Is there an expert search strategy for the world wide web?

PUBLISHER

Loughborough University

REPOSITORY RECORD

Horrell, A. Mark. 2021. “Is There an Expert Search Strategy for the World Wide Web?”. Loughborough University. https://doi.org/10.26174/thesis.lboro.14784510.v1.

Is There An Expert Search Strategy For The World Wide Web?

by

Mark Horrell, B.Sc.

A Master's Dissertation, submitted in partial
fulfilment of the requirements for the award of
Master of Science degree of Loughborough University.

September 1996

Supervisor: Alan Poulter, M.A., M.Sc., A.L.A.
Department of Information & Library Studies

Abstract

Conceived in 1969 by the American Defense Department, the Internet has become a huge source of greatly diverse information. The World Wide Web is introduced as the most recent and accessible area of the Internet, and its access method of hypertext is described, along with its limitations as a retrieval mechanism. Some examples of the sorts of information available on the World Wide Web are described in depth, and other sources of information available on the Internet, such as FTP and Gopher are touched upon briefly.

The issue of searching for information on the World Wide Web is approached and some of the retrieval tools available for this purpose, such as Web search engines, are explored. It is concluded that Web search engines are still the most reliable tools available for information retrieval on the Internet and should any expert search strategy exist then it will rely heavily on the use of them.

A classification of Web search engines based on factors such as design, retrieval mechanism and documents indexed, is formulated, and this is used to explore the possibility of an expert search strategy dependent on the particular information that one is looking for. Existing search engines, from each of the classes arrived at in the classification, are compared and contrasted with particular reference made to their size, ease of use and when they might be used.

Future developments in information retrieval on the World Wide Web are discussed from the point of view of how they may be introduced into a search strategy. It is concluded that, given the huge range of information available on the World Wide Web, any search strategy will have to be extremely general, but there are many guidelines which can be followed, and these are described. Mention is made of the speed of development of new technologies on the World Wide Web, and the importance of constantly updating the final strategy is stressed.

Acknowledgements

I would like to thank my supervisor, Alan Poulter, for his invaluable comments, advice and support throughout this project. I would also like to thank Goff Sargent, Shirley Horner and Karlis Kreslin with their help in sorting out any problems I may have had involving the use of the departmental computer network.

IS THERE AN EXPERT SEARCH STRATEGY FOR THE WORLD WIDE WEB?

MSc Dissertation
Mark Horrell
Dept of Information & Library Studies
Loughborough University.

CONTENTS

LIST OF PLATES

IS THERE AN EXPERT SEARCH STRATEGY FOR THE WORLD WIDE WEB?

MSc Dissertation
Mark Horrell
Dept of Information & Library Studies
Loughborough University.

Chapter One -

The World Wide Web and its Evolution as an Information Source

1.1 Introduction

1.1.1 The Internet

In 1969, the American Defense Department were looking for a nuclear attack resistant method of exchanging information, both scientific and military. ARPAnet was dreamed up. Consisting of a network of computers permanently linked through high-speed connections, it was the grandfather of the Internet. Little by little, the network expanded as more computers were connected and, eventually, other networks also became attached. The National Science Foundation network (NSFNET) joined in 1985 and research agencies and universities were linked across North America. In the UK, the Joint Academic Network (JANET) was established, again linking academic institutions throughout the land and, in turn, this fed into the North American network.

During the 1980s, the Internet was still very much the playground of the military and academia and certain guidelines even banned the use of it for commercial purposes under suitable use policies. As the Cold War collapsed towards the end of the decade, however, the nature of the Internet changed irreversibly. The term *information superhighway* was coined by US Vice-President Al Gore and a whole philosophy of openness in terms of provision of information was born, coupled with a powerful belief in the Internet as the ultimate democratic forum providing freedom of speech worldwide. Anybody with an Internet connection can send an e-mail message, post to a newsgroup or bulletin board, or upload a file for others to read, yet nobody need reveal their age, sex, colour or religion. In this climate, an abundance of information of vastly ranging accuracy, reliability and authority, appeared on the network. There were more developments to come too. The appearance of the World Wide Web in the 1990s provided previously unseen potential for cheap advertising and desk top publishing. Information available on the Internet expanded still further as companies set up their *Web sites* and individuals provided us with their, sometimes shamelessly egocentric, World Wide Web *home pages*. By the end of 1995, nearly 6 million computers worldwide were linked to the Internet, via more than 5000 networks in 33 different countries[1].

Today there is a veritable mine of information available on the Internet, and almost all of it is free provided one has a computer with an Internet connection. The sheer volume of information coupled with the almost total lack of organisation, however, can make retrieving the information required a real problem, particularly if one is new to the ways of the Internet. Yet because of the Internet's potential as an information source, were it possible to become accustomed to the way that information is arranged, there may well be scope for an "expert" information professional to situate his or herself between the interface with the Internet (which at the time of writing means the Netscape *Web browser* or Microsoft's *Internet Explorer*, which provide access to the World Wide Web, Gopher servers, newsgroups, FTP sites and e-mail in a relatively user-friendly form) and the end-user or client requiring information in the first place. The aim of this document is to investigate firstly whether it is possible to attain expert knowledge of information available on the Internet and, if so, whether there is a certain strategy the expert can follow when retrieving information stored there.

1.1.2 The World Wide Web

Before we start, it is worth digressing slightly. By far the most accessible area of the Internet is the World Wide Web and the rate that it is expanding means that it is invariably the starting point for anyone looking for information on the Internet.

Although an expert information professional would have to consider all areas in his or her search for information, none of which will be neglected in this document, the main focus will be on the World Wide Web. It is also worth bearing in mind that most areas of the Internet can be accessed using the Netscape web browser, suggesting that it may not be necessary for the information professional to move from this interface. For this reason, it would be pertinent to consider briefly the history and development of the World Wide Web, to give an idea of the nature of the information stored there. Armed with this, the information professional can begin to have an idea of how the information is arranged and this should help formulate his or her strategy.

The World Wide Web was developed at CERN (the European Laboratory for Particle Physics) at Geneva in the early 1990s. Its principle method of organisation is hypertext; passages of text (nodes) joined together and accessed from and to one another by a series of links. Clicking on a link in a passage of text takes the user to the next relevant passage of text. The World Wide Web's creator, Tim Berners-Lee[2] , had been working with hypertext ideas for nearly a decade. He was known to be influenced by the work of Ted Nelson, who formulated the idea of *Xanadu*, a project which involved linking together the text of every published work on computer using hypertext links[3] .

In 1980, Berners-Lee wrote a notebook programme entitled "Enquire-Within-Upon-Everything", which used hypertext links, while acting as a consultant for CERN[4] , but it was not until CERN joined the Internet in 1988 that Berners-Lee had a chance to develop his ideas into something as tangible as the World Wide Web[5].

The Web's development was rapid indeed. In March 1989, Berners-Lee first circulated his document, "Information Management: A Proposal". The document was, "an attempt to persuade CERN management that a global hypertext system was in CERN's interests," and focussed on the fact that CERN is a large organisation of many constantly changing parts, as well as a high staff turnover. Berners-Lee was concerned at the loss of information in an organisation where the average length of stay for an employee is about two years, and most information tends to be circulated by word of mouth, meaning experience of the organisation is essential. A database of information about CERN for new and existing employees was necessary, and he expressed dissatisfaction with more commonly used information systems such as tree structures, where related pieces of information are often found on different branches of the tree, and keyword searches, where two people may choose different keywords for an identical subject.

He proposed hypertext as a way forward, with entities within the organisation, such as people, concepts and projects, representing the nodes, and relationships between entities, such as, "Person A is working on Project B," representing the links. He identified a number of criteria relevant to CERN that such an information system must have:

- CERN's scattered nature meant that remote access was essential;
- the same data would have to be accessible from different platforms, such as VAX, Macintosh and Unix;
- the organisation tended to expand in unpredictable ways, so it was essential that the system should be non-centralised and allow expansion from all sides;
- existing databases had to be easily incorporated.

