The Evolving "Invisible Web":
Tried-and-True Methods
and New Developments
for Locating the Web's Hidden Content

Robert J. Lackie

ABSTRACT. A large part of the Web is comprised of databases that only reveal their information if you are sitting on their doorstep, requesting it directly. General-purpose search engines do not find this dynamically-generated information, nor do they easily locate the many different file formats loaded on the Web. Tried-and-true methods searchers can use to obtain access to this hidden Web content, as well as recent developments in search technologies, are described and explained. [Article copies available for a fee from The Haworth Document Delivery Service: 1-800-HAWORTH. E-mail address: <docdelivery@haworthpress.com> Website: <http://www.HaworthPress.com> © 2003 by The Haworth Press, Inc. All rights reserved.]

KEYWORDS. Invisible Web, hidden Web, deep Web, databases, file formats

Robert J. Lackie (MLIS, University of South Carolina; MA in Curriculum, Instruction and Supervision, Rider University) is Assistant Professor II-Librarian, Franklin F. Moore Library, Rider University, Lawrenceville, NJ 08648 (address e-mail to: rlackie@rider.edu).

Manuscript submissions for “Internet Insider” can be submitted directly to feature editor Beth Ashmore, Reference Librarian, Samford University, Birmingham, AL 35229 (e-mail: beashmore@samford.edu).

College & Undergraduate Libraries, Vol. 10(2) 2003
http://www.haworthpress.com/store/product.asp?sku=J106 © 2003 by The Haworth Press, Inc. All rights reserved.
Digital Object Identifier: 10.1300/J106v10n02_06
INTRODUCTION

At library conferences in early 2000, the new buzz phrase among librarians was the “Invisible Web,” also known as the “deep” or “hidden” Web. Since that time, the discussions have become almost a craze among librarians, teachers, and avid Web surfers, especially as they discovered that not everything that really is on the Web could be easily located, if at all, using general-purpose search engines. Unfortunately, many searchers still rely heavily on search engines’ abilities to index the Web and assume that, if their favorite search engine(s) cannot locate it, the material must not be there. Searchers need to become familiar with that portion of the Web that is hidden from them—the Invisible Web.

WHAT IS THIS “INVISIBLE WEB” AND WHY DOES IT EXIST?

When people refer to the Invisible Web, they primarily are talking about material that popular, traditional search engines cannot or will not index. This information is stored on the Web within various non-textual file formats and free content-rich databases created by government agencies, educational institutions, and other organizations around the world. Some examples are “phone books, ‘people finders,’ patents, laws, dictionary definitions, items in Web stores or auctions [i.e., eBay], digital exhibits, and multimedia and geographical files.” In addition, this hidden Web can contain “information that is new and dynamically changing in content . . . such as news, job postings, airline flights, hotel rooms, stock prices, [and] market averages” (Cohen 2002).

One major reason this Invisible Web exists is that traditional search engines—primarily designed to index simple static HTML pages that link to other pages—do not, for technical and non-technical reasons, index many pages created “on-the-fly” in today’s dynamic modern Web sites. Most search engines can find the Web sites themselves, just not their dynamically-generated content, which can only be accessed by direct queries to the Web sites’ databases. All of this can be a bit confusing. Fortunately, articles have been published recently explaining this Invisible Web and directing searchers to tools that can uncover the hidden information within dynamic sites. A very readable article published in Teacher-Librarian last year can bring most searchers up to speed (Clyde 2002), and this author’s previous article on the topic provides an annotated listing of specific tools capable of retrieving hidden content on the Web (Lackie 2002).
WHY CARE ABOUT THIS INVISIBLE WEB?

A large portion of the Web is missing from search engines’ results pages. Even so, an immense amount of “visible” information is available, so why care about the “invisible” content? In truth, the Invisible Web is growing at a much faster rate than the visible portion, and many experts believe that the Invisible Web’s resources are often of a higher quality than those found on the visible Web (Sherman 2001). Therefore, many information seekers must change the manner in which they search the Web in order to find this growing Invisible Web material.

TRIED-AND-TRUE METHODS FOR FINDING INVISIBLE WEB INFORMATION

Once librarians have convinced Web users that the Invisible Web actually is a valuable resource for scholarly research and information retrieval, they need to then show them how to efficiently use it. Good tutorial sites and articles exist (Clyde 2002; Cohen 2002; Hermann 2002; Lackie 2003) that not only explain the Invisible Web, but also provide basic search strategies and in-depth answers to frequently asked questions. All provide links to excellent Web directories and portals (i.e., FindLaw, Infomine, and LII), which are arguably the best way for most people to find quality Web databases. Sometimes, the simple search method works well, too.

