South African Journal of Information Management
[DoE accredited]

Student Work Vol.3(3/4) December 2001

The invisible Web

M. van der Westhuizen
(JD Group)
Post Graduate Diploma in Information Management
Rand Afrikaans University
infosci@rau.ac.za

Contents

1 The term invisible Web

Despite its uniform interface and seamless linked integration, the Web is not a single coherent element. There are two distinct elements: the visible and the invisible Web. The visible Web consists of manually produced, static pages. It provides the same generic information to everyone and is therefore available for indexing to all search engines. The invisible Web consists of computer generated, dynamic pages and provides customized information according to specific requirements. In other words, the Web has its own form of black holes or dark matter. This refers to a dense repository of data and information, which the average search engine cannot easily detect. 'Invisible Web' is the term coined for this rather peculiar but unexplored environment. This section of the Web is massive and in all likelihood is growing faster than the visible Web.

Material invisible to or 'hidden' from the general search tools like Alta Vista and Google is said to reside on the invisible or deep Web – a vast part of the Internet that the search engines cannot, do not or will not include in their indexes of the Web. Search engines therefore simply cannot 'see' the contents of the invisible Web.

top

2 Size or scope of the invisible Web

A new study by BrightPlanet puts the size of the invisible Web at 400 to 550 times larger than the visible Web, which is currently estimated to be more than 2.5 billion pages. Much of this material is authoritative information and invaluable in that it is largely comprised of content-rich databases from universities, libraries, associations, businesses and government agencies around the world.

Many times, you will get to the front door (i.e. the home page) but you will not find the pages behind it in a 'normal' Web search – nor will you find the content behind forms and dynamic pages.

Much of the Web cannot be 'seen' using standard search engines like Google or Alta Vista. Even the biggest search engines search less than 60% of all Web pages. The remaining 40% lie hidden behind security barriers, are too deep in a Web site's hierarchy to be indexed, or require a password. There is even a larger invisible Web, according to a study found on the Search Engine Watch site, that can be mined only by using individual database portals. In fact, the study determines that only 1/500th of the information on the Web is accessible through standard search engines! The rest lies buried in databases. The *Making of America (MOA)* Web site is an example of what lies buried in the invisible Web. Through the MOA portal, a researcher can access the full text of 6600 books and 50000 journal articles, yet not a single MOA source will be found using a standard search engine

top

3 Search techniques and search facilities to gain access to and retrieve information from the invisible Web

It is clear that software developers of search engines are seeking to exploit the thorny problem of invisible Web databases that search engines cannot 'see'. The opportunity exists, because Web pages that are generated dynamically via databases are different from what are generally known as 'flat html' pages. The latter are generated, one at a time, by people using authoring tools or coding by hand and then leaving them on a server until someone requests them. Dynamically generated Web pages do not exist as separate files, so spiders from the major search engines do not generally discern them. The problem is intensifying because of the proliferation of off-the-shelf tools to link databases to the Web, whether as whole sites or as site components. This means that proportionately less and less pages are available for search engines to see.

One response to this problem has been to divide the Web into vertical sections intended to appeal to specific interests. Kapoor (1999:1) predicts that there will be an explosion of vertical search sites, providing access to deep, tightly focused databases.

Another benefit to search precision is narrowing search domains to specific subjects, accomplished by honing the scope of what information is searched, perhaps by limiting searches to certain domains or languages, or conducting specialized searches in subject oriented search engines.

Andrews (1997:2) predicts a change in how people will use the Web in future. Instead of wandering around and bookmarking what looks interesting, he says, people are already activating their Internet connections with a specific goal in mind. He continues to say that databases are listed in categories, and users choose which to search, based on brief descriptions instead of searching through them all at once.

However, Google has quietly rolled out a new feature that allows searchers to find information contained in Adobe Portable Document Format (PDF) files, effectively

revealing a significant portion of the invisible Web. While PDF files are not as abundant as the simple HTML files that make up most of the Web content, they often contain high quality information that is often unavailable elsewhere. Most of the major search engines do not include PDF files in their Web indexes, which is why they have long been considered as part of the invisible Web. Google has therefore provided a great service to the Web community in indexing PDF files. So far, they have indexed more than 13 million files, from all parts of the Web. Though they make up only a small part of the invisible Web, the generally high quality and authoritative information they provide is a boon to serious searchers.
There is a public, or 'free', Web and a private, or 'fee', Web with virtually no overlap. This is closely related to the invisible Web discussed in the previous point. The public Web contains the sites retrieved by standard search engines. The private Web contains huge databases of journal articles and books that are password protected. It's on the private side where you will find all the high-quality sources needed for a research assignment; but not a single one will be found by using Google, Yahoo, or Alta Vista.

