– DRAFT –
Pre-print of Article to Appear in Texas Library Journal
– Do Not Cite --

The Deep Web: Resource Discovery in the Library of Texas

By Kathleen R. Murray and William E. Moen

(Note: URLs for all referenced Websites are listed at the end of this article.)

Introduction

The networked information environment is broad and deep. It includes websites, documents, databases, library catalogs, images, organizations, and much more. While users travel this landscape using a variety of tools, a common component is a Web browser that interacts with resources. A key challenge is understanding the limits and capabilities of tools that make visible the wealth of resources in this networked environment.

In recent years, the concept of the deep Web has helped people understand there is more than meets the eyes of their favorite search engine. New tools in the form of metasearch applications have begun to make available those resources not accessible to general-purpose Web search engines. The new resource discovery service of the statewide virtual Library of Texas (LOT) is an example of a metasearch application that allows users to discover and use resources not typically available through a general-purpose Web search engine.

Images of outer space and oceans are often used to illustrate the "deep" versus the "surface" Web. From the perspective of general-purpose search engines such as Google, LookSmart, or Ask Jeeves, whose crawlers are either unable or instructed not to index deep Web content, the deep Web is hidden or invisible. Therefore, its treasures are not identified in response to the 600 million or more Web searches performed on the average day from the most popular search engines.

Among the many resource collections in the deep Web are online library catalogs and licensed commercial resources such as the TexShare databases. Use of these resources is possible if someone first knows where and how to access the catalogs or databases and, in the case of the some commercial databases that are licensed for a particular set of users, if someone has authorization for access either within a library or via the Web. The LOT addresses discovery of some of the deep Web resources by offering users a central point from which to search online library catalogs and TexShare databases.

But what exactly is the deep Web? The remainder of this article will introduce the concept of the deep or invisible Web, provide some directory resources to deep Web content, and describe the LOT resource discovery service.

It's about:

- Searching the Web.
- Discovering content that is not accessible from general-purpose search engines.
- Making that content easy to get.

The Deep Web

A quick search for information about the deep Web is certain to retrieve the 2001 white paper reporting the results of the DeepPlanet study that quantified and characterized the content in the surface versus the deep Web (Bergman, 2001). The study estimates that the deep Web is 500 times larger than the surface Web and contains higher quality resources. The number of deep websites is estimated at 200,000 and characterized by rapid growth. Additionally, a full 95% of deep Web content is publicly accessible, requiring no subscriptions or licenses.

While there is general agreement that the deep Web is home to a wealth of high quality and primary source data, there is not agreement as to its size relative to the surface Web. Sherman (2001) asserts that the DeepPlanet study is flawed in its measurement techniques and that the deep Web is actually 2 - 50 times larger than the surface Web. Regardless of measurement approach, there is no doubt that the deep Web is much larger than the surface Web trolled by the major search engines and is believed to be growing rapidly (Bergman, 2001). Figure 1 graphically illustrates both the differential size and growth rates attributed to the surface versus the deep Web.

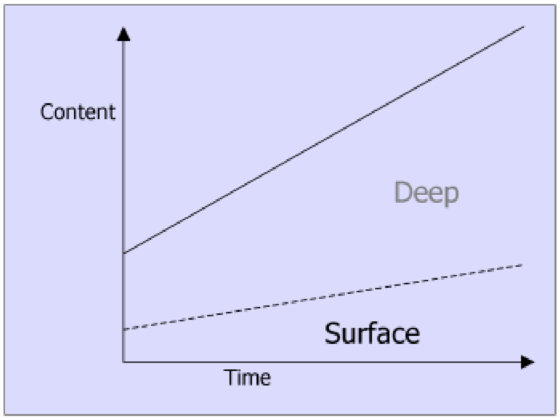


Figure 1. Web Content Growth Rate

The deep Web is principally defined by its content, that is, what resources it includes. An often-cited reference on the deep Web is a book by Gary Price and Chris Sherman called *The Invisible Web: Uncovering Information Sources Search Engines Can't See* (2001). They state that the "Invisible Web consists of material that general-purpose search engines either can not, or perhaps more importantly, will not include in their collections of Web pages".

