Nicholson, S. (1997). Indexing and abstracting on the World Wide Web: An examination of six Web databases. Information Technology and Libraries 16(2). 73-81.

Indexing and Abstracting on the World Wide Web:

An Examination of Six Web Databases

Scott Nicholson

About the Author:

(current blurb) Scott Nicholson, Ph.D., Syracuse University School of Information Studies

(original blurb) Scott Nicholson (scott@scottnicholson.com) is currently working at his doctorate in Information Science at the University of North Texas. Previously, he was the Reference/Electronic Services Librarian at Texas Christian University. He received his MLIS at the University of Oklahoma in 1996, and worked as a computer consultant before becoming a librarian. He used the conclusions from this paper to create *SCOTT - Your guide to finding it on the Internet*, a virtual reference librarian that directs users to the best place to search for their information need. *SCOTT* is currently located at http://www.askscott.com .

Abstract:

Web databases, commonly known as search engines or web directories, are currently the most useful way to search the Internet. In this article, the author draws from library literature to develop a series of questions that can be used to analyze these web searching tools. Six popular web databases are analyzed using this method. Using this analysis, the author creates three categories for web databases and explores the most appropriate searches to perform with each. The work concludes with a proposal for the ideal web database.

The Internet provides a link to many valuable information sources with no centralized database for organization and searching. Many individual web databases and their attached search engines accessible through the World

Wide Web compete to provide subject and keyword access to information available through the Internet. These databases are created by both humans and automated computer programs called "spiders" or "robots." As there is no standard (such as an AACR2R variant) for description of web pages, each engine provides access in a unique way to a different database. This article will examine the methods used to collect information about the information resources, the indexing used, and the abstracting done as of February 25, 1997 in these six web databases:

Lycos - http://www.lycos.com

Alta Vista - http://www.altavista.digital.com

Excite - http://www.excite.com

Open Text - http://index.opentext.net

Yahoo - http://www.yahoo.com

Magellan - http://www.mckinley.com

To evaluate these databases from the viewpoint of an indexer/abstracter, three aspects will be examined - collection methods, indexing, and abstracting. The following questions, selected from Auster (1986), Conhaim, (1996), Courtois, Baer, and Stark(1995), Katz (1992), Lancaster (1991), Venditto (1996), and Winship (1995), will be examined for each database:

Collection methods

- How are sites selected (human/automation)?
- What selection criteria are used?
- What types of Internet resources are analyzed?
- What is the scope of searching the Internet for sites?
- How long does it take a site to be included?
- How often are the entries updated?
- How large is the database, and how fast is it growing?

Indexing

- Which parts of the site are indexed? Are these parts appropriate surrogates for the work?
- Is a controlled vocabulary used? Is it available to end-users?
- How is the keyword indexing accomplished?
- How can users search the indexed terms?

Abstracting

- What is included in a displayed citation?
- Can the user discern from where the citation came?
- How valuable is the displayed citation in assisting a user to predict usefulness?
- Are there descriptions, abstracts or reviews presented for the site? How are they created?

Summary

- For what type of searching is this database suited?
- For what type of searcher is the search engine created?
- How could the database/search engine be improved?
- How can an author assist the database service in accurate indexing and abstracting?

LYCOS

Lycos, the "Catalog of the Internet," is one of the oldest search engines on the web. It was started at the Center for Machine Translation at Carnegie Mellon University in 1994 (Mauldin and Leavitt, 1994). Lycos is one of the most popular web databases, and was the first search engine available from the Netscape Net Search button (Notess, 1995), but currently shares that honor with Web Crawler, Excite, Yahoo, and Infoseek. Besides the web database, Lycos provides access to a subject directory, the top 5% of the web, and information on cities, stocks, individuals and companies.

Collection methods

Upon request by a user, Lycos sends out a spider that will navigate the site, recording information in the database. As quickly as possible, the spider will bring back information regarding the home page of the site, then will navigate the entire site at a later date. It takes between two and four weeks from the initial request to search a site, and all sites are visited every four weeks to keep the database current. (Lycos, 1997a).

There are no selection criteria used for determining what sites are indexed. Any submitted URL that results in an active web page is eligible for the database. Lycos indexes FTP and Gopher sites along with web documents. It will ignore Telnet and WAIS-based information (Mauldin and Leavitt, 1994). There are currently summaries of over nineteen million web pages (Venditto, 1996), and about 50,000 new database entries are made each day (Kimmel, 1996).

