

Paper presented at Electronic Age '96 A conference for the information industry 20-21 April 1996 Ranelagh House, Robertson NSW. To be printed in the AusSI Newsletter and LASIE 27(3):32-42.

Internet Indexing : pinning jelly to the wall?

Roxanne Missingham

"Don't panic" Hitchhikers Guide to the Galaxy (1)

"Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?" (2)

Introduction

We are, without doubt, in the information age. Information technology pervades our work, our homes and our recreation. Each day we use this technology to turn on our heating, cook our meals, produce reports, communicate with others, for entertainment and to obtain information.

Traditional publishing is changing in response to the new technology. Conferences convened by the Australian National Scholarly Communications Forum (3) in the past 3 years have described the move to electronic publishing, particularly on the Internet, by the academic/research and government sectors. This move has had an impact on all the links in the information chain - from the writing of ideas (now done on word processors), to editing and peer review (through word processing and electronic mail), publication (desktop publishing for print and electronic out put) and in locating or finding publications (electronic indexes).

The changes in information technology have an impact on the steps or strategies used to find information for a research inquiry. Research tools such as abstracts and indexes are now online, on CD ROM and on the Internet. Automation of information sources has occurred in two ways. Firstly existing abstracting and indexing services have made their information available electronically. This has enabled easier, quicker searching. Secondly new tools are being developed, such as Internet indexes, which automatically generate a database of information. Retrieval of information from the Internet is much more complex than for publications in print form. This complexity has stimulated the production of automated indexes. Attempts to index Internet information using standard library cataloguing or database indexing methods, seen in Cyberstacks and Netfirst (an OCLC service) are yet to be as successful as automated indexes, partly because of the sheer volume of material to be indexed, and also because of the newness of the medium. The challenges to researchers and information specialists to find information are being addressed through innovative solutions, both automated and manual.

This paper demonstrates that Internet indexing or searching is not yet fully solved. Search tools are critical to effectively accessing information on the Internet, but need to be used with care. A comparison of 7 major indexes (commonly called search engines) finds great differences in search results and effective access to information. Overall trends and issues for the future development of Internet search engines are then discussed.

Why is the Internet important?

The Internet is the most rapidly growing medium of communication and publication. The Internet Domain Survey, of January 1996 found almost 9.5 millions hosts were connected to the Internet, an increase of approximately 142% on the figure for July 1995. Academic, research and government sectors now publish on the Internet.

The survey also found that:

- the net domain is over twice as large as it was six months ago

- about 76000 systems are now named www (i.e are World Wide Web sites), up from only 17000 six months ago
- No one has any clue how many users there are, but most people would agree that there is at least one user per host.

(Network Wizards, 1996)

Estimates of those with access to the Internet range up to 65 million people (Four Corners, February 1995 estimated 35 million, Compuserve and America Online Figures, were quoted as 65 million in In, around and about, September 1995). The original Internet community comprised the university and major research sectors. (AARNET, the Australian Internet segment now owned by Telstra was originally a consortium of the Australian Vice Chancellors Committee and CSIRO). Schools, libraries, corporations and to a increasing extent the general public have joined the Internet. All these groups are using the Internet to publish and to search for information (or "surfing" the "sea" of information on the Internet)

Publishing via the Internet, whether by electronic mail or making information available through gopher, ftp or world wide web has the advantage of being:

- quick (you can communicate or publish apparently "instantly");
- widely accessible (to the very broad Internet community);
- cheap to produce material (many Web mark up tools are free or come with desktop publishing/word processing packages);
- relatively simple to produce (taking much less time than more complex desktop publishing); and
- pretty much in the total control of those publishing (as opposed to the significant controls of publishing houses and traditional editorial control).

Many of the traditional publishing and distribution processes we have relied upon for consistency and control are becoming less important as electronic publishing increases. Internet publishing of scholarly and non-scholarly material is changing forever the way we organise and seek material.