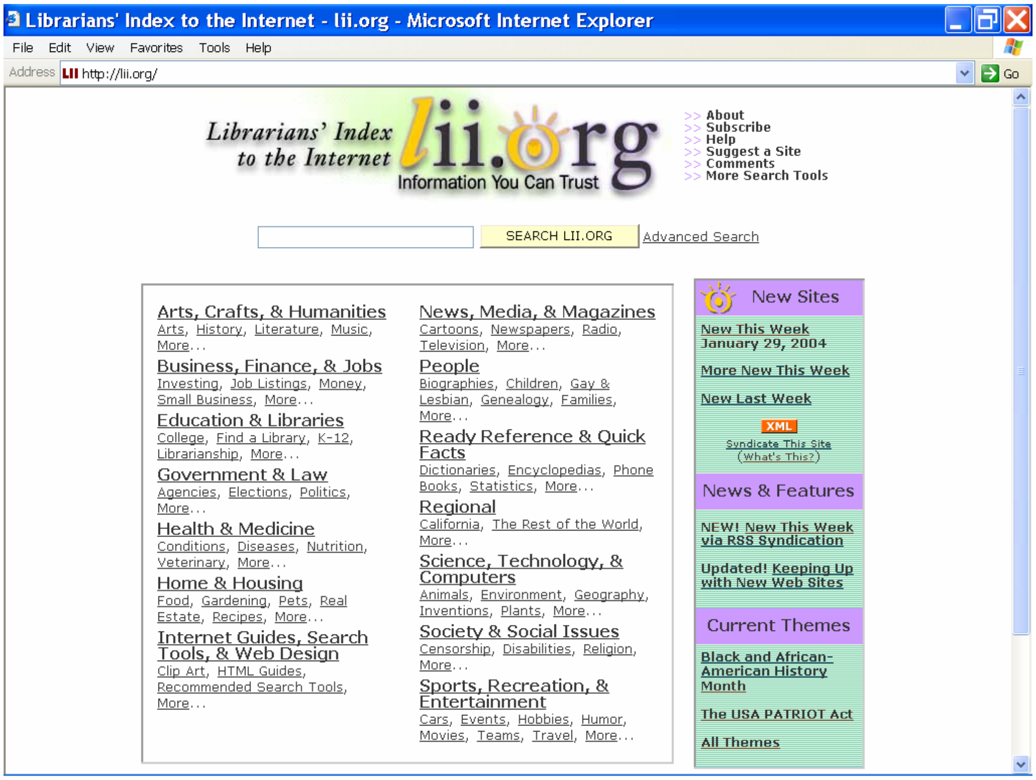
While the terms 'invisible Web' and 'deep Web' are frequently used interchangeably, some protest that 'invisible' is an inaccurate term. They argue that since most of the vast resources in the deep Web are in fact accessible and available at no cost to users, they are not invisible, save to users who limit their search behavior to general-purpose search engines. Bergman (2001), reflecting this view, asserts that: "Surface Web content is persistent on static pages discoverable by search engines through crawling, while deep Web content is only presented dynamically in response to a direct request." Simply put, the deep Web is that part of the Web not indexed by general-purpose search engines and therefore not visible to users of those search engines. Specifically, deep Web content resides in ***databases*** and is retrieved dynamically in response to specific user queries.

These databases are accessible as Web resources and users interact with them through standard Web browsers. For example, many library catalogs currently offer a Web interface to search and retrieve records. Similarly, TexShare databases are available through Web interfaces provided by individual database vendors. While there may be a static webpage for a user to input search criteria or otherwise interact with the database behind that webpage, simple indexing of that page by a general-purpose search engine does not reveal the richness of the database resource.

There are two general types of database content in the deep Web: content freely available to the public (e.g., the Government Printing Office databases) and content available only to paid subscribers (e.g., the TexShare databases). In either case, general-purpose search engines are either unable or unwilling to access and index the information in these databases. For the average Web searcher, this means that the richness of resources available in the deep Web is rarely discovered.

In contrast, librarians, educators, and professionals from all walks of life routinely use deep Web databases in the course of their jobs – without necessarily understanding that these database resources exist in the deep Web and are not revealed through general-purpose search engines. For example, librarians may search the Library of Congress catalog databases; educators may identify classroom resources via the ERIC database or the Gateway to Educational Materials; and various professionals may access government information via the Government Printing Office databases. Each of these databases contains information resources that professionals trust to be authentic and valid.

This concept of "trusted resources" is echoed in the slogan of the Librarian's Index to the Internet (LII): "Information You Can Trust." The LII directory is compiled by librarians and organized by topic (see Figure 2). It includes a searchable index to over 10,000 resources, including deep Web databases. Closer to home, the TexShare databases provide a wide range of trusted resources used by librarians and end users alike.


Librarians' Index to the Internet - lii.org - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://lii.org/
Go
Librarians' Index to the Internet
lii.org
Information You Can Trust
>> About
>> Subscribe
>> Help
>> Suggest a Site
>> Comments
>> More Search Tools
SEARCH LII.ORG
Advanced Search

Arts, Crafts, & Humanities
Arts, History, Literature, Music, More...

Business, Finance, & Jobs
Investing, Job Listings, Money, Small Business, More...

Education & Libraries
College, Find a Library, K-12, Librarianship, More...

Government & Law
Agencies, Elections, Politics, More...

Health & Medicine
Conditions, Diseases, Nutrition, Veterinary, More...

Home & Housing
Food, Gardening, Pets, Real Estate, Recipes, More...

Internet Guides, Search Tools, & Web Design
Clip Art, HTML Guides, Recommended Search Tools, More...

News, Media, & Magazines
Cartoons, Newspapers, Radio, Television, More...

People
Biographies, Children, Gay & Lesbian, Genealogy, Families, More...

Ready Reference & Quick Facts
Dictionaries, Encyclopedias, Phone Books, Statistics, More...

Regional
California, The Rest of the World, More...

Science, Technology, & Computers
Animals, Environment, Geography, Inventions, Plants, More...

Society & Social Issues
Censorship, Disabilities, Religion, More...

Sports, Recreation, & Entertainment
Cars, Events, Hobbies, Humor, Movies, Teams, Travel, More...

New Sites

New This Week
January 29, 2004

More New This Week

New Last Week

XML
Syndicate This Site
(What's This?)