It is interesting to note that only passing reference was made to multimedia documents containing graphics, sound and video. Berners-Lee stated that he, "will not discuss this latter aspect further here, although I will use the word *Hypermedia* to indicate that one is not bound to text." In actual fact, the incorporation of multimedia facility was one of the principal reasons for the Web's rapid growth in popularity.

Importantly, he suggested that, "CERN is a model in miniature of the rest of the world. CERN meets now some problems which the rest of the world will have to face soon." He clearly envisaged the expansion of his local information system into what would become known as the World Wide Web[6].

By the end of 1990, CERN had employed the help of Robert Cailliau[7] and technical student Nicola Pellow[8], to aid in its development. By 17 May 1991, the World Wide Web was up and running on CERN machines, and by October of that year, enough interest had been sparked in the World Wide Web to warrant the starting of a mailing list `www-interest` (now `www-announce`) along with much discussion in the newsgroup `alt.hypertext`. CERN demonstrated the World Wide Web to interested parties throughout 1992 and various Web browsers were developed as the interface between Web and client became gradually more user-friendly. In September 1993, Marc Andreesen[9] from the National Center for Supercomputing Applications (NCSA) at the University of Illinois completed the *Mosaic* web browser, which was the first working browser for all of the most common platforms: X-Windows, MS-DOS/Windows, and Apple Macintosh.

At last, the popular press started taking an interest in the World Wide Web as the *New York Times*, *The Guardian* and *The Economist* all published articles about it. The release of Mosaic had coincided with the arrival of colour images on the World Wide Web and, by March 1994, the browser had become so popular that its developers saw fit to leave NCSA and set up their own software company, Mosaic Communications Corporation, later to become Netscape. And still the Web develops and expands. The First International WWW Conference was held in Geneva on 25-27 May 1994[10]. In February 1995, the European Commission deemed it important enough to warrant discussion at a G7 summit meeting in Brussels. Text, colour images, sound, video and even interactive feedback via a Common Gateway Interface (CGI) are all possible on the World Wide Web and a recent development has been the use of the Java programming language to allow mini-application programmes ("applets") to be run through Netscape.

1.1.3 Some Problems

Despite its global nature, there are still many problems associated with the Web as an information source, particularly with regard to the use of hypertext as a retrieval mechanism. Some of these problems are outlined below.

1.1.3.1 Difficulties with Hypertext

- The link mechanism is overstrained as a searching device. It is fine for a small local information source with only a handful of links which can be clearly labelled in a single contents page, but in a system as vast as the World Wide Web, it quickly becomes unmanageable. It is also worth noting the similarities between a small local source which uses hypertext and a tree diagram/menu structure and asking whether it is worth using hypertext at all for such a system.

- A commonly reported problem with using hypertext is that of navigation and the possibility of becoming *lost in hyperspace*. McKnight et al. suggest a definition of "navigation difficulty" originally put forward by Elm and Woods[11]. They describe it as, "users not knowing how the information is organised, how to find the information they seek or even that information is available."[12].

They propose three levels of representation developed by a person as he or she attempts to navigate space: landmarks, routes, and surveys:

- *Landmark knowledge* is characterised by our representing our awareness of where we are in space, "in terms of highly salient visual landmarks in the environment, such as buildings, statues, etc."

- *Route knowledge*, "is characterised by the ability to navigate from point A to point B, using whatever landmark knowledge we have acquired to make decisions about when to turn left or right," and is therefore a step more advanced than landmark knowledge due to the initial reliance upon it.

- *Survey knowledge* is described by McKnight et al. as, "a world frame of reference rather than an ego-centred one." It is the ability to give directions and travel routes not previously seen based on advanced landmark and route knowledge of an area, along with the ability, "to know the general direction of places, eg. 'westward' or 'over there'."[13]

In order to apply this model to hypertext and, ultimately, the World Wide Web, it is necessary to draw parallels between landscape, route and survey knowledge in the world around us and in hyperspace.

Although aids to navigation exist on the Web, such as a home page configured through the Web browser, bookmark lists and history lists of documents that have been recently viewed, applying any of these to the above criteria is problematic.

In conclusion, although there are definite parallels between navigation in hyperspace and in the world around us, there would seem to be fewer aids to navigation for the traveller in hyperspace.

1.1.3.2 Difficulties with the Web in General

Outlined below are just a few of the problems confronting the World Wide Web in its capacity as an information source.

- There is no built-in search mechanism for locating documents on the World Wide Web. Despite the fact that there are many different types of data available, hypertext links between documents remain the only structuring mechanism for bringing them together. The Web's nature is browsing-centred rather than searching-centred and relies on Web users following appropriate (and sometimes, inappropriate) links in the hope of finding a relevant document. Furthermore, many servers contain links to only their own documents and the opportunity to browse across different servers is therefore lost.

- It is possible to find a given document on the Web one day, yet discover it not to be there the next. Frequently, useful links are moved to a different location on the whim of a Webmaster. Often, a pointer is provided to the new location at the old one, but this is by no means always the case. Another drawback is that these pointers do not remain permanently, so if one is trying to find a document a long time after it has been moved, one may not be able to find it.

- A similar problem occurs if one is applying an identical search strategy to relocate a document previously found. Search engines and Web databases tend to be updated so frequently that an identical search will generate a different result a week later.

- Pages on the Web are often updated meaning that previously available information is deleted and no longer available online. Similarly, if vital new information is added to a page, there is no standard way for a Web user to know that the page has been updated. Recently, online software has been created which addresses this problem. An example is the URL-Minder[14]. To use this service, all one has to do is submit the address of a Web page and an e-mail message will automatically arrive whenever the page is updated. Software, such as *WebWatcher*, developed by Carnegie Mellon University[15], also exists for this purpose

- Although a browsable list of links to available pages of information may seem like a reasonable if basic tool to aid searching, even this can be problematic given the Web's somewhat quirky nature. The absence of two-way links means that links cannot be traced backwards to a server's home page should the address of a site change.

- An increasing problem is the inability of Web search engines (see chapter 2) to keep track of pages that have been updated, deleted or moved.

1.2 Information Resources on the Web

A problem the Internet has always had is that of regulation, resulting in a decided lack of organisation. The Web's rapid increase in popularity over the last couple of years has resulted in a highly eclectic mix of information. It has been suggested of the Internet that the, "vast amount of information has been compiled by enthusiasts, volunteers and interested parties; as an information resource reflecting the spectrum of modern culture, society and human endeavour, the Internet is thus unparalleled."[16]. Although attempts to comprehensively categorise this massive range of information must surely be doomed to failure, it is worth examining it in a little more detail so that the information professional can at least be given some idea of its nature.

Diagram 1.1 gives a broad overview of the types of pages one may find when browsing the World Wide Web. It is roughly hierarchical in structure, having its origin in the more traditional tree structure that some information systems have tended to take. For instance, from a given academic institution's home page, one may find access to various departmental home pages within the institution. These in turn may provide links to the text of papers published by personnel in each department. What must always be borne in mind however, is the nature of hypertext, meaning that a given point on the diagram may provide a direct link to any other point on the same diagram. It should be stressed that none of these information sources need be textual. Images, video and sound are all available from most of the resources indicated on the diagram. To give a more precise idea of what a World Wide Web home page may contain, examples of some of the more typical web sites have been provided below.