Companies that add value to information by organizing, cataloging, and packaging it create these sites. Access to these sites is then sold to organizations, such as libraries. When one thinks about it, it makes sense that there would be a private Web. Billions of dollars are spent annually producing and selling books and journals. Why would publishers let that material flow freely on the Web? Typically, an information provider licenses campus-wide database access to a library, then all computers on that campus would access that database. For example, many of the Research Databases in the Hekman Digital Library reside on the private Web.

The bottom line: The Web *does* contain a wealth of information, it just can't be accessed using a standard search engine. To access that wealth of information, you need to enter the Web through the library's Web site. For example, students at Calvin would enter through Hekman Digital Library.

However, more help is at hand! Gary Price of George Washington University in the USA has compiled Direct Search – a regularly updated and growing compilation of links to the search interfaces of resources that contain data not easily or entirely searchable or accessible from general search tools like Alta Vista, Google or Hotbot. The Direct Search SearchCenter interface provides search access to all Direct Search pages as well as the following Web reference compilations: fast facts; price's list of lists; speech and transcript centre; news centre; streaming media; news and public affairs resources; and Web accessible congressional research service reports. Direct Access categories include archives and library catalogues, bibliographies/bibliographic aids, books (full-text), business/economics, government (US and international), government (US state and city), humanities, legal, news sources and serials, ready reference, recent additions to the collection, science, social sciences and additional subject-specific resources. It also gives access to advanced search engines like Alta Vista, Google, Fast, Yahoo, etc. Find Direct Search at http://gwis2.circ.gwu.edu/~gprice/direct.htm.

Another huge Web undertaking was the collection of links to special search engines and searchable directories that, in a number of cases, can be used as an alternative for the big search engines like Northern Light, Hotbot, Alta Vista, Excite and Infoseek. Most of them are discipline or subject specific, others are (collections of) national or regional search engines. This collection is preceded by a few sites where one may learn to search on the World-Wide Web, a collection of synonym dictionaries and thesauri (to find the right search terms), experts to answer questions and the URLs of a number of fee-based services, which offer to do the searching for you. Under the heading 'Search engine code texts', the user can find the addresses of some sites with pieces of code which can be pasted into the user's own

homepage to offer direct access. There are also directories of free bibliographies and bibliographic databases on the Web, as well as free journals and magazines on the Web. This collection of specialized search engines and databases was compiled by Marten Hofstede at the University of Leiden in The Netherlands and can be found at http://www.leidenuniv.nl/ub/biv/specials.htm.

top

4 Guide to specialized search engines to access and retrieve information on the invisible Web

Just because some Web pages are not included in a search engine's index it does not automatically make them invisible. Search engines use automated programs called 'spiders' to 'crawl' the Web and fetch them for inclusion in their search indexes. For a variety of reasons, crawling is often an incomplete and inefficient process.

Invisibleweb.com is a first-grade guide containing over 10000 search engines organized into 18 subject categories and hundreds of subcategories and subsubcategories. In spite of its enormous size, Invisibleweb.com is easy to use because of its clear and logical design. If you are short of time and would like to see just a sampling of the largest specialized search engines about a popular topic, for example, breath holding spells, you can click on a subject from the 'Hot List' and get the names of approximately 10 leading engines relating to one of these topics. Keyword searching for search engines is available and is often exceptionally effective.

Invisibleweb.com contains search engine collections for a variety of popular, general and academic topics. Surprisingly, there is no subject category for regional engines.

One of Invisibleweb.com's strengths is its detailed classification of subjects, which can reduce the time it takes to find search engines covering a specific subject. For example, under the subcategory investments, some of the subsubcategories are Bonds, Commodities, Futures and options, Mutual Funds and Stocks.

Search engine selection is generally excellent and comprehensiveness varies with the topic. Occasionally the same engine appears more than once under the subject because its different information collections are listed separately. Some categories with especially extensive search engine collections are Legal, Travel, Sciences and References. One can also choose to see an unusually full, informative description of each search engine. Search menus are displayed for a small percentage of the engines.

InvisibleWeb.com is particularly valuable and useful for writers, students, professionals, academics, subject specialists, and researchers of all kinds, as well as the average searcher looking for in-depth information about a subject. Inexperienced searchers will feel comfortable here because of the friendly design.

top

5 Selected directories of searchable databases

Table 1 is a guide, listed in ranked order for academic research purposes, that shows the different directories of searchable databases, which one can use to access and retrieve information from the invisible Web.

