Module 5

Introduction

For many research needs, the general purpose search engines discussed earlier in this course may be a good place to start gathering information. However, many of the major search engines index hundreds of millions of Web pages, and the most carefully phrased search statement may sometimes produce an unmanageable number of results, many of which are of poor quality. Another problem is that there are many Web resources, including the contents of searchable databases, contained in a subset of the Web called the "Deep Web." The information in these databases is "invisible" to general Web search engines.

The Internet Search Tools chart in the Module 3 Introduction outlined the differences between four basic types of search tools: search engines, meta-search engines, metasites, and subject directories. As explained in Module 3's "How Do Search Engines Work", search engine databases are created by automated programs and allow you to search for websites by keyword using Boolean search parameters. Search engines usually have minimal human oversight and do not apply selection or evaluation criteria to the Web pages they index. The search tools presented earlier add an element of quality control to the search process.

The two key differences between directories and general search engines are the human factor and the presentation format of the site listings. For directories, humans (usually librarians and/or volunteers) evaluate Web sources and create categories of information, either by subject or by type of resource. Directories are usually searchable, so at first glance their user interface may resemble a search engine. However, they also offer hierarchically organized indexes of subjects and subheadings that allow the searcher to browse through lists of subjects for relevant information. Each site listed under a subject is often accompanied by a description that helps the user determine whether or not the information provided by the site will be useful.
In addition to general subject directories, Web users may encounter more focused directories, such as metasites and virtual libraries. These directories are geared to specific topic areas or user groups.

Many search engines and directory sites have expanded to become Web portals, sites that offer a wide range of services and resources, such as Web directories, white and yellow pages, online shopping malls, email services, discussion forums, etc. Web portals emulate the services first offered by online service providers such as CompuServe and America Online. My Yahoo! allows registered users to customize their initial display. This can include the latest news from a variety of popular, local, national and international sources, Web search boxes for the search engine, maps or yellow pages, previews of your email (if you use Yahoo! email) and local information such as TV listings and weather.

A specific type of portal, a vertical industry portal, or vortal, focuses on a specific industry or subject, provides news, research, communication tools, and other resources aimed at educating users about the industry or topic.

Deep Web

As Web technology advances, more and more information is stored in databases. The "Deep Web," also called the "invisible Web," consists of searchable databases, inaccessible to the spiders and webcrawlers that compile indexes for the general-purpose search engines.

There are various reasons for limited or no access to Deep Web resources:

- Some spider crawlers may limit access to files due to efficiency rules and limitations on frequency of passes through the Internet.
- Some search engines may only retrieve certain file types.
- Spiders cannot access private sites that are protected by firewalls or password-protection.

In these instances, a spider can index only the location of the Web documents, computer files, or databases but nothing contained within the resources.

The Deep Web is rapidly increasing. Many databases are maintained by educational institutions and government agencies and contain a great deal of scholarly information. There are a number of directories providing access to these Deep Web databases:

- Complete Planet is a large directory of searchable databases, although its database also includes searchable Web pages.
- **OAIster** is a collection of freely available, previously difficult-to-access, academically oriented digital resources. Created by the University of Michigan Digital Production Service, it contains approximately 30 million records from more than 1,500 contributors.

**Google Scholar, Google Book Search, and Open WorldCat**

At the same time that Deep Web resources are increasing, Google has been working to improve access to these resources. Google has developed tools for locating the material, regardless or whether the sources are available in a proprietary database or not. Google has developed tools for locating scholarly material, regardless of whether the sources are available freely on the Internet or in a proprietary database. **Google Scholar** and **Google Book Search** are examples of such tools.

The goal of Google Scholar is to provide a search platform specifically for scholarly literature, including peer-reviewed papers, theses, preprints, abstracts and technical reports from all broad areas of research. Unfortunately, most of the results displayed are for proprietary Deep Web information that is not available for free online. These include books and journal articles that may be located in your school library or can be obtained for you through Interlibrary Loan. Items that are available online will have their title hyperlinked to that page. These results can also be found through a traditional Google search.

Google Book Search is a platform for searching, retrieving book excerpts or entire contents of books (depending on copyright and proprietary factors), and linking to sources for purchasing or borrowing books. Various libraries and publishers collaborate with Google in an effort to broaden the base of books available on the Web.