A critical issue in this new publishing medium remains over how, with such a vast number of Internet documents, relevant material can be found. Those publishing on the Internet wish to communicate with an audience and be read, as they did in print. The location of information is thus an important issue to those publishing as well as information seekers.

What is an Internet index?

The Internet consists of many files across the world. There are over 22 million World Wide Web pages, millions of gopher files, applications and databases. To provide a finding service "Internet indexes" or search engines, as they are often known, have developed. These play much the same role as traditional indexes used to find books in the library, or journal articles, conference papers and other publications.

Internet indexes all offer different methods of access (use of search terms) and different information about the files or "publications. The same process is used for the collection and distribution of information as a traditional printed index. Major differences exist in the production of the database used in the index, and its access.

The process for construction of traditional and Internet indexes is:

Traditional Print/database index Selection of sources Creation of Citations Addition to a database /abstracts/indexes End product/access - Print - CD ROM/online

Internet index Sources selected (usually by self nomination) Robots collect information (sometimes called 'harvesting') Database created Online access using a "search engine"

The terminology used by Internet indexes/search engines can be confusing. Common usage of some key terms is:

Search engine: used to describe 1) the actual software which searches the database of Internet files/resources or selected information on the files and also 2) the actual Internet index or complete searching service offered online.

Search: the matching undertaken by the search engine of the terms entered by the user against the terms existing in the database. There are a wide variety of algorithms used for this process.

Index: used to describe the service provided through the Internet search site comprising the information collected and made available through the search software or search engine for queries.

Robot: a program which automatically (usually on a regular basis) searches the Internet for a defined set of sites and collects information about the files at each site. Sometimes it may collect copies of the whole files, generally called "harvesting" after the software, Harvest. Others collect selective information, such as the title, location and first paragraphs.

In this paper the term Internet Indexes is used to describe the search service available, rather than search engines. Index indicates a process of selection, storage and retrieval using the search engine.

Internet indexing challenges

The Internet cannot be considered to be just another step in the history of indexing. It offers huge challenges and needs a very different approach to indexing to enable effective information retrieval. The need for quality access to information has increased, because of the number of searchers. There is also a need for tailored search tools for specific subjects. These tools save searchers time and provide access to reliable, quality information including evaluative comments from professional indexers, or what Internet Indexing services now called "professional web reviewers".

Indexing the Internet offers many challenges:

- the Internet contains millions of documents/files;
- the location of Internet documents/files changes frequently;
- there is no quality control for Internet information, no consistency in use of terminology, or even in the use of titles;
- keeping up with new sources is very difficult; and
- indexes are complicated by the fact that most will only index sites which nominate themselves (somewhat similar to the current cataloguing in publication process).

Internet documents have only two absolute features. The first is a location (where they are stored and can be found) and the second is content. There are no standards that require authors or title to be used, nor a requirement that heading information include title or sub title information. Internet indexing is quite different to journal article indexing where this information is clear.

Evaluation of indexes

Why do we need to evaluate indexes? There is a multiplicity of indexes for print and electronic information which have been used by researchers and librarians for many years. They have been used because they provide an effective and efficient mechanism for finding information on specific subjects. In commencing to research a topic or inquiry the best search strategies are selected (see Basch). Indexes are used to obtain the best citations (or information lists) as efficiently as possible. Information found needs to be relevant and

accurate to enable further refinement of the search or to enable the researcher to obtain the information sources listed in the results.

Indexes are finding aids for publications. Different indexes are used for different subjects. Different search strategies are also used. By evaluating indexes it is possible to find which are best suited for different types of research, as well as the best way to use them. These decisions have to be made for the use of traditional printed and automated indexes, and apply equally to Internet indexes.