Table 1 Guide to directories of searchable databases for access and retrieval from the

invisible Web

Click to go to any tool:	Size and general features	Searchable or browsable?	Evaluations of databases	Search boxes
The Invisible Web Catalog ****	Large (over 10000) collection of searchable databases. Many academic subjects and audience. Easy to use because of its clear and logical design. Quick search a concept or topic. Advanced search allows Boolean and other searches. Keep searches broad. Click GO. Browse in 'Hot List' and categories expand for convenient selection. Can sort results alphabetically or by score (relevance default). Click database name for search box. Links to matching subject categories at top of results get more databases in category. Keyword searching for search engines is available and exceptionally effective. Search menus are displayed for a small percentage of the engines. Excellent help function. Lycos 'Searchable Databases' is a subset of the invisible Web catalog through partnership.	Both	Excellent Click [more] to read complete evaluation.	Yes, usually.
Direct Search ****	Mainly a scholarly search engine guide. You need a certain degree of subject knowledge to understand the material covered. Several long pages listing and describing searchable databases on many academic topics. Subjects covered	Both, but not a searchable database. Search feature new in fall 2000 tells which page to look	Yes, from academic librarian perspective.	None.

	range from Biochemistry to Government to Humanities. Excellent collection of lists of data, many of which contain information about public and private companies. Pick the section or page from the links near the top. Done by an academic librarian with research in mind. Especially useful for academics, subject specialists and, to some extend, business researchers.	in. Use Ctrl+F to find term in page.		
Internets ***	Large (# not specified) collection of searchable databases (search engines). Also selected Web sites, often of academic interest. Especially useful for subject specialists, writers, academics, researchers, students and professionals. They may discover new search engines relating to their field of interest using this guide. Search a concept or topic. Keep searches broad. For many subjects, one can get a more complete list of search engines than any other guide. Keyword searching for search engines is available and very useful. Categories are divided into subcategories and in turn divided into many highly specific subcategories, for example, cattle databases or scuba diving databases. Use the term database interchangeably with search engine. Two somewhat confusing	Both	None	Rarely Called 'In Line Databases'.

	results are displayed: 1. If boldface, numbered, and ranked by % score, results are subject sub-categories. Pick the most appropriate category to view searchable list of databases. 2. If bulleted list not boldface, you have the searchable list of databases. Click on title to go directly to site. Click on Web sites at top for selected Web pages, often of high value in academic research.			
IncyWincy	Large (claims 100000 but few are databases). Collection of Web pages, directories and some searchable databases drawn from the DMOZ Open Directory Often search boxes are not linked to the contents of the page but to some other database (like Amazon.com). Use specific terms: supports AND, OR, and NOT, "" and *. Use them because it is a whole subject directory.	Both Can search top results sometimes in second box.	Brief descriptions.	Some, but unreliable. Is supposed to identify and link to a search box on the page. Submits one's terms in the search engine (not useful).
Collection of Search Engines ***	Fairly large, well designed and easy to use. Lacks flexibility when using list of searchable databases. List begins after a number of other links to topics related to searching. Scroll down the bar on the left to find the subjects with searchable databases. Subjects are arranged	Browsable with difficulty.	Some.	None.

	alphabetically rather than by broad category. Keyword searching for search engines not available. 40% of subjects covered are academic, the others are general or Web related. Links at top are often academic but not to searchable databases.			
Complete Planet **	Large database of searchable databases, Web pages with search boxes (not databases), and mere Web pages. Although the site speaks eloquently of the 'deep' (their term for invisible Web), many of the links are to 'visible' or 'surface' Web. Hard to know which are databases. Search using " " around phrases, and Boolean operators (complete set). Stems. Simple searches often retrieve too many documents. 'Categories' link at end of entry displays subject classifications assigned for easy access to more in a category.	Search, then use 'category' links at each entry to browse.	No evaluations. Some descriptions, some strings of keywords, some extracts from the page.	None.
Search Power **	Almost 14000 of the sites are city and state guides. Excellent resource for these. Subjects covered are popular and general topics. Useful for the general searcher who is looking for information about cities and states in the U.S.A. Some valuable academic content in the rest.	Both.	Descriptions seem to be extracted from or supplied by the sites. No evaluations by Search Power.	None

	Too tiny search box in upper left usually retrieves a more reasonable search area. Select the Search type or 'Select a Search Engine' (not the default) and resubmit your search – annoying double work! Results are a combination of subject categories with your terms and sites/descriptions. Click a 'category' for more in the category (recommended), or click a database name to search it. More options: search (at bottom of results) allows combining terms and phrase search.			
Internet Oracle **	Medium/large search engine guide. Focuses on popular and general subjects. Especially useful for finding a wide variety of search engines that cover popular subjects. Graphically attractive, extremely clear and well designed. Colourful subject groupings of often-useful searchable databases. Short lists of important Web sites are included for many topics. Keyword searching for search engines is not available. Some niche categories for example, women, gay and lesbian, and others can have useful academic value.	Browse by selecting icons or links at left.	None.	Yes. Search box sometimes has a drop-down menu.
Special Search Engines **	From the Nanyang Technological University Library in Singapore. Academic focus and Asian/Pacific emphasis.	Browse only.	Brief descriptions, with infrequent evaluations.	None.