1.2.1 Campaign Organisation - Friends Of The Earth

(http://www.foe.co.uk/)

The Friends Of The Earth Home Page is a good example of how the World Wide Web can be used both to educate and to campaign for a certain issue. Its welcome page gives a brief introduction to the organisation and what it stands for. As with many home pages on the Web, this is often quite partisan, and prominent is the hard sell: as well as justifying their existence, they appeal to the reader's conscience in explaining why he or she should join the organisation. One of the pages is entitled, "5 good reasons to support Friends Of The Earth," and details of how to join are provided, including a database of their 250 or so local branches throughout the UK, accessed by a search

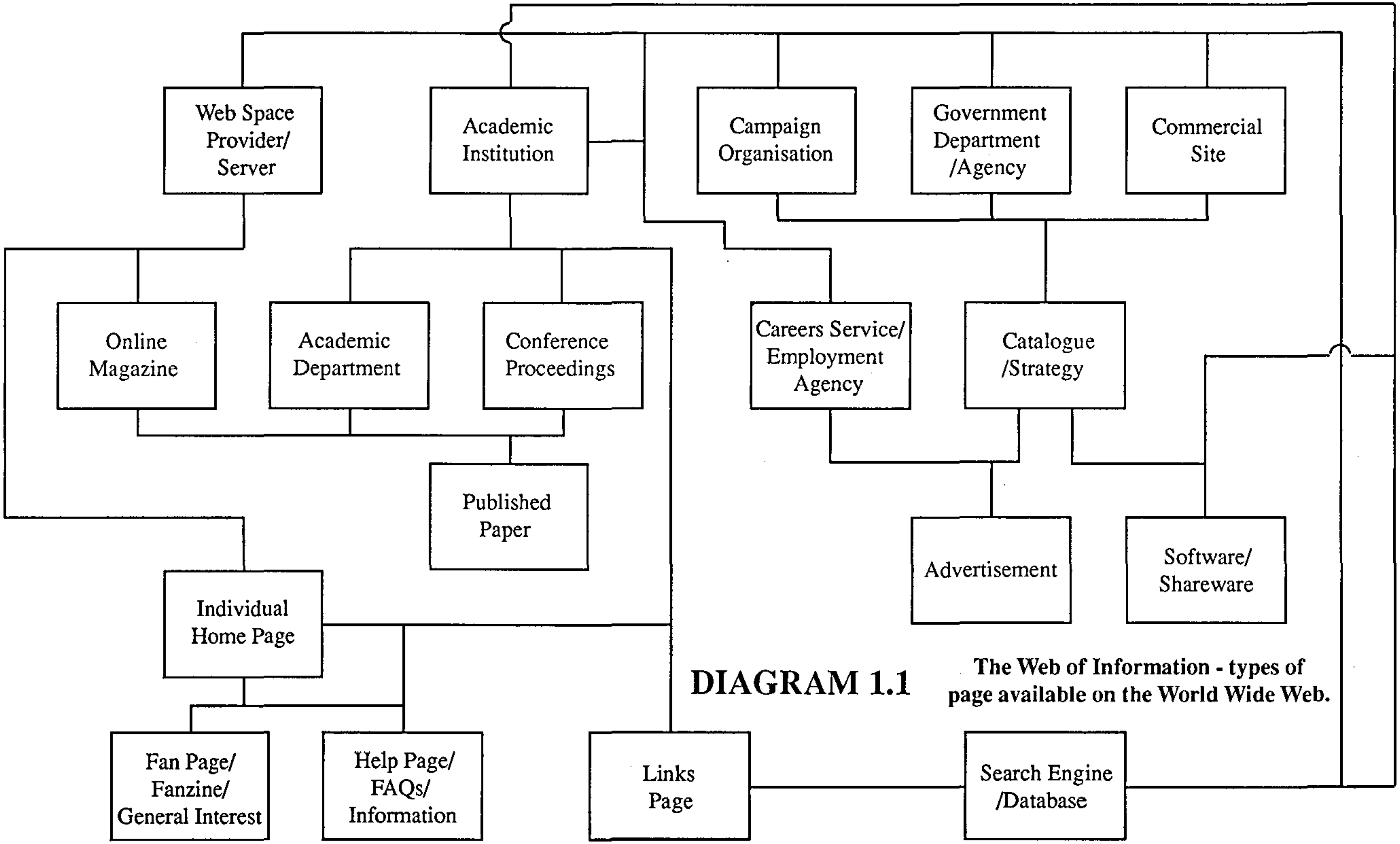


DIAGRAM 1.1 **The Web of Information - types of page available on the World Wide Web.**

engine interface which asks for the searcher's postcode.

The front page contains a graphical link to another page celebrating 25 years of Friends Of The Earth. This page gives a really quite comprehensive history of the organisation, including details of their most successful and high-profile campaigns.

There is a news service providing information on current issues in which Friends Of The Earth are involved. An example, at the time of writing, is the Newbury By-Pass Campaign. This contains links to newsgroups currently discussing the issue, such as `talk.environment` and `uk.transport`, and to other relevant home pages, such as the government Highways Agency[17].

Perhaps the most useful Internet information source provided by Friends Of The Earth is the page entitled, "Information Resources at FoE." This includes the text from press releases, information sheets and leaflets, and details of upcoming events and fundraising news. There is a page describing current FoE-funded projects, and an explanation of how they use and gather environmental data. They even include links to the entire text of various articles published by them in academic journals and conference proceedings. Finally, there is an exhaustive link list to other environmental resources to be found on the Internet. This is in subject tree format with headings ranging from *Biodiversity and Habitats* to *Transport and Planning*.

1.2.2 Government Department - Her Majesty's Treasury

(http://www.hm-treasury.gov.uk/)

The front page of the Web site for Her Majesty's Treasury presents the reader with an icon-driven menu of resources. A good starting point here is the, "About HM Treasury" page, which provides the full text of the last two departmental reports, outlining aims and objectives for the year ahead. There is an office directory listing the names of personnel, and a more detailed description of the department's organisation, describing its management structure and code of practice. There is also a listing of its publications in print (although no link to the actual text of these publications) along with details of how to obtain them.

The resource includes a section on treasury news, with such details as new appointments, recent meetings and press releases. As with the FoE home page however,

the reader must be aware of the partiality of the information provided, which tends to be in the manner of soapbox soundings. The news section also contains a link to the Central Computer and Telecommunications Authority (CCTA) Government Internet Service, which is the main Internet resource for all government departments.
Naturally, the site contains economic information, such as news of monthly monetary meetings, debt management reports and details of the past two years' Budgets, including links to other treasury-related news releases. There is an extremely comprehensive resource containing the full text of speeches by government ministers, again covering the last two years.

Finally, there is a search engine covering information contained on all UK national and local government Web sites.

1.2.3 Employment Agency - Jobserve

(http://www.jobserve.com/)

Jobserve is not an employment agency as such, but more a resource containing information collected from many genuine employment agencies. It is updated daily from faxes received from IT-related recruitment agencies throughout Europe, and features lists of current job vacancies sorted by region. Otherwise, vacancies are listed in no particular order, although the site does include its own keyword search engine.

Additional resources on this site include introductory information about Jobserve, statistics and contact details, and directory of the several hundred recruitment agencies associated with it, along with links to their home pages.

1.2.4 Academic Institution - University of Bath

(http://www.bath.ac.uk/)

Throughout the 1980s, academic institutions dominated the Internet and although commercial organisations are now making their presence felt, Diagram 1.1 shows how central academia still is to resources on the World Wide Web, providing access to their own academic departments and conference proceedings, hence to published academic papers, to individual home pages, help pages and frequently asked questions, links pages, careers services, shareware and databases. Web servers run by academic institutions are almost certainly the best source of information on the net in terms of authority and impartiality, and the University of Bath's World Wide Web resources are one of the best examples in the UK of academic institutions providing unique services free of charge.

In keeping with all good university Web services, there is general information about the city of Bath, including a brief description and history with photographs, contact addresses and telephone numbers of useful services, such as banks, taxis, public transport and cinemas, and more detailed general information about the university. The latter includes a campus map, photographs, descriptions of the facilities available, such as sporting, medical and computing, Student Record Office statistics, a noticeboard of general announcements, and details of committee meeting, minutes, etc.

More specifically, there are details of academic staff and students at the university, including a search engine for finding details by name, and links to every single personal home page on the University of Bath server, sorted by department. There is information about staff and student services and societies, the staff handbook and calendar of meetings.