Indexing

Lycos indexes the title, headings, subheadings, first twenty lines, and the 100 most "weighty" words in a document, ignoring stopwords. (Maudlin and Leavitt, 1994). Thus, depending upon the algorithm used, the keywords selected may not be appropriate surrogates for the work. In addition, the same number of words is collected regardless of the length of the work. Therefore, the consistency of indexing is not good - works with 200 words will have half of them indexed, while works with 10,000 words will only have 1 percent of the words selected.

There is no controlled vocabulary used in the database; instead, the database is created from natural language extracted directly from the text. This is not made clear to users in the search engine, and may cause searching problems. Users can search with the Boolean AND/OR (in different searches) and the natural language version of NOT. A right-side truncation symbol is supported along with an "exact" operator, so as a default, Lycos does search for typical endings on a search term. (Lycos, 1997b).

Abstracting

Results are displayed in a relevance-ranked format, which allows for the weight of the words searched, the location in the site, and the site's relative popularity (Courtois, Baer, and Start, 1995; Venditto, 1996). Here is a sample entry:

6) CNNfn - the financial network

CNNfn - the financial network **CNNfn** description **CNNfn**.com delivers comprehensive coverage of world business news and financial markets straight from the **CNNfn** newsroom...

http://www.cnnfn.com/index.html (7k)

[91% relevant]

The citation has the title of the page and an abstract made up of some of the first 100 words. It lists the keywords found in bold. The citation ends with the URL, the size of the page, and the relevancy ranking (with the most relevant document ranked with 100%). However, it is not easily discernible by the user from where this information came.

There is no human-created part of this record - it is all created by the Lycos spider. If a word is not one of the first ten keywords or in the first 100 words of the page, but is in the database, the word will not show up on the citation. Therefore, this citation is of some use, but could be much better if it included all instances of keywords in the database records in some type of KWIC index.

There is a nice feature on the search results page - a button marked "Related Sites" will take the user into Lycos' Internet Directory in a related subject area. In addition, Lycos allows the user to easily search for pictures or sounds of the topic searched by providing appropriate buttons.

Summary

Lycos, has a large, well-indexed database. Lycos offers higher-precision searches than databases that index the full text of each page, but lower precision than those based on a controlled vocabulary. The search engine has few features, but is intuitive to new users. The ranking information is useful, and the abstracts are fairly lengthy, just missing a few features. Lycos is a good intermediary search step, but with a variable number of representative keywords and more options in the search engine, could become more useful for advanced searching.

An author can improve the accuracy of indexing by using a controlled vocabulary in the creation of the document. Smaller documents will be more thoroughly indexed than larger ones, thus to have better coverage, documents

should be broken into multiple web pages. As the first 100 words are shown as the abstract, the author should keep that in mind and ensure that the scope of the page is covered in that text.

ALTA VISTA

Alta Vista is the code name for a web search program that navigates the web by using multiple spiders that systematically search and index web sites (PR Newswire, 1996). It was originally created to demonstrate the power of Digital's Alpha line, which is a 64-bit computer that runs at 266 MHz and utilizes 256 MB of RAM. It is currently claims to have indexed the largest number of web sites (Venditto, 1996), with twenty-eight million accesses per weekday as of February 25, 1997 (Digital Equipment Corporation, 1997d).

Collection methods

Alta Vista is completely automated. It currently claims to have thirty-one million pages indexed (Digital Equipment Corporation, 1997d), and searches the web at 2.5 million pages per day (PR Newswire, 1996). However, there are no selection criteria used - it indexes every page the spiders can find. It only indexes web sites and newsgroups, ignoring Gopher and FTP sites (Venditto, 1996). Its index is updated once a day with the newly submitted URLs and the pages found by the search robot, Scooter (Digital Equipment Corporation, 1997c).

Indexing

Every word on the page is indexed by Alta Vista (Digital Equipment Corporation, 1997b). There is no quality control used to determine what might be an important word in indexing (although there is some used in the relevance ranking). Thus, Alta Vista has tremendous indexing depth. However, a user may have to search through pages and pages of hits to find a relevant citation. The lack of control in indexing makes for very low precision when searching Alta Vista unless the user is well-versed in search tactics or is using very specific terms.