News & Features

NEW! New This Week via RSS Syndication

Updated! Keeping Up with New Web Sites

Current Themes

Black and African-American History Month

The USA PATRIOT Act

All Themes

Figure 2. Librarian's Index to the Internet

Search Engines and Spiders

There are technical reasons as well as organizational policies that dictate why search engines do not identify deep Web content in their databases. General-purpose search engines consist of very large searchable databases or indexes compiled by spiders or crawlers. Spiders are programs that move from webpage to webpage throughout the Web following hyperlinks. Spiders carry out instructions to either ignore or retrieve for indexing the webpages they encounter. Typically, they are limited in their activities to static webpages and have no ability to “crawl” through the databases existing in the networked information environment.

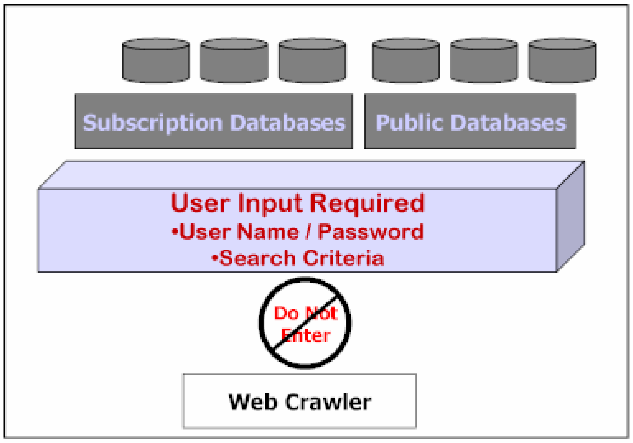


Figure 3. Database Inaccessibility to Web Crawlers

As Figure 3 depicts, because web crawlers possess neither the inclination, the intelligence, nor the information necessary to input search criteria or authentication data, they cannot access database resources. Their activities are limited to static webpages hence they do not include any dynamically generated webpages. Additionally, as a matter of policy, they are generally instructed to ignore any dynamic webpages they do encounter in order to avoid "spider traps" that could result in endless loops and ineffectively engage a spider for long periods of valuable indexing time. For these reasons, real-time information and high-quality, authoritative content resulting from deep Web database queries are generally excluded from general-purpose search engines. Table 1 summarizes content not typically indexed by these search engines.

Table 1. Content Excluded from General-Purpose Search Engines

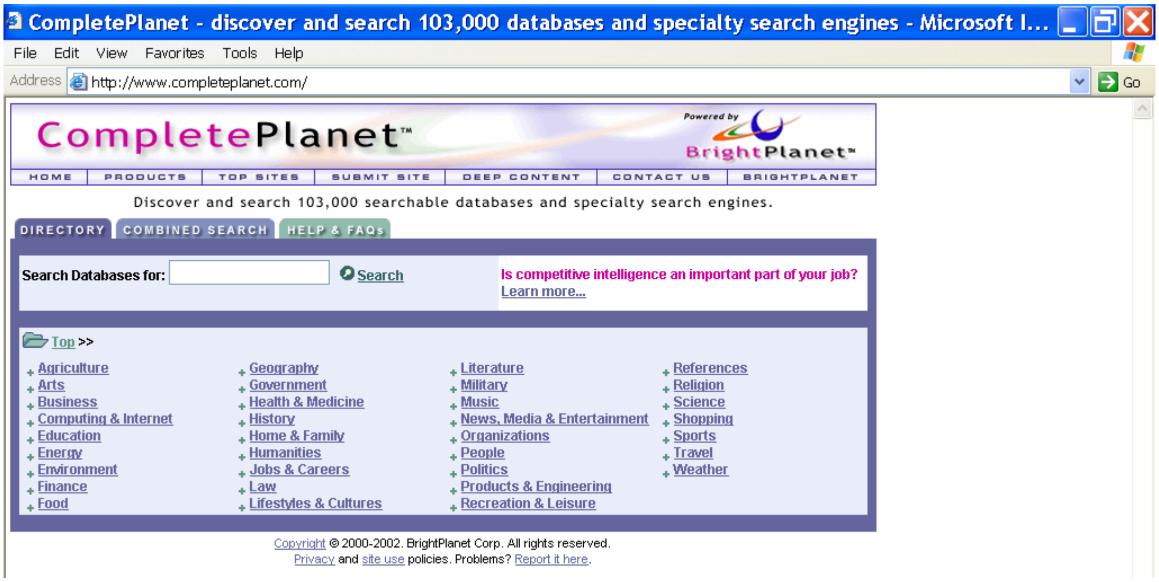
Dynamic Webpages	Created by search term input (e.g., keywords or authors) or in response to real-time information parameters (e.g., weather information and stock quotes)
Restricted content	Licensed databases
File formats that cannot be parsed by spiders	Webpages primarily containing media content with little HTML (e.g., Shockwave or certain image formats)
Policy exclusions	Webpages resulting in part or whole from script execution (e.g., identifiable by a question mark in the URL); or pages that include a robot exclusion tag

Finding Deep Web Resources

Certainly not everyone needs a deep Web database to satisfy his or her information need all of the time. Many information needs are competently addressed by general-purpose search engines and specialized or hybrid resource discovery tools. It is important to match the information need to the search tool.