	Its over 1000 search engines are divided into 25 categories. The regional category contains the largest number of search engines. Alphabetical subject categories of a selection of searchable databases by subject start after the international general Web search engines. Especially useful for anyone doing business with companies located in Japan, Singapore, or China, or with firms in other Eastern or European countries.			

Ranked as: ** Very useful for academic research, *** Useful for academic research, ** Less useful for academic research**

In addition, apart from Invisibleweb.com and the others mentioned in Table 1 other searchable directories are listed in Table 2.

Table 2 Searchable directories and their usefulness

Fossick.com – Interesting mixture of popular, academic, general, and Internet-related search engines. Handy for the general and academic searcher.
WebData – Lists sites that are mainly commercial along with the search engines. Some use for researchers, professionals and general searchers. Rather select a search engine than a commercial site.
Beaucoup – Oldest specialized search engine guide. Useful for the average searcher who wants to find many different aspects of a subject.
SearchIQ – Covers all types of subjects. Useful for general searchers and, depending on the subject, people doing research on the Internet.
MetaIQ.com – Contains mainly popular and general specialized search engines. Useful for the general searcher.
Virtual Search Engines – Offers the general searcher a good variety of search engines specializing in some professional subjects, such as legal and health search engines. Useful for an introduction to a subject.
About.com - Web Search – Wide range of subject categories. Emphasis on popular and Internet-related topics. Includes a small but useful collection of academic engines relating to science, the arts and the humanities. Beginners in searching will find it useful because of its general search information and advice, and clear design.
Search Engine Guide – All types of subjects are covered. The business category contains engines and directories pertaining to various businesses. Very useful to

the general searcher.
FinderSeeker's – Strength lies in its ability to search for search engines about a topic from a specific country, for example, legal search engines from Australia. Also lists engines from individual cities and states in the USA.
SearchBug.com – Useful for Internet beginners or inexperienced searchers in finding a small but high-quality collection of search engines about commonly searched for subjects. An unusual category is packages, which includes search engines concerned with package tracking and drop-off locations, for example, FedEx or UPS.
AllSearchEngines – Popular and general subjects make up the majority of topics. There is a wide difference in the quality of search engine selection for different subjects, with business and government-related subjects covered comprehensively.
Search Engine Colossus – The collections of general and specialized search engines from some of the larger countries (particularly the USA) are extensive. Useful when looking for search engines originating in specific countries.
Search Engines Worldwide – Search engines from countries of every size all over the world are included. Useful for finding information originating in various countries.
My Search Engines – Part of Reference.com, a general directory and reference site. Mostly popular topics. Useful for searchers who want to look at just a few search engines in a subject category.
The Ultimate WWW Search Engine Collection – Only popular subjects. Search engine selections are small but useful. For searches who want a simple guide with a fairly small selection of search engines.
Little-Red-Schoolhouse Library – Specialty Search Engines – Subject categories are especially designed to appeal to children's interests or that are relevant to their schoolwork, for example, SchoolHelp and Just-4-Kids.
ZeekSearch – Valuable specialized search engine guide that accesses search engines especially useful to high school, junior-high school and older elementary school students.
Kids Search Tools – Useful specialized search engine guide for children, particularly ages 7 through 12.
TekMom Search Tools for Students – Specialized search engine guide for students from elementary school through high school.

top

6 Links to a few selected sites about searchable databases and the concept of the 'invisible Web'

- SearchAbility. Descriptions of many directories and lists of searchable databases, extensively annotated, rated, and described. Excellent background on specialized searchable databases on the Web.
- The Invisible Web Revealed and The Invisible Web Gets Deeper

top

7 References

Andrews, W. 1997. *Challenges for spiders: searching invisible Web*. [Online]. Available WWW. http://www.internetworld.com/print/1997/02/03/industry/spiders.html.

Kapoor, J. 1999. *Web search engines*. [Online]. Available WWW: http://yallara.cs.rmit.edu.au/~achatter/search-engines/future.htm.

ISSN 1560-683X
Published by InterWord Communications for the Centre for Research in Web-based Applications, Rand Afrikaans University