The tools are available to write detailed search requests. There are two types of queries available - a simple query and an advanced query. The simple query is a natural language search engine, where a + is indicated before a term to ensure that term exists in the record, a - is used to ensure a word in not in the record and a user can search for a phrase by placing it in quotes (Digital, 1997b). The advanced query allows Boolean searches using variants of AND, OR, NEAR, and NOT. It also allows the user to search for phrases and use parentheses to group terms. A non-Boolean element of the advanced query allows the user to specify one term in a "ranking field," which will give that term a much higher weight (Digital, 1997a).

A new searching feature is called "Live Topics." After a search, a user may select a "Live Topics" organization, which will analyze the results and group common words on the pages into categories. The user is then prompted to search the database again, including or excluding these common terms. This is a unique feature of Alta Vista, and is one of the first attempts in web search tools at the computer creation of entries similar to subject headings. This attempt will hopefully lead search tools into a new era of using artificially intelligent searching features.

Abstracting

Alta Vista's weak point is in its citations. It lists the first twenty-five words of the page (or a user-supplied

description with no quality control), along with the title, the address, the size, and the last date of indexing. It gives no indication of ranking, important words, or even where the search terms occurred in the document. Here is a sample entry:

CNNfn - the financial network

CNNfn.com delivers comprehensive coverage of world business news and financial markers straight from the CNNfn newsroom. Major corporate stories, live

http://cnnfn.com/ size 8K 9 Feb 97

http://www.cnnfn.com/ - size 8K - 5 Feb 97

The user will not be able to tell much about the site, nor from where the abstract came, and if there is no pertinent information in the first twenty-five words, the user may not know why the site was selected. The entries are ranked in a relevance order according to the location of the words in the document, closeness of keywords, and any user-given ranking (Digital, 1997a). However, there is no ranking information indicated to the user. In addition, as the number of search terms found is not indicated, the user has no way with an OR query to know when the citations only satisfy some of the keywords. This facet of Alta Vista is where the most work needs to be done.

Summary

Alta Vista's full-text indexing and huge database play heavily in its favor for those looking for find a high recall search of web resources on a topic. However, as the full text is indexed, searchers should include variant word forms and synonyms. Because it indexes every word, Alta Vista excels in searching for specific names, quotations, or phrases. The two Query options make this tool appear attractive for novice and advanced searchers, yet novice searchers will not realize the records they are missing with an ambiguous search. In addition, the abstracting done is very limited. By expanding the citations and adding a controlled vocabulary for pre-coordinated searching, Alta Vista would easily become the premiere web database.

As Alta Vista indexes every term in the database, an author can control how high the page will appear in a ranked relevance output by using the same form of words. In addition, it is important to use many synonyms to receive better search results. However, terms that have nothing to do with the work should be avoided. The user can include keywords in the META tag to help his entry be accurately indexed. The first twenty-five words of the page should be highly relevant, as that is what the user has to judge the usefulness of the record.

EXCITE

Excite was one of the first "all-purpose" search service on the web. Excite contains a search engine, reviews to selected sites, a subject directory, city information, current news, and other information. Excite focuses on making their site easy to use with many features for the end-user, with selectable types of searches and a good way to refine searches (Venditto, 1996).

Collection methods

Excite's robot examines popular sites to find links. Every week, Excite's robots go out and check every site in the database and examine various What's New pages on the web. Thus, the quality control used for Excite is whatever is popular. In addition, users can suggest a URL to the Excite engine, and that page will be added to the database during the next web search. Its database is large, with fifty million pages indexed and 60,000 pages reviewed(Excite, 1996).

After this weekly data is collected, the database is then rebuilt. Thus, web pages found through Excite are most likely going to be up and the user will have current information regarding the page (Venditto, 1996). Excite looks at Web cites and indexes about half of the Usenet newsgroups. It does not index Gopher or FTP sites (Excite, 1996).

Indexing

Excite uses "Intelligent Concept Extraction" to analyze a site and determine keywords for indexing, by searching for concepts instead of keywords. It uses an algorithm to look for themes in the page and find relevant concepts on a page. It does not examine the HTML META tag, which allows a web author to assign keywords to their own pages, which is a grave mistake. Instead, the terms from the title and the "relevant" keywords, according to the robot, are then indexed (Excite, 1997).

In addition, popular and user-suggested sites may be reviewed by a "team of journalists" (Venditto, 1996). When a site is reviewed, it is added into the subject tree directory and listed with a sentence-long review. Thus, the users may navigate the tree of controlled vocabulary to find a few selected references about a topic. (Excite, 1997).