Match the Search Tool to the User Need

While information professionals and general users alike can be frustrated with the high volume and low quality of resources returned in Web searches, general-purpose search engines do a respectable job of retrieving resources for specific information needs as long as the content required is not contained in databases. Likewise, topic specific portals and directories contain selected lists of high quality content that may well meet a given user need. Deep Web databases are yet another alternative for finding resources not accessible by other means. Directories of deep Web databases assist users in locating database resources that match their needs. Two examples of directories of deep Web databases are CompletePlanet and Invisible-Web.net (see Figures 4 and 5).

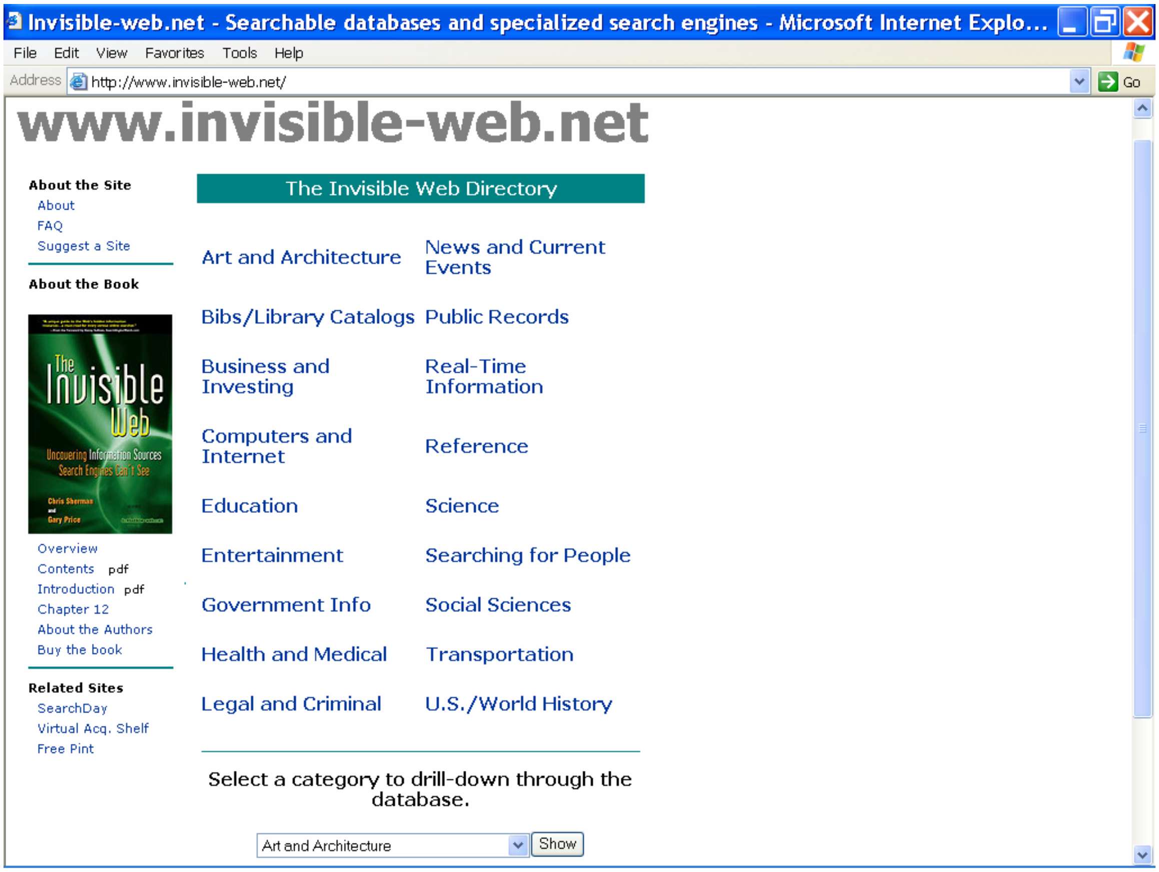


Discover and search 103,000 searchable databases and specialty search engines.

Is competitive intelligence an important part of your job? Learn more...

Copyright © 2000-2002. BrightPlanet Corp. All rights reserved.
Privacy and site use policies. Problems? Report it here.

Figure 4. CompletePlanet: A Searchable Deep Web Directory



Select a category to drill-down through the database.

Art and Architecture

Figure 5. The Invisible-Web Directory: A Browsable Deep Web Directory

Finding the content in the deep Web first involves finding websites or specific pages that provide access to the deep Web databases. There are two methods. The first is to specify a subject term and append 'database' to it at a general-purpose search engine. For example, if you would like to identify databases related to mammals, you could search Beaucoup! with the phrase "mammal database." The second method is to use a directory of databases and/or specialized search engines, which are also called pathfinders.

In both methods, users perform at least two searches. The first search identifies relevant databases and subsequent searches are directed to the selected databases. The number of searches depends on the number of databases selected. Typically, users will search the databases one at a time through different interfaces specific to the databases (see Figure 6).