Google Book Search harvests records from the **Open Worldcat Project** often providing links to the libraries that lend the books. WorldCat is a proprietary database that contains book records and links to associated library catalogs. Web researchers log into WorldCat through their local library, find books and make an online request for materials to be delivered to their local library. WorldCat was a Deep Web resource available to a limited number of people, but the Open WorldCat Project has made this resource available to the world. A WorldCat search yields a list of books and, for each book, a list of the lending libraries closest to the searcher's zip code.
Google Scholar Coverage, Access, and Advanced Searching

Google Scholar retrieves information from academic and professional publications, repositories, colleges and universities, and other web sites. Some Google Scholar search results will include open access materials and others will be behind a paywall. You should avoid purchasing articles because, as an IRSC student, you have access to a wealth of articles for free through the IRSC databases. Articles not available through the IRSC databases can be requested for you via interlibrary loan. Ask your librarian for help accessing articles behind paywalls.

You can add IRSC to your Google Scholar settings so that search results contained in IRSC databases can easily be accessed for free. If you already have a Google/Gmail account, direct your browser to scholar.google.com and login. Click on settings in the top left. Click on Library Links on the left. Type “Indian River State College” in the search field and press enter/search. Ensure that Indian River State College is added below “Open WorldCat” and that the box next to it is checked.

Like Google search, Google Scholar offers an advanced search featuring the following limiters:

- Additional search boxes with Boolean Operator functions
- Author
- Publication
- Date
Directories

Directories are important search tools to consider because of the human interface that organizes and categorizes the Web for easier accessibility of information. Google is a leader in the spider-based automated search engines and has become so commonplace that we now “google” information on the Web. However, search tools with databases that contain human-evaluation resources should not be overlooked because these tools often provide a more efficient alternative to the huge result lists generated by a Google search.

Directories categorize information on the Web. We are familiar with phone directories. Each one contains the white and yellow pages. The white page listings categorize information by people and organizations while the yellow pages categorize by subject. Similarly, Web directories categorize information. Directories usually come with keyword search capabilities, but the main feature of any directory is the browsing capabilities.

Directories are divided into three types: (1) general subject directories, (2) metasites, and (3) virtual libraries.

**General Subject Directories**

General subject directories divide the entire Web into subject categories and sub-
categories. General subject directories are useful for browsing information on the Web by subject in the same manner some people prefer to browse the shelves of a library. This may be helpful if you feel that your more general searching has been ineffective. It is important to note that subject directories will only represent a small fraction of the information available online. Subject directories are collected, organized, and maintained by “experts” (sometimes with credentials, sometimes hobbyists), with the intent to curate the highest quality sources on a topic.

Typically, a search in a general subject directory begins at the top level. For instance, "Arts and Humanities," "Education," or "Health" are examples of top-level subject categories. After choosing a subject at the top level, click on links in order to move through lists of submenus and narrow the search.

The following general subject directories are other alternatives to Yahoo’s directory and can be useful research tools:

- ipl2
- Open Directory Project
- Galaxy
- Best of the Web
- Digital Librarian

Additionally, many libraries will offer research guides that act as subject directories to the best resources for your projects

**Metasites**

Metasites are more focused directories devoted to specific subjects or topics, blogs, movies, recipes, software downloads, etc. For example, Medline Plus is a metasite for health and medicine; Findlaw is a metasite for law and legal issues. A metasite of recipes on the Web is Allrecipes.

When searching the Web in specific subject areas or for specific types of resources, metasites are most useful. An encyclopedia or almanac or other reference source may be required for research, so metasites like LibrarySpot, Britannica.com, Bartleby.com, or refdesk.com may be the best tools. Poetry, poets, and criticism of poetry could be found at Poetry.org, another metasite. Metasites of electronic books (Project Gutenberg) and electronic periodicals (Directory of Open Access Journals) are examples of tools that limit searching to smaller subsets of the Web.
Metasites for various subjects include:

- Arts: ArtsEdge
- Health and Medical: Medline Plus
- Humanities: Voice of the Shuttle
- Government: USA.gov
- Law and Legal Studies: Findlaw
- Science: Public Library of Science

Although a bit outdated, a good search engine for finding metasites is provided at Research Central: Internet Search.

When regularly researching in a particular subject area, a Web searcher becomes familiar with metasites in specific disciplines. Librarians regularly search the Web and become familiar with various resources. They are also a good source of information about specific metasites. Librarians publish information about websites useful for their academic institutions. A searcher at a college or university is wise to consider using the institution’s virtual library – the third type of search tool for retrieving human-evaluated information.