In searching the database, users can select whether they are searching the Web database, the indexed newsgroups, or just the "for-sale" newsgroups. Searchers can use the Boolean search terms AND, OR, and AND NOT, or can add a + in front of a word to require the word and a - to find hits that don't contain the word (Excite, 1997). In addition, users can quickly navigate the subject tree of reviewed pages and the current news.

Abstracting

The search produces an entry like this:

100% CNNfn - the financial network [More Like This]

URL: http://www.cnnfn.com:80/index.html

Summary: FREE download Mercury's woes deepen. For Mercury Finance, Friday struck like a chain reaction pile-up on the freeway as investors sent the auto financing company's shares spiraling downward.

Thus, the entry consists of the title, a summary, and the percentage of relevance the cite has to the search. In addition, there is a "More Like This" link with each entry. If this button is pressed, Excite takes the keywords from that database entry and researches the database on those terms. It is an excellent way for the user to refine a

search.

The summary is created by the search program. Once the keywords are defined, the robot looks for the lines that are the most appropriate using the keywords as a guideline and concatenates those for the summary. One problem with this approach is that if a page has a keyword many times in a sentence, that sentence may be used as the summary (Excite, 1996). Thus, the summary may be excellent or unacceptable depending on the syntax on the page, and the source of the summary cannot be discerned by the user.

Both entry forms are very sparse, with very little information about the page. Without having the URLs on the page, it's difficult to easily tell what pages come from the same site. While the summary may be more useful than just the first words from the page, it may also be misleading. This entry form is of variable use to the searcher, depending upon the accuracy of the summary or review.

Summary

Excite's combination of a subject tree and keyword searching is very good for searchers. The database indexes concepts or keywords instead of the full text. Thus, the searches on Excite are good for searching on non-specific searching and a good second tool to use, much like Lycos. Excite is a good search service for a beginning searcher that is ready to move away from the subject-tree approach of Yahoo.

The abstract will usually be more useful than just pulling the first X words from a page, and the ability to find similar documents is a useful feature. By increasing the database, but still maintaining some quality control, Excite could become a very useful site for the non-academic searcher. As they only index popular sites, however, their use in academia is limited.

To be accurately indexed, an author should avoid phrases in the document that are not descriptive of the work. As only the keywords will be represented, the author should apply a controlled vocabulary to the document when possible. However, Excite warns against a user repeating a term many times in a row to give it a higher ranking. This will cause the abstracting software to pull those sentences as the abstract (Excite, 1997).

OPEN TEXT

According to Internet World magazine, Open Text has "arguably the best-designed search interface on the Web" (Venditto, 1996). This web database has two different ways of searching - simple and power - and provides useful features with the abstract listings. The search engine was created to demonstrate the Open Text Corporation's search software, which has been used for years in full-text indexing projects (Open Text, 1997).

Collection methods

Sites are collected by an automated program which explores about 50,000 pages per day. It has a large database with over ten billion words, as it collects the full text from every page. Users can submit URLs for indexing, and there is no criteria used in the selection of web sites for the index. Open Text indexes Web sites, gopher menus, and FTP sites, but no Usenet newsgroups (Venditto, 1996).

Indexing

The entire web page is indexed - there are no stop words. There is no controlled vocabulary applied, so the only searching available is natural language searching. The default search option is a Boolean AND search - this can be frustrating, as most web search engines default to a Boolean OR operator. Users may wonder why they are retrieving so few hits if they entered synonyms of a term, which might be successful on other engines.

This is a good search engine for the advanced searcher. The "Power Search" allows the searcher to use the Boolean AND, OR, NOT, and the adjacency concepts of near and followed by. In addition, each search term can be limited to a "field" of the web page. This was one of most powerful search engines currently on the Internet in 1995 (Courtois, Baer, and Stark, 1995); however, at that time, Open Text allowed the user to weight each term in the Power Search. That weighting option has been removed; nevertheless, the Power Search is still quite useful.

Abstracting

Here is a sample entry taken on February 27, 1997 from the Open Text database:

1.CNNfn - the financial network (score: 1001, size: 8.0k)

From: **http://www.cnnfn.com/**

No summary available.