There are links to individual department home pages, which in turn include details of ongoing research projects and links to the text of published papers, which can be downloaded in Postscript format or via FTP. As well as academic departments, there are details of the various centres run by the university. One of the most important of these from an information professional's point of view is the UK Office for Library and Information Networking (UKOLN), which describes itself as the, "national centre for support in network information management in the library and information communities."[18]. This particular link includes the full text of reports and papers produced by the centre.

In addition, there is a section entitled, "Searching the Internet," which includes a local University of Bath search engine and a brief guide to some of the more popular worldwide search engines, as well as links to online library catalogue services such as Hytelnet[19], and information regarding e-mail discussion lists. There is even general information about the World Wide Web itself and links to useful Web information sources such as HTML manuals and Thomas Boutell's, "World Wide Web Frequently Asked Questions"[20]. About the only section on the University of Bath web server which could possibly be regarded as commercial is that giving conference booking details, which is really just an advertisement to potential clients and delegates.

Perhaps the most significant feature of all on this server can be found on the link to the University of Bath Library. As well as general information regarding opening hours and the entire library catalogue, there is a gateway to Bath Information Data Services (BIDS), one of the world's most comprehensive online databases of academic journal abstracts, and the full text of *Infobytes*, the Bath University Information Services' newsletter. Most important of all for anyone searching for information on the Web, there is the Bulletin Board for Libraries (BUBL) Information Service[21], which includes a subject tree, available in alphabetical or Universal Decimal Classification (UDC) arrangement, of resources available on the Web. This service is so comprehensive that it will be dealt with in much greater depth in following chapters.

In summary, it can be seen that, in this case, the information which can be obtained from just a single server is enormous and much of it is not something the uninitiated would necessarily associate with the University of Bath. This description has not even attempted to explore the large amount of information stored on individual home pages. Depending on the commitment and zeal of the owner, these pages can often be an important source of information in their own right, particularly in terms of links to other resources, and the chances of finding the really useful ones simply by browsing through are extremely slim. This is a good illustration of the erratic organisation of information on the Web, and the need to tackle the problem in later chapters.

1.2.5 Commercial Site - Netscape

(http://www.netscape.com/)

What better site to examine than the home of World Wide Web browsing? The front page contains a comprehensive list of contents in any one of four languages - English, French, German and Japanese. In keeping with any commercial Web presence, its site features a history of the company along with a comprehensive catalogue of company products and information. Being a software company, this part features a shareware site from where Internet users can download software for use on their own computer. A feature called, "My Page," demonstrates the many different facilities available using the latest beta version of Netscape Navigator. Unfortunately, the Internet surfer naturally needs to be using this particular version in order to be able to view this feature, but if this is not the case, then a prominent icon can be clicked to download it.

The site features substantial information regarding the dozens of new Internet developments Netscape are involved with, both on their own and in partnership with other companies. In the rapidly expanding world of multimedia communications, Netscape stand at the forefront, so this site can be a good place to look for up to the minute information on current Internet technology. One area covered particularly well as that of *Intranets* - Web servers run by companies over their own Local Area Networks rather than the Internet so that company employees use in-house resources rather than external sites.

Netscape also run an extensive Assistance and Learning Center, accessible from their home page. This includes links to their Web developer's newsletter, HTML and web site authoring resources throughout the Internet, ranging from the beginners' to the advanced, and information about the Internet, such as the *Internet Society*[22] and *Electronic Frontier Foundation*[23]. Their Technical Support Team provides technical support both to clients and to servers, and technical documentation is available, such as software licensing standards, network security and hardware requirements. They have details of Netscape User Groups' discussion lists and even run day release courses, information about which can be found here.

Perhaps Netscape's biggest difference over most commercial Web sites is in the number of links they provide to general information resources available on the Web. A section entitled, "Netscape destinations," provides extensive links to such sources as news pages, such as *Reuters*[24] and the *New York Times*[25], technology news, financial news, hardware and software company home pages, sports news, travel companies, entertainment sites, and even places to go Internet shopping. Some of the pages listed here can be obtained directly from buttons visible on the Netscape Navigator browser interface. This includes *Netsearch*, a link to a number of popular Web search engines; *What's New?*, a resource listing many brand new Web pages; and *What's Cool?*, a subjective list of favourite sites selected by Netscape staff.

1.3 Alternative Information Sources on the Internet

Finally, now that some sample Web pages have been described and a reasonable picture is beginning to form as to the type of information available on the World Wide Web, it is worth looking at the other information sources to be found on the Internet. Although these sources are not presented in such an attractive fashion to the user, they can nevertheless be equally useful and can all be accessed through the Netscape Web browser. They should therefore not be neglected by the information professional in his or her search for information. Diagram 1.2 shows the information sources available from various client front-ends. As can be seen, a Web browser such as Netscape can pick up many more sources than just http (World Wide Web) documents. These will be described in a little more detail.

1.3.1 Archie

Archie is the equivalent of a vast searchable catalogue of documents available on "anonymous" FTP sites throughout the world. There are a number of sites around the world offering an Archie search utility, accessed by a variety of methods, although the searcher would normally have to know the precise name of the file being searched for. Archie search software allows one to search for files using a command line interface such as DOS or Unix. In this case the command would usually be something along the lines of "archie -s filename" and the output would give names for the file's site and directory. If Archie software is not available, numerous sites offer addresses one can telnet to carry out the search. A UK example using this method can be found at Imperial College, London, whose address is *archie.doc.ic.ac.uk*. But Archie request forms can also be found at http addresses throughout the Web, where the search procedure is considerably more user-friendly, and the interface is not dissimilar to the various Web search engines described in the following chapter.

1.3.2 WAIS

A Wide Area Information Service (WAIS) is a large set of indexed databases on a variety of topics. These databases are distributed across the Internet and can be searched by keyword. Most interfaces to WAIS clients work in a similar way, whereby the user selects a set of databases to be searched and formulates a keyword enquiry. Again, they can be accessed from the Netscape Web browser. Numerous guides to using a WAIS can be found on the World Wide Web, an example of which can be found at the *Infoville Schoolhouse*[26].