Traditional indexes, whether in print, CD ROM or online, have been evaluated in terms of:

1. geographical coverage country of publication of material included, for scientific indexes this is usually US/UK/European orientation
2. language English/French etc
3. dates of coverage normally indexes cover material published within specific dates eg within 2 years of the date of the volume
4. discipline (subject area) such as zoological material, biological material
5. scope
6. inclusiveness (breadth) a wide range of material within subjects or selected material based on quality of research or sometimes accessibility of material
7. comprehensiveness (depth) for instance, does it include all material on zoology, including grey literature or only citations of major journals in the field
8. level (e.g. for the general public, for researchers)
9. format (CDROM, Online, print etc)

(from Lane 1989)

This is quite a comprehensive list, but on a day to day basis probably the most important criteria used in deciding which index to use are:

- subject/discipline
- scope
- date
- inclusiveness
- comprehensiveness.

All of these criteria apply to Internet indexes, together with additional factors.

Debates rage across the Internet over indexing. There is a very wide range of opinion on what should be done to ensure that information can be found through Internet indexes. At one extreme there are those who propose a completely new approach to information discovery or "mining" through a completely automated system. They argue that the only human components of information discovery will be the humans creating information on the Internet. In order to fully automate this process creators will be required to provide good quality information on title, authorship, publication date and subjects of their Internet publications. Without this information the ability of any Index to provide an excellent search result is limited. With high quality information personal robots could be used to find information on a regular (perhaps daily) basis according to personal profiles. This approach is based on a philosophy that "everyone can get what they want" and that information gathering is a long term activity gradually building each individual's knowledge. It presumes that some form of quality control can be built into the Internet to enable an easier detection of relevant information.

At the other end of the spectrum are the indexing and searching professionals who see their role is changing, but becoming more important in the "information mine" of the Internet. An example of the application of traditional indexing procedures (a very time consuming approach) to the Internet is Cyberstacks starting with a selection of subject areas, such as agriculture. This approach has its difficulties in application too, mostly due to the volatile nature of the web where indexing a resource can really be like pinning jelly to the wall, as it may be there today and gone, or completely changed tomorrow. Not only can the resource's name, contents and location change regularly, but its accessibility and format change easily as well.

If the Internet is growing, becoming an increasingly important research publishing format and more accessible, how can we determine what to use for each inquiry? A set of criteria to evaluate Internet indexes is proposed below, based on Internet index use and feedback from research skills students at the University of Canberra.

1. contents:
 1. subjects areas covered
 2. inclusiveness (breadth)
 3. comprehensiveness (depth)
2. (ftp, gopher, www) (see note 4)
3. search techniques available
4. interface design
5. accessibility (speed of access, convenience of form, regularity of updating of index)
When accessibility is very poor, however (eg response time is very slow) this may be the first factor to eliminate an Internet Index from use.
6. format of results (useful information includes title, address, extracts of text or a summary)

To test the proposed criteria, it has been applied to some actual search engines. In order to evaluate the usefulness of the Index three specific searches were also undertaken:

1. an actual research topic: a library client has asked for research on the effect of forest disturbance on fauna;
2. a general question: find "newjour" - the list of new Internet journals; and
3. a general search for the Australian Society of Indexers.

All searches were undertaken in the last week of March 1996 and results are valid only for this time.

The Evaluation

Alta Vista Excite Infoseek Inktomi Lycos Magellan Open web Yahoo

Location www.altavista.digital.com www.excite.com www.infoseek.com inktomi.berkeley.edu
www.lycos.com www.mckinley.com www.opentext.com www.yahoo.com

Contents:* Access to 11 billion words found in 22 million Web pages. A full-text index of over 13,000 news groups updates in real-time. Fully automated index Access to more than 11.5 million web documents. Reviews database (over 50,000 web site reviews). Usenet more than 1 million articles from 10,000 newsgroups Classified advertisements (past two weeks). One of the very large web indexes. Has main database, subject lists, world news. Web sites, gopher sites, (4) Reputedly over 5 million web 'a2z' directory to browse by category or find by keyword. Includes short descriptions of each site and links to related sites. Point Top 5% list by professional web reviews Contains a wide range of sites with reviews by professional web

reviewers who award 1 to 10 points in three areas: Depth (Is it comprehensive and up-to-date?), Ease of exploration & Net appeal Large index Well established, one of the most popular with end users.