[Visit the page]

The relevance-ranked citations show the title, the address, the size, relevancy score, and an abstract from the page. The abstract is created by the robots upon indexing the page - it looks at the first 100 words on the page, discards the HTML, and the rest becomes the abstract(Open Text Corporation, 1997). In the case of this example, the web site is graphics intensive and thus there was no summary available. Without reading the help files, the users cannot discern from which area of the page the abstract came.

One unique feature (disabled as of February 1997 due to a server upgrade) is that users can look at the places on the page where the search term was found in a KWIC index. Searchers can also select an entry to find similar documents, which will run a search using the most frequently used words on that page, giving them weighting according to their use (Courtois, Baer, and Stark, 1995, 15). Thus, by using this tool, a user can easily find a group of similar sites. These tools are a valuable part of this database. With a better abstract, this would be a very useful citation display.

Summary

Open Text has a very powerful search engine, with a database smaller than Alta Vista. It collects the full-text of each page, so that carries higher recall from the Internet and lower precision than those databases based on keywords. As with Alta Vista, this tool is ideal for finding specific terms, quotations, or phrases. The search engine is not for novice users, especially since the default search option is not OR or AND, as expected. Its size

and lack of selection criteria are drawbacks - it either needs the database size of Alta Vista or needs to apply some quality control in choosing the indexed sites. The reporting format is useful, and the ability to generate a KWIC index is unique.

As with other full-text databases, the author can improve the indexing of a page by using controlled word forms, including synonyms, and avoiding spurious phrasing, and including many synonyms. The author should also focus on filling the first 100 words with a good description of the page to provide an accurate abstract. If the first 100 words of the page are HTML code, Open Text will not produce a suitable abstract.

YAHOO

Yahoo, which stands for "Yet Another Hierarchically Officious Oracle" (Kimmel, 1996), is one of the best-known web databases. It was started by two graduate students at Stanford, and now is managed by Netscape (Yahoo, 1997a). It is built around a subject tree directory, but has a search engine which links into the Alta Vista web database if nothing is found at Yahoo.

Collection methods

Yahoo is created and updated by human indexers. A spider program collects web site information from various "What's New" sites around the web and they take user submissions. Yahoo does employ some quality control in its selection of pages to add to the database of 200,000 entries (Courtois, Baer, and Stark, 1995 and Steinberg, 1996). Yahoo looks at web sites and newsgroups, but doesn't index Gopher or FTP sites. Entries are usually only updated when a user points out a problem (Yahoo, 1997b), instead of there being any automatic examination.

Indexing

Yahoo indexers examine the whole site, and then reference the home page in their database, placing it under appropriate subject headings. The searchable database comes from a very short (if any) description of the site, the URL, and the title of the page. This is usually not an appropriate surrogate for the work, as most titles are poorly written. If no description is given by the Yahoo indexer, the surrogate is inadequate.

The searching is done with the Alta Vista's search engine, using phrase and natural language searching options. It searches the Yahoo subject headings and the individual page titles, URLs, and descriptions. Yahoo's strength is in the subject tree navigation available to the user, but a useful feature of the search engine is that it can be focused just upon one branch and lower of a subject tree.

Abstracting

Yahoo's entries are sparse. Here is a sample entry:

CNNfn @@ - the financial network.

(there is a pair of sunglasses at the @@ to indicate it is a "cool" site).

Not all entries have the sentence-long abstract. The entry contains the title of the page (linked to the page) and the abstract, with a symbol to indicate if it is "new" or "cool." These abstracts are developed at the same time as the subject heading. Beyond the abstract and title, the user has no more to use in selection than the subject heading for the citation. Overall, the entries are not useful in assessing the usefulness of the source, and the user has no way to track the source of the description.

Summary

Yahoo is a wonderful starting point for Internet exploration. Ideal for new users, it allows searchers to wander through subject trees, finding small groups of relevant sites under subject headings. It is best for casual browsing or finding a short list on a desired topic. Yahoo is a good place for the first step, but holds little beyond a few starting points. An interesting point is that if it finds nothing in its dababase, it runs the search in Alta Vista. It is one of the oldest tools for web navigation, but has the smallest database of all databases reviewed. Yahoo could be greatly improved by providing more information on each page indexed, and providing each entry with a description.

In order for an author to be accurately represented on Yahoo, the web page title must be descriptive of the document. To guide the Yahoo indexers into placing a page under a desired subject heading, terminology from that heading should be used in the title as well as the document itself. A user submitting a page to Yahoo should search it first for the subheadings that contain similar pages to suggest the most appropriate places for their page.