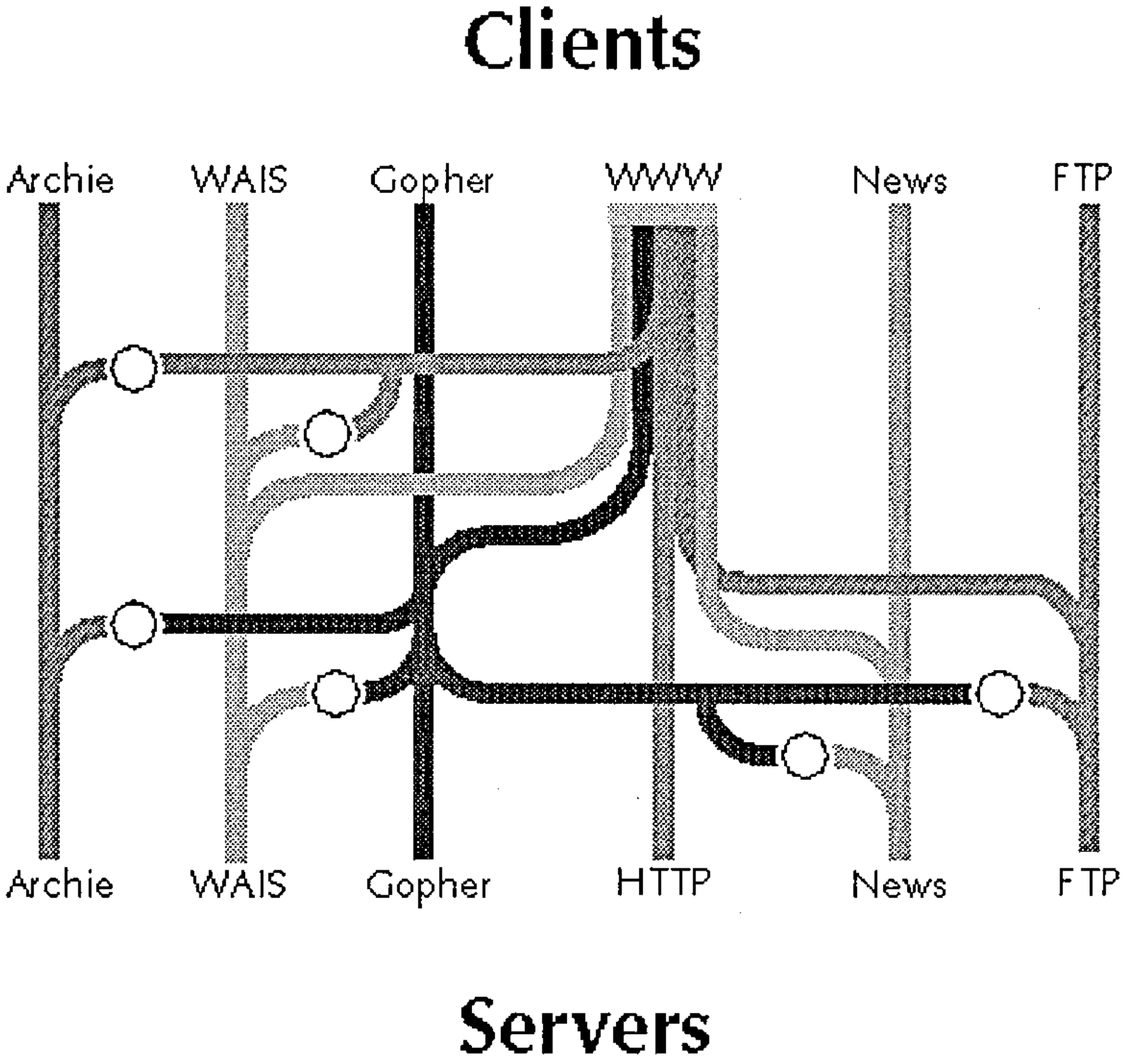


DIAGRAM 1.2
Internet information sources and how to access them.

Diagram taken from CERN's Media Day,
(http://www.cern.ch/CERN/WorldWideWeb/RCTalk/howitworks.html)
and reproduced courtesy of Robert Cailliau.

1.3.3 Gopher

Gopher can be looked upon as the forerunner of the World Wide Web and is a text only information source accessed by a system of menus. It is very easy to browse and also has its own search tool, Veronica (rather comically standing for Very Easy Rodent-Oriented Netwide Index to Computer Archives) which is again available at a number of sites dotted around the world. Unfortunately, many Gopher sites have now been superseded by Web equivalents and the information found on them may have become out of date, but in many cases Gopher should still be regarded as a valuable information source.

1.3.4 USENET Newsgroups

USENET is a huge bulletin board of independently submitted information on a whole range of topics, arranged under a system of *newsgroups* or *discussion groups*. Subscribers to these newsgroups send messages (or *postings*) to the server administering the group which in turn puts the messages up on the bulletin board for other subscribers to read and reply to. Newsgroups are accessible from Netscape, and there are even resources on the Web, such as *tile.net*[27] which offer complete lists of USENET newsgroups available.

1.3.5 FTP

FTP or File Transfer Protocol allows one to log onto another computer somewhere in the world and retrieve a file to bring back to the user's own computer. Usually, permission to access the remote computer would require a login identification and password, but a number of "anonymous" FTP sites can be found on the Internet wherein access is allowed to all, usually by their logging in as "anonymous" and using their e-mail address as the password. As mentioned earlier, these anonymous FTP sites have been indexed in the huge archie catalogue[28].

1.4 Summary

It can be seen that there is a vast range of information available on, not just the World Wide Web, but the whole of the Internet. The information professional must be aware of the very different sites where this information can be found, and the lack of organisation can make this very difficult indeed. Not only is some sort of strategy needed before one can begin to wade through it all, but clearly some tools will be needed to assist in the task.

In the following chapter, one such tool, a Web search engine, will be examined in detail, but before moving on to describe this it remains to define what is meant by an "expert search strategy". An *expert*, in this case, is not someone possessing expert knowledge in a specific field, rather, it is assumed that he or she is a skilled searcher with a background in information retrieval but is new to the concept of searching the World Wide Web. The searcher knows what he or she is looking for, but how do they go about finding it on the Web? Are there simple steps which can be followed and, if so, what are they? This report will examine the tools available to the searcher, depending on the information sought, and outline the appropriate steps to take in retrieving this information from the World Wide Web.

REFERENCES FOR CHAPTER ONE

1. **Poulter, Alan.** *Web search engines: a critical review.* In: *Program* (in press).
2. *People involved with the World Wide Web Consortium.* URL: http://www.w3.org/pub/WWW/People/W3Cpeople.html#Berners-Lee. 2 Sep 1996.
3. **Nelson, Ted,** 1986. *A technical overview of the Xanadu electronic storage and publishing system* [Videocassette]. Texas: Fredericksburg.
4. **WWW Consortium.** *A little history of the World Wide Web.* URL: http://www.w3.org/pub/WWW/History.html. 2 Sep 1996.
5. **Segal, Ben M.** 1995. *A short history of Internet protocols at CERN.* URL: http://wwwcn. cern.ch/pdp/ns/ben/TCPHIST.html. 2 Sep 1996.

6. **Berners-Lee, Tim**, 1989. *Information Management: A Proposal.* URL: http://www.w3. org/hypertext/WWW/History/1989/proposal.html. 2 Sep 1996.

7. For biography see: *People who have contributed to the World Wide Web project.* URL: http://www.w3.org/pub/WWW/People.html. 2 Sep 1996.

8. Ibid.

9. Ibid.

10. *First WWW conference.* URL: http://www1.cern.ch/WWW94/Welcome.html. 2 Sep 1996.

11. **Elm,W. & D.Woods**, 1985. *Getting lost: a case study in interface design.* In: *Proceedings of the Human Factors Society 29th Annual Meeting.* Santa Monica, CA: Human Factors Society, pp.927-931.

12. **McKnight, Cliff, Andrew Dillon & John Richardson**, 1993. *Space - the final chapter, or why physical representations are not semantic intentions.* In: *McKnight,C, A.Dillon & J.Richardson, 1993. Hypertext: a psychological perspective.* New York: Ellis Horwood, p.170.

13. **McKnight, Cliff, Andrew Dillon & John Richardson**, 1991. *Hypertext in context.* Cambridge: CUP, p.69.

14. *The URL-minder! Your own personal Web robot.* URL: http://www.netmind. com/URL-minder/URL-minder.html. 2 Sep 1996.

15. *WebWatcher home page.* URL: http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/ People/webwatcher/. 2 Sep 1996.

16. **Tseng, Gwyneth, Alan Poulter & Debra Hiom**, 1996. *The library and information professional's guide to the Internet.* London: Library Association. p.91.

17. *Highways Agency home page.* URL: http://www.open.gov.uk/hiagency/highhome. htm. 2 Sep 1996.

18. *UKOLN: UK Office for Library and Information Networking.* URL: http://ukoln.bath.ac. uk/about.html.

19. *Hytelnet information page.* URL: http://www.lights.com/hytelnet/. 2 Sep 1996.

20. *World Wide Web FAQ.* URL: http://info.ox.ac.uk/help/wwwfaq/index.html. 2 Sep 1996

21. *BUBL Information Service Web Server.* URL: http://www.bubl.bath.ac.uk/BUBL/ home.html. 2 Sep 1996.

22. *Internet Society home page.* URL: http://www.isoc.org/. 2 Sep 1996.

23. *EFFweb - the Electronic Frontier Foundation.* URL: http://www.eff.org/. 2 Sep 1996.

24. *Reuters - the business of information.* URL: http://www.reuters.com/. 2 Sep 1996.

