, but why ears? We're so conditioned to think of the Web as a visual medium that we often forget we can find exceptional resources using means other than our browser. Listen to the media. Are new sites from authoritative organizations or people being mentioned? If so, you must personally explore them. Also, listen to those around you. Are they discussing an intriguing site? Does it sound like it has possibilities? If so, go explore it on your own.

2 STRATEGY 2: USE SEARCH ENGINES

Wait a minute—if Invisible Web content is supposedly hidden from search engines, how can AltaVista, Google, Northern Light, or their kin possibly help?

Even though Invisible Web content stored in databases is largely hidden from search engines, many have Web interfaces consisting of simple HTML pages that are perfectly visible to search engine crawlers. Once you have found a "front door," you often have full access to the riches within the database using its own internal search services.

How do you find these front pages or gateways? Run preemptive searches in general-purpose search engines using terms such as "searchable database," "interactive database," "interactive tool," "customizable database," and other similar phrases. Use the search engine's Boolean AND operator with the above terms with keywords for the topics or subjects you're interested in. Most of the results you'll turn up using this strategy will be visible Web content, but occasionally you'll hit pay dirt and discover the gateway to an Invisible Web database.

3 STRATEGY 3: DATAMINE YOUR BOOKMARK COLLECTION

You may have already discovered Invisible Web resources and added them to your bookmark collection without even realizing it. Sometimes Invisible Web content coexists on the same site with visible Web content. This happens all the time with huge sites like the Library of Congress or the World Bank. Though search engines can find a lot of the material on these sites, they completely miss the content included in databases available at the sites.

To find Invisible Web content that's cloaked within a visible Web site, adopt a datamining approach and really dig in to what the site has to offer. Look at the site map (if one is available). Does it mention a database? Explore the front page, looking for announcements of new resources. Is there a link to a statistics page? Statistics and other numeric data are often kept in Invisible Web databases. Finally, adopt the same approach using the site's internal search tool that we outlined in strategy two: search for "database" and other telltale keywords that suggest Invisible Web content.

4 STRATEGY 4: USE THE NET'S "BAKER STREET IRREGULARS"

Sherlock Holmes was widely regarded for his powers of observation and deduction. But he also relied on an extensive intelligence network he called the "Baker Street Irregulars," a motley band of streetwise urchins who could be relied upon for up-to-the-moment news of city life in London—information that Holmes couldn't possibly have discovered on his own in such a timely fashion.

The Internet has a similar group of people who serve as an exceptional "early warning system" for new Invisible Web resources. These are the people who participate in discussion lists and pride themselves on their ability to be the first to report on interesting or useful new sites. They're driven to relentlessly scour the Net in search of undiscovered resources and derive a deep satisfaction from delivering "scoops" to their peers, savoring their reputations as being way ahead of the curve.

There are no "essential" lists for the Invisible Web. We suggest that you monitor lists for subject areas in which you are interested. To find relevant lists, try the Directory of Scholarly and Professional E-Conferences (http://www.n2h2.com/kovacs), now in its 14th edition, which permits both keyword searching and browsing. Topica (http://www.topica.com) provides a one-stop shop with subscription instructions and management tools.

Some librarian-oriented discussion lists we find valuable for finding Invisible Web content include: Govdoc-L (Government Documents), Buslib (Business Librarianship), and Newslib (News and Media Librarianship). You can find subscription information and more details for these and many other lists of use to information professionals on the Library-Oriented Lists and Electronic Serials page (http://www.wrlc.org/liblists).

Another excellent list, often including discussion of new resources, is CARR-L (Computer Assisted Reporting), a list for reporters who use the Web

The 7 Invisible Web Strategies

- Adopt the Mindset of a Hunter
- Use Search Engines
- Datamine Your Bookmark Collection
- Use the Net's "Baker Street Irregulars"
- Use Invisible Web Pathfinders
- Use Offline Finding Aids
- Create Your Own "Monitoring Service"

➤

and the Invisible Web as research and background resources that is maintained at the University of Louisville.

This strategy is somewhat akin to panning for gold, hoping to turn up one nugget among the dozens of rocks you'll reject. Keeping current and monitoring for new materials is a continuous process, often taking time, energy, and yes, a bit of luck.

STRATEGY 5: USE INVISIBLE WEB PATHFINDERS

Invisible Web pathfinders are, for the most part, Yahoo!-like directories with lists of links to Invisible Web resources. Most of these pathfinders, however, also include links to searchable resources that aren't strictly invisible. Nonetheless, they are useful starting points for finding and building your own collection of Invisible Web resources.