When hunting for quality information that is likely to be stored within a database, such as lesson plans or articles within a certain academic field or on a particular topic, the word that searchers need to keep in mind is “databases.” General search engines can still help searchers locate free searchable Web databases. Remember, the databases or searchable sites are not invisible to the engines, only the information they contain. So, how does a searcher get the engine to lead him or her to the right database? Simply type a broad topic or academic subject into the search block followed by the word “database” (i.e., in Google Advanced Search, type lesson plans database into the “with all of the words” search block, and two prominent lesson plan database sites will appear on the first page of the search results: AskERIC Lesson Plans and Gateway to Educational Materials). This approach produces many irrelevant hits, but it can also quickly lead to some free, valuable searchable databases on a particular subject so that direct queries can then be made. However, since the results are from a general search engine, and not from a quality directory (i.e., LII), searchers should evaluate the database and its content.
SOME RECENT DEVELOPMENTS IN DATABASE CONTENT RETRIEVAL

Tutorials and search tools listed earlier do assist in the hunt for Invisible Web material. However, with the recent exponential increase of content stored in free Web databases and file formats, Web search technology companies saw the need to develop new products and methods to efficiently capture this hidden information. In fact, many initiatives have recently received publicity for their efforts in bringing hidden content in Web databases into the open. For instance, Flipper.com (http://flipper.com), introduced by Quigo Technologies in 2002, includes Invisible Web content in its search results by directing users to dynamic HTML pages on Web sites (Ward 2002). This author has used it to find free full-text article databases, including FindArticles.com, an archive of published articles from almost 500 periodicals that is a product of a joint venture between the Gale Group and LookSmart (www.findarticles.com).

A Web directory called the Resource Directory Network (RDN), a collaboration between many educational and research organizations, was another good find. RDN links to over 35,000 deep Web sites, with material identified and organized by subject experts. Their subject portals will be featured at the 2003 Internet Librarian International Conference (www.rdn.ac.uk).

BrightPlanet, a player on the deep Web scene for quite a while with their CompletePlanet site, have recently developed a new tool to help organizations find and manage information from thousands of deep Web databases. This Deep Query Manager (DQM2) can “search more than 55,000 unique, [deep Web] specialty content sites” (www.brightplanet.com). Free demonstrations are now available, but no pricing is listed yet.

A local company on the rise in the greater Philadelphia area is NicheUSA, LLC (www.nicheusa.com) with its ZoomerOne search and retrieval products. For instance, in a “school setting, it can collect information from multiple Web sites based on educational needs and teachers’ input so students can be presented with the best material” (Fox 2003). ZoomerOne-for-scholars has gone through a successful trial at the University of Pennsylvania’s biochemistry department, and multiple ZoomerOne-for-kids are offered to schools in New Jersey with the help of the Mercer County ETTC. “Feed it a query and it goes off, looking only at reliable sites, to retrieve a folder full of information,” including deep Web databases.

Last, but not least, is an interesting project underway at the University of Buffalo that is attempting to move online catalog content—a sizable
portion of the Invisible Web—onto the open, or visible, Web. Their studies have concluded that dynamically-generated pages from their online catalog can be stored on the Web, allowing their students to use their favorite general search engines to find reliable information and materials located within their library. This “NetCatalog was released . . . as an experimental catalog for [the] university community during the spring 2003 semester” (Ludwig 2003). It will be interesting to follow up on this.

NEW DEVELOPMENTS
FOR FILE FORMAT CONTENT RETRIEVAL

As stated earlier, besides database-driven pages, there is “information on some sites presented in formats other than static HTML that many search engines cannot read. Files stored in the Adobe Portable Document Format (PDF) is an example of this, as are Adobe Postscript; Microsoft Word, Excel, or PowerPoint; and Rich Text formats,” all available through a limit feature on Google’s advanced search page since late 2001 (Lackie 2002). Many new developments regarding file format searching have occurred since that time. In a race to catch up to and even surpass Google in this area, some search services have revamped their advanced search capabilities in order to capture content from various file formats.

In late 2002, AllTheWeb (www.alltheweb.com) became the first service to search for text within Macromedia Flash (animation) files. “Google has been indexing links within Flash documents for some time, but it has not been extracting text from these files” (Sullivan 2002). AllTheWeb now indexes PDF files and digs deeper into sites to find hidden content, including content within audio, video images, Flash, Java applets, and JavaScript. A document directory depth limit and a personal home page limit were added, too, assisting AllTheWeb to be very competitive in the deep Web search race (Notess 2002).

In addition to Google and AllTheWeb, MSN Search has recently added new advanced search features to its site, “most notably, the ability to limit searches to Microsoft Office or PDF documents. Results also feature contextual descriptions, with highlighted keywords” (Sherman 2003). Like MSN Search, Singingfish (www.singingfish.com), has also won over many searchers with its new capabilities. Their “proprietary technology allows existing search engines, portals, destination sites—any Web site at all—to deliver superior multimedia search results . . . [especially] for those seeking streaming audio or video files from across the web.”
CONCLUSION

Although the Invisible Web is only one of a variety of great places to find materials searchers demand, it is quickly becoming a very important one to those seriously looking for quality information. The Invisible Web does not take the place of the many well-known databases that libraries purchase through providers such as EBSCO and Elsevier, databases which probably should be the first online destination for those searching for recently published periodical articles. Even so, not everyone has access to these large academic databases, so many searchers depend on the free Web search tools for their information needs. Information professionals need to help connect those who have been using only general search engines to the quality and rich content of the Invisible Web—including the government documents, books, curricula, etc., stored in collections within libraries, accessible through online library catalogs.

Invisible Web search tools do need to improve, but a lot of progress has recently been made through resource collaboration and advances in search technologies. Let’s hope this trend continues and more hidden content becomes revealed, because right now the Invisible Web contains billions of quality documents hidden within thousands of free databases. And with the continued exponential growth of the Invisible Web, “serious information seekers can no longer avoid the importance or quality of Deep Web information” (Bergman 2003).

QUICK BIB

REFERENCES