25. *The New York Times on the Web.* URL: http://www.nytimes.com/. 2 Sep 1996.

26. *Infoville Schoolhouse: WAIS.* URL: http://canyon.ucsd.edu/infoville/schoolhouse/wais.html. 2 Sep 1996.

27. *tile.net.* URL: http://tile.net/. 2 Sep 1996.

28. Rankin, Bob. *Beginners' [guide to the Internet].* In: *The Web*, **6**. May/June 1996, pp.52-53. Macclesfield: IDG Media. URL: http://www.wcentral.co.uk/listings/beginners/index.html, 21 May 1996.

Chapter Two -
Web Search Tools

2.1 Web Searching - Some Solutions

It has been pointed out in the first chapter that there is no built in search mechanism attached to the World Wide Web. Furthermore, the eclectic mix of information coupled with its lack of any sort of organised structure mean additional tools are required before one can begin to think about locating Web documents.

2.1.1 Fish Searching

Many solutions have been proposed. De Bra and Post[1] describe the concept of *fish searching* which is a function of the browser software (in this case *Mosaic for X*) allowing one to keyword search a given page and its links. The fish metaphor is used because the fish search algorithm is supposed to simulate a school of fish breeding and searching for food.

A good starting point is found by using a well-connected document (ie. one with plenty of relevant links) from experience or from an existing keyword search engine. This represents a school of fish and each URL leading from it a single fish. Each time a fish detaches itself from the school, how long it survives and how many offspring it spawns is dependent on the amount and quality of food that it finds. Similarly, how long one follows a search trail picking up linked hypertext documents depends on the number of linked URLs from each document and the relevance of them to the subject being pursued.

Fish searching follows this same principle. A single document is selected as the starting point and depth-first navigation is used to find the largest possible number of cross-referenced links. The retrieved documents are then scanned for relevant information at the client end. Those documents deemed relevant are also scanned for links to further relevant documents, and this process is followed until the number of relevant links has been exhausted (or to use the school of fish metaphor, the fish population has died out).

The practice of highlighting links to nodes which have already been visited is used to prevent repetition. The whole process differs significantly from most Web searching facilities in one important respect: the Web browser at the client end of the network is used to perform the information request, rather than a remote server[2].

2.1.2 WAVE

Kent and Neuss[3] have used the mathematical theory of concept analysis. They describe WAVE - *Web Analysis and Visualisation Environment* - a 3D interface for Web information visualisation and browsing. *Concept scales* are used to rate characteristics such as location, title, keywords, topic, size, and modification time, and these provide a mechanism by which *objects* (or documents) can be grouped together in a conceptual information space. These groups are called *conceptual classes*, and Kent and Neuss describe them as, "any group of entities or objects exhibiting one or more common characteristics, traits and attributes." Objects and their characteristics are connected by a *formal context*: a triple <G,M,I> containing the two sets G and M, and the relationship I, where the elements g of G are objects, and the elements m of M are characteristics. Thus "gIm" asserts that, "object g has characteristic m." WAVE uses both concept analysis and the more traditional library tool of classification and the entire process follows three distinct phases: data acquisition, analysis and classification, and visualisation (or "interactive browsing").

2.2 Web Search Engines and Databases

While such examples may exist in theory, although fish searching was programmed into the *Mosaic for X* Web browser, neither solution has been implemented to any great degree in practice. There have been, and indeed still are, more practical tools however, and this chapter introduces the most important of these - the Web search engine. Before the existence of the search engine, Internet users had to resort to less reliable and systematic methods of retrieving information, such as asking subscribers to mailing lists or newsgroups whether or not they knew where to find given resources on the Internet, or simply by browsing through links in the hope of finding something relevant, an increasingly unreliable and time-consuming method to employ given the present hugely rapid expansion of the Web.

As time went on, regular Web users started assembling their own directories of useful sites and adding links to these sites on their own Web pages. Some of these grew into important resources. Scott Yanoff's *Internet Services List*[4], for instance, is still used by many people today, and a list called *Yahoo!*[5], started in April 1994 by Jerry Yang and David Filo, two Ph.D. students at Stanford University, is now a highly profitable limited company and one of the most important information resources available on the Web. Both of these are examples of Web search engines and it would be appropriate at this moment to outline a definition.

A *Web search engine* has been defined as:

> *"a retrieval tool, consisting of a database (or databases) including resources available on the World Wide Web, search software and a user interface available via WWW."*[6]

The earliest search engine, Archie, an FTP file retrieval tool originally based at McGill University, has already been described in chapter one. The original Archie database was built by sites registering their FTP files using a special template containing the fields necessary for Archie to construct each record. The onus was therefore on the owners of the information to make their resources known to the outside world. This is unusual. In the past with directories which have been compiled in hard copy, it has been the job of the compiler to assemble the database of information resources. Most Web search engines, however, combine the two approaches. Database compilers use software or browsing to assemble their catalogues but usually include an option available via their search engine's Web interface which allows owners of information to submit the URLs of Web pages.

2.2.1 Types of Web Database

There are now quite literally hundreds of search engines available for retrieving information on the Internet and World Wide Web and choosing which one to use can be based on a number of factors including personal preference, ease of use, the type and subject of information being searched for, or even just awareness of the search engine's existence.

2.2.1.1 Database Assembly Method

At the basic level, Web databases tend to fall into one of two categories:

- *Manually-assembled databases* such as *Yahoo!*[7] are compiled by human experience. Sites are read and reviewed (very often being given a rating) by staff and, if deemed worthwhile enough are classified into a given subject category and added to the database. Documents are then retrieved by following branches of a subject tree until relevant information is found. Such databases tend to be smaller given that it takes the time and effort of human beings to compile them but, for the same reason, they can nearly always be relied upon to point to information of a guaranteed quality.

- *Automatically-assembled databases* such as *Lycos*[8] and *AltaVista*[9] are compiled using a piece of software known as a *robot* (also known as a *spider* or a *wanderer*). A robot functions by starting with a given URL then recursively retrieving linked URLs and adding them to its database, thus a large catalogue can be built up very quickly with minimal human effort. Such databases can be massive (Alta Vista was set up in the Autumn of 1995 and by May 1996 estimated its database to contain some 30 million URLs!), but because of its automated nature no editorial control is exercised over the quality of documents indexed.

The above description is described as being at the basic level because as search engines are developing and improving, elements of each tend to merge and collide. Yahoo, for instance (as indeed do all good manually-assembled Web databases) contains a keyword search facility for accessing documents on its database. Conversely, Lycos contains a sister service, known as Lycos A2Z[10], of selected documents from its automatically-assembled database reviewed and arranged by subject, which can then be retrieved by selecting links from a menu.

The key issue here, therefore, is one of quality versus quantity. Such is the size of a database such as Alta Vista, for instance, that one can very often assume that it has a relevant document inside it somewhere. Indeed, a single keyword search can often generate quite literally thousands of results. A good query language allowing more precise pinpointing for a qualified information specialist (and Alta Vista does indeed offer a very powerful *advanced search* feature) may therefore be all that is required. On the other hand, a searcher may not wish for thousands of results to sift through for relevant information. He or she may want only ten results returned as long as at least some of them are relevant. In this case a manually-assembled database such as Yahoo! may prove the better option.

2.2.1.2 Document Retrieval Method

This brings us conveniently to the next distinction to make among the various Web search engines available: that of the keyword search engine versus the classified index.

- *Keyword search engines* can vary enormously in the complexity and power of their search interface. At one end there is the very basic keyword search form such as that offered by the *CUI W3 Catalog*[11], although this does offer a more advanced search facility for those familiar with the Perl programming language. The basic keyword search usually allows the searcher to enter a string of words on a single line and then defaults to dividing them by the boolean OR operator. The keywords can then be searched for in the document title, URL, HTML *meta* tag, first few lines of text or, in some cases, the full text of an article.