Subjects areas covered All- sites added by nomination ditto ditto ditto ditto ditto ditto

Inclusiveness Very large number of sites ditto again very inclusive large number of sites very inclusive high up on this scale

Comprehensiveness Down to document text level ditto also down to document level ditto ditto listing more of sites

Scope of coverage ftp, www, gopher, newsgroups all plus review database and classified database web, ftp, gopher, listserv, other web, gopher main database covering ftp, www, gopher, etc, 'a2z' database, top 5% database sites and reviews sies and reviews/ subject lists

Interface design Drop down menus where appropriate, simple & clear Uses frames, coxes, radio buttons. You may find the form does not display in Mosaic 2. Uses tables for mock columns, subject categories are at this level. Simple input box. Uses drop down menus & single input box Drop down menus where appropriate, simple & clear, subject categories are at this level List of subejct categories and input box, uses frames and tables Simple entry form with room for 4 words or phrases, fields selection drop boxes & Boolean for each term Input box, list of subject categories, no frames, table

Search techniques available

Simple queries (can search phrases, use "+", truncation,) specify fields(eg .gov). Advanced searches (Selection Criteria: AND, OR, NOT and NEAR, phrase Results Ranking Criteria: words for sorting), Start date, End date

Advanced

searching also

available Simple queries (terms are '+d), Advanced queries (+ in front of a search word for ALL of the documents, to Excluded Words e.g Jaguar -car , full Boolean operators and syntax. All searching - capitalisation may be used, "" for a phrase, [] for near searches (within 100 words), +word must be in results, -word must change

not be in results Simple queries (terms are '+d), Advanced queries (+ and - to include and exclude words). Simple, enhanced (can "and" or "or") and formless searching. Refine search on bottom of page. Assumes "and"for terms, no capitalisation, "-" (minus) symbol between words, the search will find records containing one word but not the other. Drop downmenus allow forsearching in specific fields, can search by phrase, &offers full Boolean searches. Can"customize"

restrict search to titles, URLs, and/or comments; search case-sensitive or case insensitive; "or"; search; change result limit from 100.

Accessibility Sppeed of access comparable with most US sites, not overly heavy use of graphics Speed of access comparable with most US sites, frames slower, than no frames frames option is noticeably slower than no frames quite quick, better than most index sites quite quick quite quick

Format Contains title, basic information location relevance ranking, title, summary title, location,relevance, score, size, abstract, Similar page link relevance ranking, title, occurrences of search words, location title, relevance, ranking, outline, abstract, location, size title, contents, summary, location title, relevance score, size, location & find similar pages title (hyperlinked) and single sentence

Search #1

Comment

28 pages of material was found 2 very useful citation

Comment

Again hit rate in the 400 range. This site had 2 useful citations

Comment

Many hits, limited usefulness

Comment

Many hits WCMC material scored first, not as much relevant material as on excite

Comment

Again many hits, useful material located (2)

Comment

29 pages of material, no useful citations

Comment

only 2 citations found - neither useful

Comment

Search produced no useful results - navigated to society and culture, environment, then forests

Search #2 references to individual files in the site come up before the main site individual files on the site come up before the home page pointed to newjour page #3 in list directly pointed to correct page in the top 2 hits pointed directly to correct page #1 on list #3 on the list, #s 1 & 2 are postings to ADFA epub list not found

Search #3 ASI came up quickly ranking algorithm has American society higher than Australian! ASI came up quickly at the top of the list. ASI came up quickly as #1 on list ASI came up quickly as #1 on list 20 pages of results, #1 was Australian Computer Society Gopher, ASI was #6 on list #5 on list not found

Overall Excellent (****) Very Good (***) searching techniques limited Very Good (***) unusual use of search language (near means within 100 words) Good (**) appears not to have as much in the index, search techniques limited Excellent (****) search techniques limited, excellent format Good /Very Good (***) ranking hid some useful material, limited searching techniques Good /Very Good (***) ranking hid some useful material, does not retrieve as many citations as Lycos Very popular and interesting in approach to subject descriptions

* see note 4

Internet myths

The Internet is clearly becoming an increasing important information source. Using the Internet productively for information can only be achieved through finding the relevant information. Indexes are the key to this process. The indexes have not yet solved the problem of how to find quality information. Many factors indicate that finding Internet information is at a very early stage of development. Some are outlined below.