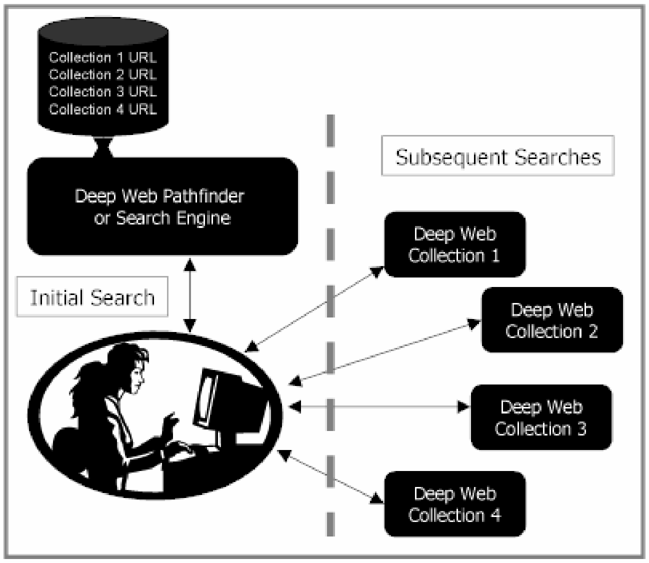


Figure 6. Identifying and Searching Deep Web Databases

Deep Web database directories identify and organize databases, often by grouping databases into subject categories. The directories can be searched and/or browsed, and users can link directly to specialized databases. Many of the databases are free, while some require registration and have a fee structure for access and/or resource acquisition. To get you started, Table 2 lists a sampling of deep Web database directories.

Table 2. Deep Web Database Directories	
CompletePlanet http://www.completeplanet.com	A searchable database of 103,000 deep websites and about 11,000 surface Web search sites.
Profusion http://www.profusion.com/	A searchable database of over 10,000 databases and specialized search engines.
Invisible-Web.net http://www.invisible-Web.net	The companion Website to Sherman and Price's book on the Invisible Web. It is a “browsable” directory of deep Web databases.
SearchAbility http://www.searchability.com	Provides a directory and guide to thousands of specialized search engines. The site describes the scope, organization, and usefulness of each of the directories.

Metasearch Engines

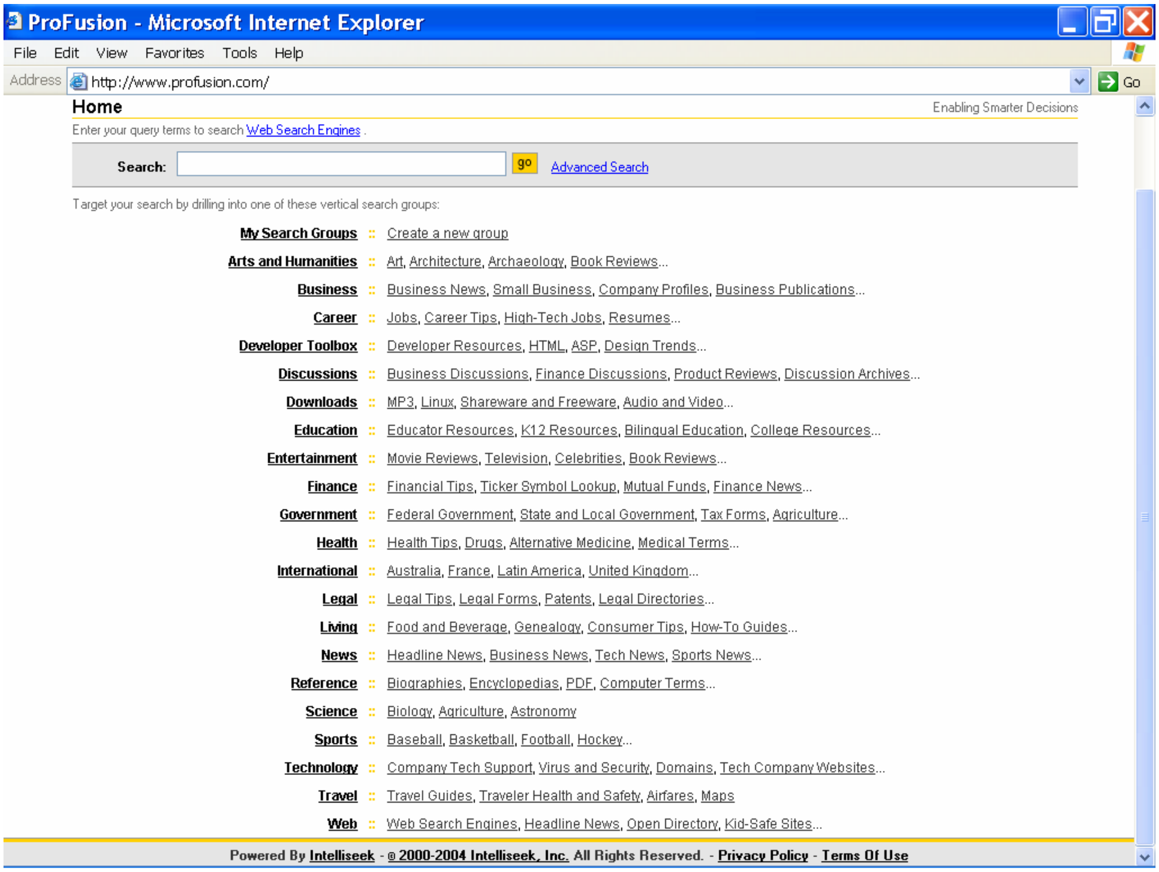
The concept of “metasearch” has been around for a number of years and describes applications that allow a user to submit a search and have that search executed by different Web search engines. Such metasearch applications eliminate the need to perform multiple database searches using multiple interfaces. Unlike Web search engines (e.g., Google or AltaVista), metasearch applications do not build searchable resource databases using crawler technology. Instead they send search criteria entered at their site to several Web search engines and present the results either separately by search engine or in some ranked relevance order. The surface Web has a number of metasearch applications (e.g., Vivisimo and Kartoo).