  Quality of retrieval in this case can vary enormously depending on the particular *relevance ranking* algorithm used by the search engine (see section 2.2.3), and a particularly annoying feature common to many keyword search engines is their tendency to *substring search* (ie. to truncate search terms automatically) so that a search for, say, "anthem," might retrieve articles on the singer Van Morrison containing the expression, "V<u>an the M</u>an"!

  At the other end of the scale are extremely powerful search interfaces offering all the usual tools available to professional online searchers, such as nested boolean queries containing the operators AND, OR and NOT, truncation (but not automatic truncation), and the NEAR operator for finding words commonly associated with one another. Examples of search engines with particularly powerful search interfaces include Alta Vista and *Open Text*[12].

- *Classified indices* can also vary in the form that each one may take. Some, such as Yahoo!, use their own self-devised subject-based system of classification, listing documents alphabetically in some sort of organised tree structure. How useful one finds these resources from the point of view of searching and browsing depends very much on how accustomed one gets to the various subject categories available, and how logically their subcategories appear to be linked together. Comprehensive cross-referencing can usually be an enormous help when using such a resource.

Other services take a more traditional approach by arranging documents according to various library classification schemes. This can be problematic for some databases given the huge diversity of documents available on the Web, from academic documents through hobbyists home pages to commercial sites and advertisements, and the range of (often unrelated) information available on a single page may also make classification difficult. However, when a resource is indexing academic documents only, such a method seems to prove quite effective.

A good example of this is the *Bulletin Board for Libraries (BUBL) Information Service*[13], based at the University of Bath, which indexes mainly UK academic resources and can be arranged either as a long alphabetical list or under Universal Decimal Classification (UDC) order. A more eclectic example is the *World Wide Web Virtual Library*[14], which arranges its hundred or so subjects in an alphabetical list or under Library Of Congress classification. Beyond each single heading classification can vary considerably since the resource is maintained by many different people around the world, each assigned a single specialist subject area. Also bracketed under this category are subject-based services such as *SOSIG, the Social Science Information Gateway*[15], which can be arranged under alphabetical or UDC order.

2.2.1.3 Search Engine Hierarchies

Other varieties of search engine have emerged as Web operators have striven to perfect the Internet search tool. Hierarchies of search engine exist. These can be in the form of a multi-engine search page, a meta-search engine, or a search engine of search engines.

- *Multi-engine search pages*, such as the *All-in-One Search Page*[16] contain search forms for a number of different search engines collected together in the same HTML document.

- *Meta-search engines* (sometimes known as *Simultaneous Unified Search Indices* or *SUSIs*) go a step further than multi-engine search pages by actually utilising a single search entered by the end user, searching a number of different search engines and combining the results from all of them in a single response. A good example of this is *MetaCrawler*[17].

- *Search engines of search engines* are are another variant on the multi-engine search page. *C/Net's Search.Com*[18] and the *Internet Sleuth*[19] both appear on the surface to be two more examples of a manually-assembled Web database just like Yahoo, containing both a comprehensive list of subject categories with which the user can browse for documents, and a keyword search facility for more precise pinpointing, but in reality neither search engine contains its own database of Web documents at all; they merely point the searcher in the direction of a subject-specific Web search engine for the particular field they are interested in. Both of these search tools contain databases of over a thousand alternative Web search engines.

2.2.1.4 Vertical Search Engines

Then there is the so-called *vertical* search engine. Again, this can take several forms.

- *Subject-specific search engines.* Some search engines limit themselves to a single topic, supposedly compiling a comprehensive database of Internet resources confined to a narrow subject area, such as SOSIG, mentioned earlier in this section, as well as other services resulting from the eLib *Access to Network Resources project*[20], including *HUMBUL*, the *Humanities Gateway*[21], *ADAM*, the *Art, Design, Architecture and Media Information Gateway*[22], and *OMNI*, *Organising Medical Networked Information*[23]. These can be hierarchical too. The World Wide Web Virtual Library, also mentioned earlier, can be described as a large cooperative database of over a hundred smaller independently run databases each confined to a single subject area.

  Subject-specific search engines exist for a hugely diverse range of topics, from the serious to the distinctly less important, and offer many novel methods of searching. Indeed, those seeking greater censorship of the Internet would be extremely interested to hear about the *Amateur Hardcore Search Engine*[24], a facility which allows users to locate various adult Web sites by keyword searching on particular fetish!

- *Geographically-specific search engines.* Other search engines limit themselves geographically, and even this can be done in a number of ways. Many sites have their own keyword search facility limiting its database to documents contained on a single server, while others search for documents only in a given Internet domain. For instance, the *AC/DC* search engine[25], released upon the Web in May 1996, limits its search to documents contained on servers within the domain `ac.uk`, ie, only those Web servers within the UK academic community. Language can be a particular problem with search engines. While some, such as *SavvySearch,* another meta-search engine[26] claim to index in dozens of different languages, the real question is not whether the index itself is in a given language, but whether the documents it indexes are written in that language. Recently this problem has been overcome by the appearance of national search engines such as *Flipper*[27] which is in German and indexes only German documents, and *Recursos de Internet en Español y Portugués*[28] which is a multi-engine search page claiming to contain all 31 search engines of the Spanish, Portuguese and Catalan worlds.

- *Media-specific search engines.* The concept of using the Web as a multimedia information source will be touched upon later, but the present category should include those resources that specifically seek non-textual information on the Web. A good example of one of these is the *Query By Image Content* service[29].

2.2.1.5 Price

Another issue is that of price. The vast majority of Web search engines still offer their services free to Internet users. The sheer number of accesses to Yahoo! each day means that the company can make huge profits purely from online advertising, and it therefore sees no need to start charging users for its services. Presumably were it to do so then the number of people visiting the site would drop dramatically and the company could no longer justify charging such large sums for advertising. It is much easier for an advertiser to see how successful its online advertisements are being compared with a hard copy advertisement in a magazine. While a Web site may charge an advertiser (say) 1 cent for each time an advertisement is *eyeballed* by an Internet user (ie. the user has merely seen the advertisement on the page), it might charge 10 cents for each time the advertiser's own Web site is accessed by an Internet user clicking on the advertisement banner.

Despite this, many search engines have started charging for some or all of their services. *Infoseek*[30] offers a free service available via the World Wide Web which contains a limited database of Web documents, newsgroups, FTP and gopher sites, and a keyword search interface allowing up to a hundred results to be retrieved. Its main service is fee-based, however, with a package that includes many additional databases. Usenet News, Cineman Movie, book and music reviews, news wire services, Hoover's Company Profiles, CorpTech Directory of Technology Companies, and MDX Health Directory are available at the time of writing, and new databases on a wide range of topics are added regularly. In many cases, the full text of articles stored in these databases can be retrieved. Membership costs can vary from (for example) 20 cents per search to a monthly subscription rate of $9.95. *NlightN*[31] and *IBM infoMarket*[32] are two more fee-based Web search services. NlightN has its own *Universal Index* which contains access points to Web pages, news wires, reference sources and, in its own words, "hundreds of public domain and proprietary databases."[33]. IBM infoMarket offers a way of searching several commercial databases across the Internet simultaneously. Neither offers quite as comprehensive a service to non-subscribers as Infoseek however. In both cases the free service is more of a "taster" to entice new subscribers, while Infoseek's free service is a useful Web resource in its own right.

2.2.1.6 Quality of Resource Description

Another important difference between the many different search engines available on the Web is the amount of information available on the various documents and resources in their databases. Currently, two particular methods of providing resource detail seem to be popular: the review, and the record.

- *Review*. Several search engines have emerged in recent months which employ a degree of editorial control over the documents appearing in their databases to the extent that a review, and in some cases a rating, is supplied for each resource. Two of the most popular of these are *Magellan*[34], and *Excite*[35]. Both of these services have an extensive database of Web pages over which no editorial control has been exercised, and a more limited database of documents which have been reviewed. Although the two services have now been merged into the same company[36] the search engines remain independent as a Web presence.