**Virtual Libraries**

Virtual libraries are similar to general subject directories and metasites, but are typically affiliated with one library and the specific resources that library provides for its researchers.

The determination for placing specific resources in an academic library collection is made by librarians who work with faculty and students in providing reference services and teach academic research methods. The links and other sources identified and maintained on academic library websites are very specific subject directories tailored to the needs of faculty and students of these institutions. Utilize these sources for specific academic projects.

**Summary**

**Characteristics of directories:**

- Usually compiled and maintained by people
- If run by a computer program, maintain some type of automated selection criteria established by humans
- Smaller databases than those of the general purpose search engines
- Catalog a small segment of the Web's millions of documents
- Provides quick and easy search by subject or other category, and often with a keyword search engine
Advantages of using directories:

- Focused research
- Results usually more relevant than those of general search engines
- Extremely useful if researcher has no idea where to start searching
- Determine types of resources available on the Internet for a particular subject
- Often have fewer, or no, advertisements

Disadvantages of using directories:

- Smaller subset of information
- Less comprehensive search

Text Documents and Special File Formats

The majority of files found through Internet searching are HTML files. These look similar to the page you are currently reading. Special attention is needed when searching for information that may be found in text-based file formats, such as PDF or Microsoft Word. Examples of these files would be company annual reports or forms, such as IRS tax forms (e.g., Form 1040). While Yahoo! and Google search text within PDF and some Microsoft Office documents, files created with other programs may not be located. All crawlers do not search through all documents posted on the Web. If searching for something that may be printed in a file format other than HTML, it is worthwhile to take a look at your chosen search engine’s advanced search screen. These normally provide a quick summary of the engine’s search capabilities.

PDF files are increasingly prevalent on the Web and are used by many government, corporate and educational sites to provide resources that were originally created in other file formats, such as word processing, spreadsheet or desktop publishing formats. PDF, which stands for Portable Document Format, was developed by Adobe Systems and provides an international open standard for document distribution. PDF files require the free Adobe Acrobat Reader for viewing or printing.

Multimedia

Multimedia on the Web consists of images (photographs or graphics, single display or limited animation), audio files, videos, and a combination of all these. There are many types of files available, including image files (jpg, gif, etc.), audio files such as MP3s, streams of historic speeches or live radio broadcasts and multimedia files that incorporate both audio and video, such as a music video or
a live TV news stream. Unfortunately, unless someone has also included in the Web page a written text of the speech or transcript of the show you are searching for, it will not be found using a normal search strategy.

The Web is a rich source for graphics and photographs, but be aware that most images on the Web have cryptic filenames that may not correspond to the subject of the image, such as libimg.gif or comp.jpg, so a standard keyword search is not likely to be successful. General searches usually do not produce audio or video files since these files often lack corresponding text descriptions to connect with your search terms.

You can search for these special file formats by using one of the general-purpose search engines that provide multimedia searching, or you can use a metasite devoted to multimedia files or a particular type of file format.

The following chart provides a list of common media file types found on the Web, along with some of their extensions. Some file formats are not supported by some operating systems or Web browsers, or may require a browser plug-in.

<table>
<thead>
<tr>
<th>File Extension</th>
<th>File Type</th>
<th>Media Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>.au</td>
<td>Audio</td>
<td>Audio</td>
</tr>
<tr>
<td>.avi</td>
<td>Audio Video Interleave</td>
<td>Audio</td>
</tr>
<tr>
<td>.bmp</td>
<td>Bitmap</td>
<td>Graphic</td>
</tr>
<tr>
<td>.gif (pronounced jiff or giff)</td>
<td>Graphics Interchange Format</td>
<td>Graphic</td>
</tr>
<tr>
<td>.jpg or .jpeg (pronounced jay-peg)</td>
<td>Joint Photographic Experts Group</td>
<td>Graphic</td>
</tr>
<tr>
<td>.midi</td>
<td>Musical Instrument Digital Interface</td>
<td>Audio</td>
</tr>
<tr>
<td>.mov</td>
<td>QuickTime Video Clip</td>
<td>Video</td>
</tr>
<tr>
<td>.mp3</td>
<td>MPEG, Audio Layer 3</td>
<td>Audio</td>
</tr>
<tr>
<td>.pdf</td>
<td>Portable Document Format</td>
<td>Text, Graphics</td>
</tr>
<tr>
<td>.png</td>
<td>Portable Network Graphic</td>
<td>Graphics</td>
</tr>
</tbody>
</table>
For most files on the Web, you will need to download the following software: Adobe Acrobat Reader, Adobe Flash Player, Shockwave, RealPlayer, and Quicktime. There are many general search engines and resource directories that allow you to search for various types of multimedia formats. Some are included below.