Not surprisingly, we recommend direct search (http://gwis2.circ.gwu.edu/~gprice/direct.htm), a growing compilation of links to the search interfaces of resources that contain data not easily or entirely searchable/accessible from general search tools like Alta Vista, Google, and HotBot, kept current by Gary Price.

The InvisibleWeb Catalog (http://www.invisibleWeb.com) from Intelliseek contains over 10,000 databases and searchable sources that have been frequently overlooked by traditional searching. Also from Intelliseek is ProFusion (http://www.profusion.com), which, in addition to providing a sophisticated simultaneous search capability for the major general-purpose search engines, has direct access to the Invisible Web with the ability to search over 1,000 targeted sources of information.

The Librarians' Index to the Internet (http://www.lii.org) is a searchable, annotated subject directory of more than 7,000 Internet resources, categorized as Best Of, Directories, Databases, and Specific Resources. Databases, of course, are Invisible Web resources. The primary purpose of AlphaSearch (http://www.calvin.edu/library/searreso/internet/as) is to access the finest Internet "gateway" sites. The authors of these "gateway" sites have spent significant time gathering into one place all relevant sites related to a discipline, subject, or idea.

STRATEGY 6: USE OFFLINE FINDING AIDS

Somewhat counter-intuitively, books, magazines, journals, and trade publications can be exceptionally valuable sources for locating Invisible Web content. How so?

Every day, more and more high-quality print publications are migrating to the Web. Some are simply repurposing their content, but others are taking a more thoughtful approach, building value-added interactive tools to supplement their core content. By definition, sites using interactivity are often part of the Invisible Web. Take a look at your favorite publications—they may have recently published to the Web.

Publications that feature reviews of Web sites are also helpful for tracking down Invisible Web resources. The trick here is to not rely on the printed material to inform you of Invisible Web resources, which as we said are often hidden within visible Web sites. Databases and other Invisible Web resources often go unnoticed and unmentioned in Web site reviews. If a reviewed site seems promising, take the time to thoroughly explore it yourself. We find it remarkable how many first-rate Invisible Web resources we've discovered using this strategy.

STRATEGY 7: CREATE YOUR OWN "MONITORING SERVICE"

This strategy has two parts. The first is to identify the Invisible Web sites that you personally find the most valuable for your information needs. Monitor the "What's New" or press release pages from these sites so that you're always aware of new developments.

For example, we monitor sites like the World Bank, the U.S. Government Printing Office, and others that we consider among the "top producers" of Invisible Web content. This may seem like a daunting task, but it's easy if you use automated alerting software. We use C4U, a simple, free utility that lets you monitor as many Web pages as you like. Whenever C4U detects a change on a page you're monitoring (such as a "what's new" page), it alerts you, indicating changes in text, new images, keywords, and so on. These clues let you quickly determine whether the change is worth checking out before actually loading the page in your browser (see http://www.c4u.com for more information).

The second part of this strategy is to subscribe to and read "what's new" lists. These publications don't focus exclusively on the Invisible Web, but do turn up consistently high-quality resources, that, using the strategies we've described above, can often lead you to parts of the Invisible Web you never knew existed.

Ones that we like include Librarians' Index to the Internet New This Week (http://www.lii.org/search/file/mailinglist); Research Buzz from noted Web Researcher Tara Calishain (http://www.researchbuzz.com); Neat New Stuff I Found on the Web This Week from "Librarian without Walls" Marylaine Block (http://marylaine.com/neatnew.html); Infomine New Resources Alert from the University of California (http://infomine.ucr.edu/email); TVC Alert from legal Super Searcher Genie Tyburski (http://www.virtualchase.com/TVCAlert/subscribe.html); and The Virtual Acquisition Shelf from Gary Price (http://resourceshelf.blogspot.com).

We're also in the final planning stages for our upcoming Invisible Web newsletter, to be launched this summer (http://www.invisible-web.net).

GO FORTH, HUNTER

Exploring the Invisible Web isn't really that difficult if you're willing to invest some time and get caught up in the excitement of the hunt. The seven strategies we've outlined will help you get started on your own journey of exploration of the Web's hidden content. The bottom line, though, is that you must be ready to explore on your own to find the quality resources that are meaningful to you. Once you've built up your own virtual reference collection, finding what you're looking for on the Invisible Web will be as easy as using a search engine to navigate the visible Web.

***Gary Price** (gprice@invisible-web.net) is a reference librarian at the Gelman Library of George Washington University. **Chris Sherman** (csherman@searchwise.net) is associate editor of SearchEngineWatch.com. Their forthcoming book, published by CyberAge Books, is* The Invisible Web: Uncovering Information Sources Search Engines Can't See.

Comments? Email letters to the editor to marydee@xmission.com.