MAGELLAN

The McKinley Group, a group of publishers and information scientists, are responsible for Magellan but it is currently owned by Excite. Named after Ferdinand Magellan, this database relies upon humans for its creation and upkeep (McKinley Group, 1997). Magellan provides an easy-to-use starting point for Internet novices and those just wanting "quality" sites on a topic. When searching Magellan, the default search looks at a large database - the user must select "rated and reviewed sites" after typing in the search to search the "reviews" database.

Collection methods

Magellan reviewed sites are selected by people. Sites are examined to see that they are easy to navigate, are thought-provoking, and contain current and in-depth information. Magellan indexes not only web sites and newsgroups, but also gopher menus, FTP sites, and Telnet sessions (McKinley Group, 1997). As the reviewing is done by humans, Magellan does not keep up with the adding or updating rate set by the automated services.

Indexing

The entire site is examined for indexing for subject heading assignment, but only a selection of terms is taken from the site for inclusion in the searchable database. The main Magellan menu has a search box at the top and subject headings at the bottom, thus users can easily perform whichever type of search they prefer. As the keywords are assigned by humans, they will not be as consistent as those chosen by an automated program, but they will usually be more appropriate to the meaning of the work.

The keyword searching allows the natural language operators + and -, although through an advanced search, the user can select between a Boolean AND and OR. The default search operator is the OR, like many of the search engines. There is no phrase searching available.

Abstracting

Magellan offers a two-part abstract for each entry. When a search is completed, there is a short except of the full abstract displayed for each record. For example, here is the listing from the search results screen:

CNNfn -- The Financial Network

Review: If you're interested in bypassing the central CNN site and want to get straight to its business and finance coverage, go here. You can stop by to get the day's top business headlines and market reports or take your time and browse through the various links...

http://www.cnnfn.com/

In addition, on the left-hand side of the entry is the number of stars (from 1 to 4) the reviewer gave the resource, and on the right-hand side is a green stoplight if the site was deemed suitable for children (McKinley Group, 1997). This citation makes it clear that the searcher is reading a review of the site, and not an extracted section from the site.

Clicking on the "Review" link will take the user to a review page for that site, containing the controlled vocabulary terms selected for that site, a description, keywords, audience, language, producer, cost, commercial status, and moderator status. These are assigned by the McKinley group, and are based off user submissions. This information may not be easily accessible on the web site itself, thus the abstract and bibliographic citation are extremely useful. Therefore, the entries in this database are extremely helpful in allowing a user to learn about the information resource without actually going there.

Summary

Magellan is a wonderful resource for those (such as librarians) that want to know detailed bibliographic information. The reviews, the star ratings, and the stoplight feature make this an idea place for school librarians to quickly create lists deemed suitable for children by the McKinley Group. The application of criteria to sites reviewed makes this a useful place to find some quality works. The combination of a powerful search interface and subject tree allows users flexibility in navigation. This is currently the best tool for a high precision search with quality results.

Authors will have to be careful to ensure quality on their site to be reviewed by Magellan. As all of the indexing and abstracting is done by a person, there are no rules to follow to predict what will be chosen as a keyword. However, in order to apply for indexing in Magellan, the user must write a review and assign themselves a subject heading and keywords, and this is used as a basis for the indexer. Thus, the accuracy of the information depends upon the author of the application.

Conclusion

Which is the best? There is no answer to that - it depends upon the user's need. As the search engines are fairly equivalent in their power, the single more important factor in a successful web search is the selection of the proper web searching tool. By using the questions at the beginning of this article and the definitions presented below, new search tools can be categorized as *directory*, *full-text*, or *abstracting*. By classifying the search tool by the indexing method used, predictions can be made about the type of searching to do with that tool.

The *directory* search tools are those that provide subject headings for navigation, usually created by a human. No text is taken from the page for indexing; the pages are examined and classified into a subject heading hierarchy. Yahoo is a directory search tool, as are parts of Lycos, Excite, and Magellan. Directory tools are useful for casual browsing. If a user is looking for a short, on-topic list of pages on a topic, the directory tools are also the ideal place to start a search.