**General Search Engines with Multimedia Search Capabilities**
Advanced search features in Google include limiting results to various textual file formats, including PDF, Microsoft Office, PostScript, and other file types in their database. Google also offers the ability to "View as HTML," allowing users to view the contents of these file formats even if the appropriate application is not installed. Google multimedia search engines are Google Images and Google Video. Google Images includes searching by size, file type and color.

**Multimedia General Search Engines (Web-crawler based)**
Getty Images is a keyword search engine especially for images.

picsearch locates images and animations.

**Human Evaluated Multimedia Resources**
Fagan Finder Image Search allows image searching in image databases and search engines.

**Multimedia Databases with Resources Contributed by Internet Users**
Resource directories created by Internet users are exploding on the Web research scene. Similar to the webcrawler-based search engines, these resources are available on the Web, but few qualitative guidelines exist for the usefulness of the information contained in these databases. Wikipedia is an example of a text-based resource that contains user-defined information. In the multimedia area, blinkx is the self-proclaimed world's largest video search engine. YouTube is the most well known, but iPod users have created various directories, too (for example, Podcast Directory).
YouTube is a useful tool for presentations in the academic classroom because of its limitations to streaming video. However, beware of copyright restrictions (see below) even in the classroom. A more detailed description of YouTube is provided below:

**YouTube** is a metasite that searches videos contributed to the YouTube database, including individual creations and snippets from professional productions. The primary focus of YouTube is its online video streaming feature rather than as a downloading service. There are services that bypass the restrictive streaming capabilities of YouTube and enable downloading, but beware of copyrights on some videos. Such copyrights make it illegal for downloading and, particularly, re-distribution of the material. Besides searching any Internet video content, Google Video searches the YouTube database.

**Beware of Images!**

You have the option of turning on the setting for Safe Search when searching images. Safe Search filtering is not a perfect process. The goal is to keep out materials that may be considered objectionable; however they will also often exclude reputable sites because of terms used within the page. This is especially true when researching certain medical conditions such as breast cancer. Generally, an expert searcher can avoid these materials by evaluating their search results screen before clicking on the mentioned links. However, when searching for images, your results screen will usually display thumbnail formats of the retrieved images. Therefore, search engines will generally default to a higher level of filtering for multimedia searching. If you are sensitive to graphic content, think carefully about your search terms and notice whether or not your search engine is using a filter before executing an image search.

**A Word on Copyright**

Documents found on the Web are protected by copyright law. This means that text, images and/or media files should not be reproduced without permission from the owner. Some government sites, such as the [American Memory Project](https://memory.loc.gov) from the Library of Congress allow limited use. Always check for a copyright, permissions, or rights page before reproducing any information from the Web. Generally, use of information from the Web is considered fair use (which means it is okay) when used for an academic project, live class presentation, or paper, as long as proper credit is given to the author or creator. This is normally shown through a Bibliography or Works Cited page.

Remember: using intellectual content without proper attribution is plagiarism.
So what is a surefire method of obtaining legal material from the Internet? The safest bet is to use...

**Creative Commons**

The [Creative Commons](https://creativecommons.org) project is a non-profit community where users allow use, sharing, and remixing of their original content at no charge to the public. The content creators can set limits to how their material (audio, video, photographs, literature, etc.) may be used and attributed. **Remember: all creative commons licenses require crediting the original work. Merely using the content without attributing it is plagiarism.**

Read about [best practices for attribution from the Creative Commons wiki](https://creativecommons.org). This will guide you in how to give credit to the creator of any Creative Commons-licensed materials you find and use.

Licensed under the [Creative Commons Attribution Share Alike 3.0 License](https://creativecommons.org/licenses/by-sa/3.0/) Copyright © 1997-2015 Florida College System, Council on Instructional Affairs, Learning Resources Standing Committee. Last revised June 2015 by the LIS 2004 Course Revision Committee.