1) Anyone can do it

While the basic Internet software is becoming more sophisticated it can still be difficult to use. There can be problems in accessing sites on the Internet, and to search well takes a great deal of time. Searching can be complex and require the use of a number of Internet

indexes, as well as a degree of serendipity. Libraries can expect an increasing role as intermediaries in the search for information from the Internet.

2) All the information you need is on the 'net

Many major organisations connected to the Internet provide a wealth of publications and information. Many others however have only selected information available at present. Any inquiry for "published" information needs to be evaluated to see what format is the most appropriate. Care needs to be taken to make sure that essential information, which may be in print, is not missed in the search.

3) You can find resources on the 'net

As is clear from the comparison of the 8 Internet Indexes, there is a need to determine which index is the most appropriate for a search. Criteria of contents, scope, search techniques, interface design, accessibility and format of results are the most relevant. Finding resources is complicated but sites being unavailable or moving.

4) Researchers will use the 'net all the time

Many researchers are more than fully occupied undertaking research and policy work. There is a need for good searchers to build up skills in Internet searching. While many who have the Internet at their desktop will do their own searching, an Information Centre or specialist searcher can provide a cost effective result (especially in terms of searching time).

5) Indexers will be obsolete

While Internet Indexes are predominantly automated, the last 12 months has seen a trend to using people to evaluate sites. Excite, Lycos, Magellan and Yahoo now all provide searchable site descriptions or keywords allocated by "professional web reviewers". While these are limited by short descriptions, or American use of terms, they are a growing service.

Conclusions

Automated Internet indexes are:

- limited by the lack of a controlled vocabulary for searching and the lack of consistency found in good quality citation indexes;
- limited by the sites/files on the Internet themselves;
- generally quite comprehensive (ie index whole documents) and general rather than specialised by subject;
- essential to begin any research using the Internet; and
- moving to more quality, subject information.

The sheer volume of data in a relatively uncontrolled form is like providing the front door key to the whole of the Library of Congress to our clients without a catalogue or floor map. Adding the current search tools is in some ways like giving them a torch and saying "go for it".

Limitations of Internet indexes:

- do not list sites which exclude access to robots;
- do not list sites which are down at the time the robots are working and therefore are missed;
- have great trouble finding any title composed primarily of stop words, such as "I can do it";
- may be confused by "spamming" by creators who include huge lists of terms (which may not be related to the content of the file) in the top of their files.

There are personal preferences in using Internet Indexes, just as there are in using online and printed indexes.

An ideal Internet index should:

- contain an indication of the sites covered by the index;
- allow for full Boolean searches (AND, OR, NOT);
- allow for phrase searching;
- be able to be used through a wide range of software (Netscape, Mosaic, Internet Explorer);
- index all files to document level;
- contain a minimum of graphics;
- weight site home pages higher than individual files;
- rank above 50% only those using all words in the Boolean search
- use a thesaurus which can be easily accessed by searchers;
- provide the following information in the results screen:
 - title
 - URL
 - short summary
 - relevance ranking
 - file size.

The Internet resembles jelly in its constantly moving contents. Indexes for the Internet are also changing quickly. The information structure of traditional indexes, including clear statements about authorship, title and publication details do not exist in Internet files. Some have suggested the use of metadata at larger "pins" to hold the jelly to the wall, but while publishing on the Internet remains uncontrolled it is unlikely that standards such as metadata can be made mandatory. New solutions for indexing have involved a mix of automated and human effort. The flexibility of the new indexes is stretched by the volume and range of files, including some creative attempts by authors to ensure they are indexed under terms not necessarily related to their files.