While there are efficiencies to be gained using metasearch applications, there are tradeoffs. In particular, the richness of query syntax (e.g., Boolean logic or phrase searches) can produce useful results from search engines that support advanced search queries but irrelevant results from search engines that do not. Table 3 identifies additional tradeoffs of metasearch applications.

Table 3. Tradeoffs of Meta Search Engines

	Number of Interfaces	**Functionality**	**Difficulty**
Individual Search Engines	Multiple	Richness of native interface	Harder to master multiple interfaces
Metasearch Engines	Single	Subject to the search functionality available in the Web search engines it passes the search to; Inconsistent parsing of search criteria	Easier to master a single interface

Recently, metasearch applications that provide a single search interface to deep Web database resources such as the Internet Movie Database and the U.S. National Archives and Records Administration (NARA) databases have emerged. One example is ProFusion, a metasearch application that includes deep Web databases as well as surface Web databases (e.g., the databases of Google or LookSmart). Deep Web databases within ProFusion are topically grouped. Utilizing search criteria entered by a user, ProFusion searches each of the databases within a user-selected main or sub-topic (see Figure 7). Results can be viewed by relevance or source.


ProFusion - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://www.profusion.com/
Go

Home

Enabling Smarter Decisions

Enter your query terms to search Web Search Engines.

Search: go Advanced Search

Target your search by drilling into one of these vertical search groups:

My Search Groups	Create a new group
Arts and Humanities	Art, Architecture, Archaeology, Book Reviews...
Business	Business News, Small Business, Company Profiles, Business Publications...
Career	Jobs, Career Tips, High-Tech Jobs, Resumes...
Developer Toolbox	Developer Resources, HTML, ASP, Design Trends...
Discussions	Business Discussions, Finance Discussions, Product Reviews, Discussion Archives...
Downloads	MP3, Linux, Shareware and Freeware, Audio and Video...
Education	Educator Resources, K12 Resources, Bilingual Education, College Resources...
Entertainment	Movie Reviews, Television, Celebrities, Book Reviews...
Finance	Financial Tips, Ticker Symbol Lookup, Mutual Funds, Finance News...
Government	Federal Government, State and Local Government, Tax Forms, Agriculture...
Health	Health Tips, Drugs, Alternative Medicine, Medical Terms...
International	Australia, France, Latin America, United Kingdom...
Legal	Legal Tips, Legal Forms, Patents, Legal Directories...
Living	Food and Beverage, Genealogy, Consumer Tips, How-To Guides...
News	Headline News, Business News, Tech News, Sports News...
Reference	Biographies, Encyclopedias, PDF, Computer Terms...
Science	Biology, Agriculture, Astronomy
Sports	Baseball, Basketball, Football, Hockey...
Technology	Company Tech Support, Virus and Security, Domains, Tech Company Websites...
Travel	Travel Guides, Traveler Health and Safety, Airfares, Maps
Web	Web Search Engines, Headline News, Open Directory, Kid-Safe Sites...

Powered By Intelliseek - © 2000-2004 Intelliseek, Inc. All Rights Reserved. - Privacy Policy - Terms Of Use