The *full-text* search tools index every word on every page of the database. Alta Vista and Open Text fall into this category. These tools are very useful if the user wants information on a specific person, company, or other unique name. They are also good for information on a quotation or song lyric. If the user can think of a phrase that would contain the information for which they are searching, such as "national debt per capita," a phrase search (usually surrounded by quotation marks) will be ideal. This application of phrase searching is ideal for answering short, factual, "ready reference" questions. These tools are not good for general subject searching, and that is one of the greatest frustrations for users.

The *abstracting* search tools take a selected portion of the target site for indexing. These tools use some type of algorithm to select frequently used or prominent words on a page and index them. Some of these tools include Excite, Lycos, Magellan, Web Crawler, and Hotbot. These tools are suited for general subject searching, and are a good place to go as secondary tools if a *directory* tool or a *full-text* tool provides unsatisfactory results. The user can choose a reviewing tool, such as Magellan, for sites that have been chosen for some level of quality, or some of the other sites for a more comprehensive search. In addition, these tools will provide the greatest level of success when combining facets in a search.

Thus, by examining the indexing methods of new web tools, they can be placed into one of these three categories. Once a tool has been categorized, it can be used appropriately and compared to other tools that fall in that category for further suitability to task. Table 1 summarizes these conclusions, and can be used as a quick guide for selection of the most appropriate tool.

Summary of Search Tool Types and Appropriate Searches

Directory Search Tools	Full-Text Search Tools	Abstracting Search Tools
(Pages placed into a hierarchy of subject headings)	*(Searchable database of every word of indexed pages)*	*(Searchable database of selected words from indexed pages)*
*Browsing the Web for pleasure	* Specific names or unique terms	* General or ambiguous topics

*Short but relevant lists on topics used as subject headings	* Phrases that would appear on the desired page	* Secondary searching after using one of the other tools
*Refining a search term	* Quotations or lyrics	* Multi-faceted searches
*Beginning web research	* Ready Reference questions	* Other searches

Table 1: Summary of Search Tool Types and Appropriate Searches

The Future

These tools for electronic information were not created from the aspect of a librarian, although librarians have been the key force in the organization, indexing, and abstracting of print information. These databases are useful, but can be frustrating to use. There is still obviously room for the library to create a superior database. What might such a database look like? How would it work? Here is a hypothetical look at the features from such a database:

Collection methods

Sites are gathered by multiple spiders, processed, and brought to humans for final approval. A selection policy will be used in order to create a database of quality sites. Any reliable data source accessible by the Internet will be a candidate for coverage. Updates will occur whenever a spider detects a change in title or headings. The goal is to create a database for academic information - similar to the scope found at a research library, and in the process, collect information about all reliable resources.

Indexing

The entire web page (including URLs) will be indexed. In addition, the spider software will analyze the sentences to determine keywords, which will be matched against a thesaurus and mapped to a controlled vocabulary. In addition, an abstract will be created from the areas of the page where the keywords are concentrated.

At this time, a human indexer will examine the record to see if it meets selection criteria, if the keywords and abstracts are appropriate, and if the controlled vocabulary has been applied correctly. After this, the record will be added to the appropriate databases. All documents will be added, and those that meet the selection policy will be marked so that the user can do an academic search if desired.

There will be a basic search, which will search the keyword database. Upon returning results, the user may select to broaden the search (which will replicate the search in the full-text database) or narrow the search (which will select possible controlled vocabulary after examining the previous hits, and let the user select the appropriate vocabulary). The search may be further broadened by bringing in synonyms, which will also be suggested by the software.

There will also be an advanced search, where the searcher can search in any combination of the databases using a full complement of post-coordinating search tools (Boolean, weighting, adjacency). In addition, the controlled vocabulary will be browsable as a subject tree, which will also have the search engine applied to it as a whole or individual branches.

Abstracting

The citations will contain a clearly labeled record with all of the gathered information, as well as a link to a KWIC index and a way to find similar matches. They will be ranked using term weighting, with the most relevant entries at the top of the list (with relevancy information and how the record matched the search). There will be an option to download all of the retrieved citations up to a selected relevancy weight, or to have all of those citations listed on one page. In addition, there will be buttons across the bottom that will run the search in other search engines and will integrate the results, removing duplicates if desired.

This type of search engine would bring in the concept of precision and recall, and would be flexible enough to let searchers do what kind of search they wanted, but would also have the user-friendly mode that would take care of many decisions. The user could do a search just for academic works, or look at everything. By using some traditional tools and concepts, the library can work towards creating an excellent web database, and hold an essential electronic position in this information age.