For Internet Indexes to be even more useful more specific searching needs to be available, i.e. Boolean operators, phrase searching and an indication of contents of the individual index. Connections between professional indexes and searchers and those creating Internet indexes will ensure that Internet indexes develop to become more useful than ever before.

Note: In the time taken to finalise this paper Inktomi has become "Hotbot" with an increased coverage of sites. Keep watching this space!

(1) Adams, Douglas (1979) *The Hitchhikers Guide to the Galaxy*, London, Pan Books p 24

"...he also had a device which looked like a largish electronic calculator. This has about a hundred tiny flat press buttons and a screen about four inches square on which any one of a million 'pages' could be summoned at a moment's notice. It looked insanely complicated...had the words DON'T PANIC printed on it in large friendly letters."

(2) From "Choruses from 'The Rock'" Eliot, T S (1936) *Collected poems 1909-1936*, Faber & Faber, London p 157

(3) The proceedings of the first forum were published as Mulvaney, J. and Steele, C (eds) (1993) Changes in scholarly communication patterns : Australia and the electronic library, Australian Academy of the Humanities, Canberra.

(4) * From Inktomi (<http://inktomi.berkeley.edu/counting.html>) (1 April 1996)

"There are three ways to count URLs for an index:

We count the number of documents actually retrieved from the Web and added to our index. This is a fraction of the total number of distinct URLs contained in those documents. We feel this is the most honest measure of coverage.

Lycos counts unique URLs whether their documents have been retrieved or not. If the lycos crawler comes across a link pointing to a document it gets counted as being in their index. Most of the documents in Lycos's index have never been retrieved. Here is a quote from their news page that provides the breakdown; by our metric they have 1.178 million documents:

July 15th, 1995 Latest catalog: The newest big catalog contains

- 5.07 million URLs, including 1,177,750 files downloaded.
- 5,077,834 unique URLs, including:
- 1,177,750 documents fetched totaling 8,703,484,067 bytes
- 3,900,084 unexplored URLs with descriptions
- 1,834,323,446 bytes of Lycos summaries
- 1,078,127,917 bytes of inverted index.

A third way to count URLs, used by Open Text, is to count the total number of non-distinct URLs in the collection. Thus, if a link to a popular document (such as Yahoo) appears 100,000 times, then it counts as a 100,000 documents."

Bibliography

Adams, Douglas (1979) The Hitchhikers Guide to the Galaxy, Pan Books, London

Basch, Reva (1994) Secrets of the super searchers, Eight Bit Books, Wilton, CT

BELNET (1996) A selection of Internet search tools, <http://www.mtm.kuleuven.ac.be/Services/search.html> 1 April 1996

Courtois, Martin, William Baer, Marcella Stark (Nov 1995). "Cool tools for searching the Web - a performance evaluation" Online Magazine, 19(6) : pp.14-32

Eliot, T S (1936) Collected poems 1909-1936, Faber & Faber, London

Lane, Nancy (1989) Techniques for student research: a practical guide, Longman Cheshire, Melbourne

Lebedev, Alexander (1996). Best search engines for finding scientific information in the Net, <http://www.chem.msu.su/eng/comparison.html> 1 April 1996

Mulvaney, J and Steele, C. (eds) (1993) Changes in scholarly communication patterns : Australia and the electronic library, Canberra, Australian Academy of the Humanities

Network Wizards (1996) Internet domain survey: January 1996, <http://www.nw.com/zone/WWW/report.html> 26 March 1996

Seidelman, R. In, Around and about, September 1995

Stanley, Tracey (1996). "Alta Vista vs. Lycos", Ariadne on the Web, Issue 2, <http://ukoln.bath.ac.uk/ariadne/issue2/engines/> 1 April 1996

University of Michigan School of Information and Library Studies (1996) Matrix of WWW Indices. A comparison of Internet indexing tools, 1 April 1996

Winship, Ian (1995) World Wide Web searching tools - an evaluation, 1 April 1996

© 1996 Roxanne Missingham. To be printed in the AusSI Newsletter and LASIE 27(3):32-42.