Figure 7. ProFusion: A Deep Web Metasearch Site

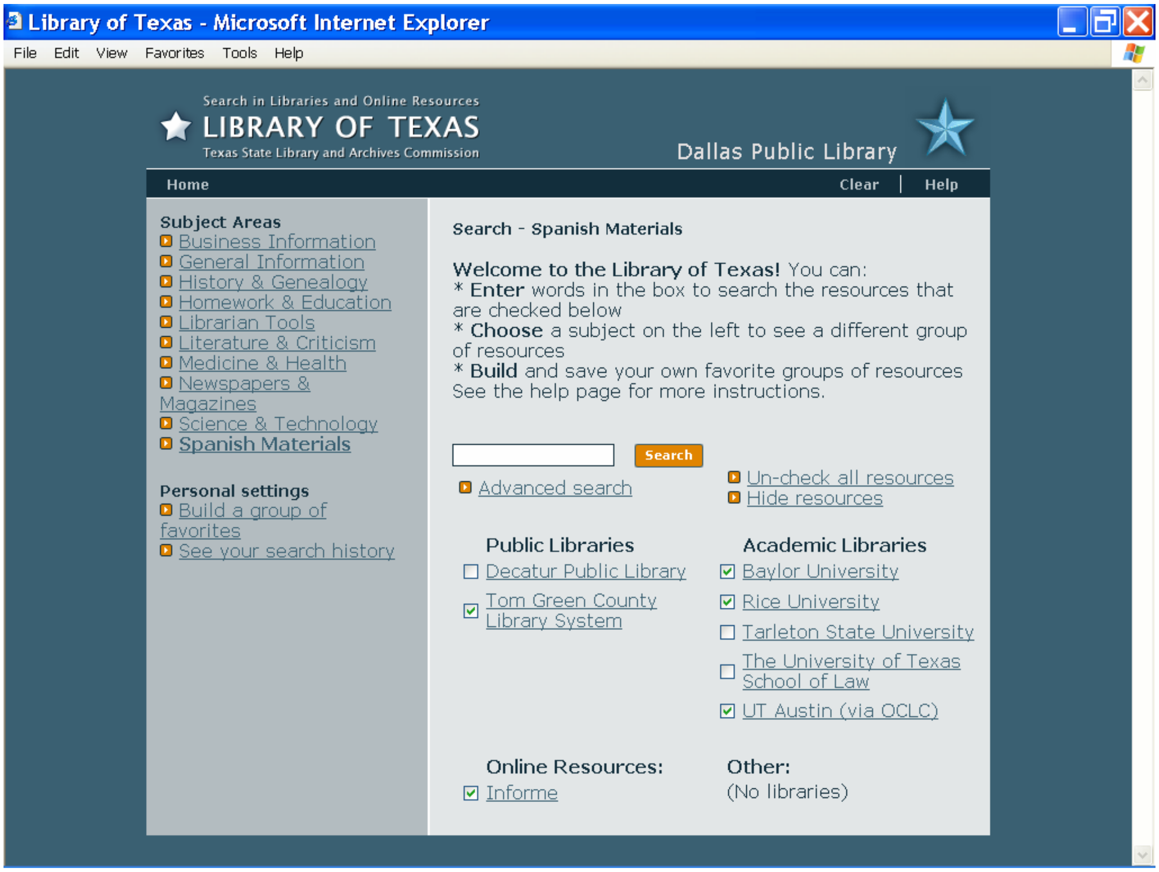
In the past several years, we have seen the emergence of metasearch applications for discovering the deep Web resources in library catalogs and commercial databases. Typifying this metasearch application is the resource discovery service of the Library of Texas.

Library of Texas

The Library of Texas (LOT) is a statewide virtual library. Although the LOT does not "own" collections of resources, it provides services to help connect users to information, and more specifically, to extend the reach and range of Texans to resources not visible to general-purpose Web search engines. Currently the LOT enables access to:

1. Texas library catalogs from over 60 public and more than 25 academic libraries
2. Forty TexShare core databases licensed for statewide access by the Texas State Library and Archives Commission

The Library of Texas is a deep Web metasearch application. Similar to other metasearch applications, the LOT topically organizes its library catalogs and commercial databases. (See Figure 8.)


Library of Texas - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Search in Libraries and Online Resources
LIBRARY OF TEXAS
Texas State Library and Archives Commission
Dallas Public Library
Home
Clear | Help
Search

Subject Areas
- Business Information
- General Information
- History & Genealogy
- Homework & Education
- Librarian Tools
- Literature & Criticism
- Medicine & Health
- Newspapers & Magazines
- Science & Technology
- Spanish Materials

Personal settings
- Build a group of favorites
- See your search history

Search - Spanish Materials

Welcome to the Library of Texas! You can:
* **Enter** words in the box to search the resources that are checked below
* **Choose** a subject on the left to see a different group of resources
* **Build** and save your own favorite groups of resources
See the help page for more instructions.

- Advanced search
- Un-check all resources
- Hide resources

Public Libraries	Academic Libraries
□ Decatur Public Library	☑ Baylor University
☑ Tom Green County Library System	☑ Rice University
	□ Tarleton State University
	□ The University of Texas School of Law
	☑ UT Austin (via OCLC)

Online Resources:	Other:
☑ Informe	(No libraries)

Figure 8. The Library of Texas: A Texas Deep Web Metasearch Site

From a single search interface, users can submit a search to multiple Texas library catalogs and TexShare databases. Users enter their search criteria and search terms just one time, and the LOT resource discovery service application translates the search parameters appropriately for each database or online catalog (i.e., the search targets). The need for translations between the LOT application and many of the search targets is mitigated by the implementation of a standard search and retrieval protocol, Z39.50.

Since the metasearch application sends the search to the various search targets, it focuses on common functionality implemented by the licensed database provider (e.g., EBSCO) and the library automation system vendor (e.g., SIRSI). The LOT metasearch application offers users basic keyword searching; author, title, and subject keyword searching; Boolean searching; and other advanced search features. As a resource discovery service, the LOT is intended to reduce the barriers to database searching and information access. Users do not need to learn the native interfaces of the TexShare databases or online catalogs to start searching. While the LOT does not replicate the functionality available through the native interfaces, it does provide links to them. This allows users to move from the common, single LOT interface to a vendor's native interface in order to take advantage of any additional search functionality.

Because the TexShare databases are licensed resources, the LOT has a login procedure that authenticates users and enables access to licensed content. Users log in once and are authorized to interact with the LOT search interface, access full text materials from a TexShare database, or link to a TexShare database vendor's native interface.