BIBLIOGRAPHY

Conhain, W. W. (1996). The Internet. [On-line]. Link-Up, 13(1), 5, 40-42. Available: DIALOG File:ABI/INFORM Item:98-19600

Courtios, M. P., Baer, W. M., & Stark, M. (1995). Cool tools for searching the web. ONLINE, 19 (6), 14-32.

Digital Equipment Corporation. (1997a). Alta Vista: Help for advanced query [WWW document]. URL http://www.altavista.digital.com/cgi-bin/query?pg=ah&what=web (visited February 25, 1997).

Digital Equipment Corporation. (1997b). Alta Vista: Help for simple query [WWW document]. URL http://www.altavista.digital.com/cgi-bin/query?pg=h&what=web (visited February 25, 1997).

Digital Equipment Corportation. (1997c). Alta Vista Search [WWW document]. URL http://www.altavista.digital.com/cgi-bin/query?pg=tmpl&v=faq.html (visited February 25, 1997).

Digital Equipment Corporation. (1997d). Alta Vista Search: Main page [WWW document]. URL http://www.altavista.digital.com/ (visited February 25, 1997).

Excite. (1996). Common answers to existing questions [WWW document]. URL http://www.excite.com/FAQ.html (visited April 12, 1996).

Excite. (1997). Excite Help [WWW document]. URL http://www.excite.com/Info/searching (visited February 25,

1997).

Lancaster, F. W. (1991). Indexing and Abstracting in Theory and Practice. Champaign, IL: University of Illinois.

Lycos. (1996a). Lycos, Inc. Information [WWW document]. URL http://www.lycos.com/addasite.html (visited February 25, 1997).

Lycos. (1996b). Lycos Search Help [WWW document]. URL http://www.lycos.com/search-help.html (visited Febrary 25, 1997).

Katz, W. A. (1992). Introduction to Reference Work (6th Edition). New York: McGraw Hill.

Kimmel, S. (1996). Robot-generated databases on the World Wide Web. DATABASE 19(1). 40-49.

McKinley (1997). Magellan's Frequently Asked Questions [WWW Document]. URL http://www.mckinley.com/feature.cgi?faq_bd (visited February 27, 1997).

Mauldin, M. L. and Leavitt, J. R. R. (1994, August 4). Web Agent Related Research at the Center for Machine Translation. [WWW Document]. URL http://fuzine.mt.cs.cmu.edu/mlm/signidr94.html (visited April 12, 1996).

Notess, G. R. (1995). Searching the World-Wide Web: Lycos, Webcrawler and more. ONLINE 19(4). 48-53.

Open Text Corporation. (1996). The Open Text Index-Frequently Asked Quesitons [WWW document]. URL http://search.opentext.com/main/faq.html (visited February 26, 1997).

PR Newswire (1996). Digital's 'Super Spider' Becomes Internet's Fastest-growing Search Tool [On-line]. Full-text from: DIALOG File:IAC Promt(R) Item: 05937400

Steinberg, S. G. (1996, May). Seek and ye shall find (maybe). Wired 4. 108-114, 177-180.

Tenopir, C. (1989). Issues in Online Database Searching. Englewood, CO: Libraries Unlimited.

Venditto, G. (1996) Search engine showdown. Internet World 7(5). 78-86.

Williams, M. E. (1986). Criteria for evaluation and selection of databases and database services. In E. Auster (Ed.), Managing Online Reference Services (pp. 27-39). New York: Neal-Schuman Publishers. (Reprinted from Special Libraries 66(12). 561-569.).

Winship, I. R. (1995). World Wide Web search tools: An evaluation [WWW Document]. URL http://www.bubl.bath.ac.uk/BUBL/IWinship.html (visited April 12, 1996).

Yahoo (1997a). History. [WWW Document]. URL http://www.yahoo.com/docs/pr/history.html (visited February 27, 1997).

Yahoo (1997b). Yahoo FAQ [WWW Document]. URL http://www.yahoo.com/docs/info/faq.html (visited February 27, 1997).

Style Guides

American Psychological Association (APA) (1994). Publication manual of the American Psychological Association (4th ed.). Washington, D.C.:American Psychological Association.

Land, T. [a.k.a. Beads] (1996, March 31). Web Extension to American Psychological Association Style (WEAPAS) (Rev. 1.2.4) [WWW Document]. URL http://www.nyu.edu/pages/psychology/WEAPAS/.