Complementing general-purpose Web search engines, the LOT provides a new level of access to deep Web resources. It addresses:

- Resource quality concerns with general-purpose search engines by providing access to authoritative and trusted databases
- Multiple interface issues by providing a single interface to the range of its databases
- A range of users by providing three interfaces:
  1. Simple public interface to free resources
  2. Texas library patron interfaces to free and subscription resources supporting both simple and advanced searching
  3. Native database interfaces

Evolution and Growth in the Distributed Information Landscape

Deep Web content has life-like qualities of movement and growth. Over time, the application of new technology is making more and more deep Web content visible to the users of general-purpose search engines. Resources that could not be indexed by Web crawlers yesterday can be indexed today. For example, document files such as PDF and MSWORD files were previously excluded from general-purpose search engine indexes but are now routinely included.

General-purpose search engines continue to add non-HTML formatted files to their indexes. Google's advanced search interface allows users to search for specific file formats, including postscript (.ps), rich text format (.rtf), Microsoft PowerPoint (.ppt), and Microsoft Excel (.xls). In its advanced searches, AlltheWeb allows inclusion or exclusion of several file formats, including Macromedia Flash files (.swf). As these previously non-indexed file formats become accessible via general-purpose search engines, the relative proportion of Web resources in the surface Web grows. (See Figure 9.)

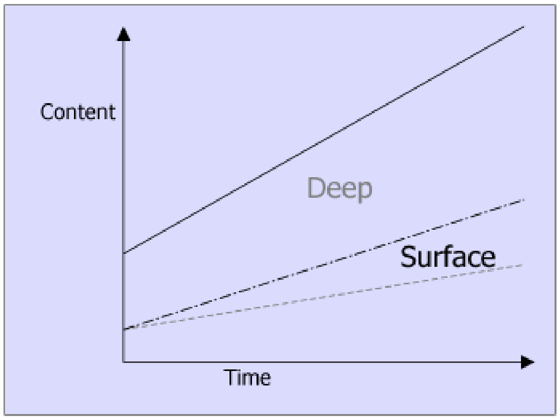


Figure 9. The Changing Nature of the Surface Web

Google is currently beta testing a search interface to online merchandise that enables users to research products prior to purchase. The application is called Froogle and is accessible from the Google advanced search page as well as its own Web location. While not directly searching remote databases, Froogle searches its own database, which is developed either by Google's Web crawlers or by vendor data feeds. In the latter case, the vendor information can be updated daily. Froogle begins to approximate real-time access to deep Web information and represents a blurring of the exclusive label of general-purpose search engine to Google.

Metasearch applications, including the Library of Texas, provide new visibility to deep Web resources not available through general-purpose Web search engines. The technical infrastructure of the LOT provides a solid foundation for future growth. While it currently makes visible the depth of Texas libraries' standard collections and TexShare databases, the LOT resource discovery service can make visible the many special collections housed in libraries and museums throughout the state that are not now readily accessible. The LOT will also be able to interact with specialized search systems that focus on a specific subject such as the Texas Records and Locator (TRAIL) service of the Texas State Library and Archives Commission and the Portal to Texas History Project at the University of North Texas Libraries. Based on an open-systems platform and utilizing standards-based information retrieval protocols, the LOT has a solid growth path to enable increasing accessibility to the rich resources in the deeper Texas Web.

Kathleen R. Murray is president of KRM Consulting and an associate fellow of the Texas Center for Digital Knowledge. Previously she was project manager for the ZLOT Project. William E. Moen is an associate professor at the School of Library and Information Science and a fellow of the Texas Center for Digital Knowledge at the University of North Texas.

References

Bergman, M.K. (2001). “The deep Web: surfacing hidden value.” *Journal of Electronic Publishing*, 7(1). Available from: http://www.press.umich.edu/jep/07-01/bergman.html.

Sherman, C. (2001, September 6). “Search the invisible Web.” *The Guardian*. Online. Available from: http://www.guardian.co.uk/print/0,3858,4250864-110837,00.html.

Sherman, C. & Price, G. (2001). *The invisible Web: Uncovering information sources search engines can't see*. Medford, NJ: CyberAge Books.

Referenced URLS

AlltheWeb	http://www.alltheWeb.com
Ask Jeeves	http://www.ask.com
Beaucoup!	http://www.beaucoup.com
Complete Planet	http://www.completeplanet.com
ERIC	http://www.eric.ed.gov
Froogle	http://froogle.google.com

Gateway to Educational Materials (GEM)	http://www.geminfo.org
Google	http://www.google.com
Government Printing Office (GPO)	http://www.gpoaccess.gov
Invisible Web.net	http://www.invisible-Web.net
Internet Movie Database	http://www.imdb.com
Kartoo	http://www.kartoo.com
Librarian's Index to the Internet	http://lii.org
Library of Congress	http://catalog.loc.gov
Library of Texas (LOT)	http://www.libraryoftexas.org
LookSmart	http://search.looksmart.com
NARA	http://www.archives.gov
ProFusion	http://www.profusion.com
SearchAbility	http://www.searchability.com
TRAIL	http://www2.tsl.state.tx.us/trail
Vivisimo	http://vivisimo.com