Magellan provides the more comprehensive information for each document, even providing details for further reference, while Excite tends to provide a shorter, more punchy review. Some sources, such as *Online* magazine, believe this type of search engine to be, "spearheading the next wave of WWW search services."[37].

• *Record.* Another method of describing a resource indexed in a Web database is to use the traditional library concept of a catalogue record. Recently, various initiatives have been made in this direction. In the US, the Online Computer Library Center (OCLC) launched a government Department of Education-funded project, "Building a Catalog of Internet-Accessible Materials,"[38] which called upon a voluntary cooperative of academic libraries nationwide to each assemble their own catalogue of Internet resources, "in accordance with local interests, collection policies, and constituents' needs," using USMARC format bibliographic records to describe each resource[39].

OCLC have also created their own search engine, *NetFirst*[40], which uses the review technique for describing a resource by providing a 50-80 word abstract while providing Library of Congress Subject Headings (LCSH) and a Dewey Decimal Classification number. At the time of writing, NetFirst contains around 55000 records of, "resources of interest to libraries and their users," available from the Web, mailing lists, newsgroups and fee-based services.

Meanwhile, in the UK, the British Library Research and Development Division (BLRDD) funded a project called Cataloguing and Retrieval of Information Over Networks Applications (CATRIONA). The aim of CATRIONA is to look at, "the development of applications programs and procedures to enable the cataloguing, classification and retrieval of documents and other resources over networks"[41]. Although not specifically designed for information retrieval on the World Wide Web, it is easy to see its relevance to the subject, and the results of the project will be of interest to Web database developers. By May 1996, the practical results of CATRIONA had made an appearance on the Web in the form of LINK - Libraries of Networked Knowledge[42] - a keyword search engine based on the BUBL Subject Tree.

Finally, it is worth mentioning HTML's own concession to creating a Web page catalogue record, the *meta tag*. The meta tag is a pointer in the HTML code of a

Web document which helps automatically-assembled Web databases to index the resource. A Web robot indexing documents will then look at any information contained in the meta tag and place it into various fields in its database, such as, "keywords," or, "description." How any information stored in a meta tag is indexed will thus depend on how the Web robot has been programmed[43].

2.2.1.7 Other Factors Affecting the Choice of Web Database

2.2.1.7.1 Retrieval Protocol.

It is worth considering the type of Internet resource one wishes to search. Which part of the Internet does the search engine catalogue? Both Veronica for Gopher documents and Archie for FTP documents have been mentioned. Most Web databases contain http (World Wide Web) addresses, but some index only Gopher, FTP sites or newsgroup postings, while others index most areas of the Internet. An easily forgotten source of information on the Internet is other users, and very often the answer to a specific question may be just an e-mail message away. One must not forget, therefore, search engines providing the e-mail addresses of experts. In addition to these particular Internet protocols where documents and other sources of information may be stored for immediate retrieval, it is also worth remembering that the Internet may also be used for locating alternative resources, such as hard copy. Most academic library catalogues in the world, for example, have gone online and are currently available via telnet. A search engine which catalogues telnet sites, therefore, can be a useful resource since although it may not lead directly to a particular piece of information, it may provide a pointer which can in turn be followed up. In summary, any search engine with a Web interface may index any combination of the following:

- Web documents
- FTP sites
- Gopher documents
- Newsgroup postings
- E-mail addresses
- Telnet sites.

2.2.1.7.2 Retrieval Medium

The type of medium one wishes to retrieve can also be a factor. Mechanisms for text retrieval are advanced by comparison with those for sound, video and images, but many Web developers and information professionals are working on possible methods for locating these latter three. The *Yahoo! Image Surfer*[44] is an example of such a tool. Subject categories can be selected for the particular image one wishes to locate and a browsing-orientated procedure is then followed until a suitable picture is found. As with most browsing on a resource the size of the Web, a search can be time-consuming, but the *Yahoo! Image Surfer* is certainly a step in the right direction, and much more effective for image retrieval than standard search engines, which tend not to take account of images at all.

In summary then, it has been made abundantly clear over the last year or so that the World Wide Web is a multimedia resource and we should therefore not merely confine ourselves to textual information. It should be borne in mind that end-users may be searching for any of the following media:

- text
- images
- sound
- video
- software.

Although it will not be discussed here, it is conceivable that very soon people will also be using the Web to retrieve Java applets from the growing number of public archives available.

The following section outlines some examples of the current generation of search engines based on the above descriptions. There are many more diverse Web search facilities not yet covered and hard to categorise, and some of these will also be explained.

2.2.2 A Classification of Web Search Engines and Databases

A search tool's functionality can depend on a number of factors affecting the user interface, how its database is put together, the methods used for information retrieval, and many other aspects not easily classified. Some examples are as follows:

- Whether its database is manually or automatically assembled

- Does the database review sites in its index?

- Whether information retrieval is by subject tree or by keyword search

- Does the search tool have its own database, or is it mainly an aid to retrieving information from existing databases such as a multi-engine search page containing search forms for various Web search engines, or a "search engine of search engines"?

- Is the tool a Simultaneous Unified Search Index (SUSI) which simultaneously applies an identical search to a number of search engines, collating their results into a single response?

- Is the database selective, specifying its resources, for instance, by subject, geographical location or Internet domain?

Most of these examples have already been described in 2.2.1, but other factors can also affect a search tool's functionality. *Firefly*[45], for example, a facility designed for locating Web sites relevant to a given individual's recreational tastes, is supposedly "intelligent" in that it attempts to build a user profile of each single searcher based on previous queries. Some of these more unusual search engines will also be touched upon later.

Before we can go on to describe a number of Web search facilities with a view to assessing which one(s) to use depending on the information required, we must classify them based on the distinctions made in the previous section.

It is proposed, therefore, that a given search engine is defined by one or more of the following elements:

(i) Assembly Method
    Manually-assembled

Automatically-assembled

(ii) Document Retrieval Method
    Keyword search engine
    Classified index

(iii) Hierarchical Arrangement
    Meta-search engines / Simultaneous Unified Search Indices (SUSIs)
    Multi-engine search pages
    Search engines of search engines

(iv) Vertical Search Engines
    Subject-specific
    Geographically-specific
    Media-specific
    Linguistically-specific

(v) Price
    Free databases
    Subscription databases with limited free service
    Subscription only databases

(vi) Quality of Resource Description
    Reviews
    Records

Additionally, when assessing the suitability of each search engine, its capacity to index the following sources must also be considered:

(i) Retrieval Protocol
    World Wide Web
    File Transfer Protocol (FTP)
    Gopher
    Newsgroups
    E-mail addresses
    Telnet

(ii) Retrieval Medium
- Text
- Images
- Sound
- Video
- Software

2.2.3 Relevance Ranking

While most results retrieved from more traditional online database services tend to list results retrieved in chronological order, or alphabetically by author or title, virtually all Web databases use mathematical algorithms to *relevance rank* any documents retrieved based on such criteria as the number of times a requested keyword appears in the document, how many times the keyword appears in the first paragraph of a document (since the first paragraph is very often a summary of the whole document), or whether it appears in the document's title.

In the absence of accurate artificial intelligence, one would expect relevance ranking to be a poor method of assessing the quality of an article from the point of view of an end user, but many of the large automatically-assembled database services have performed well in tests in this respect. A study carried out in the autumn of 1995 by reference librarians at the Elihu Burritt Library, Central Connecticut State University, measured five Web databases - Alta Vista, InfoSeek, Lycos, Magellan, and Point - for the relevance of the first ten documents retrieved by each based on 200 different subject requests. Alta Vista came up ahead with an average of 9.3 relevant documents in their top ten *hits*, and only Point, with 2.1, showed a poor degree of accuracy. However, it is worth noting that Point's database, which aims to cover the, "Top 5% of all World Wide Web sites," is relatively small and geared towards the hobbyists and casual users - their aim is, "to point out the good stuff, save you time, and help you to achieve 100% pure surfing pleasure"[46] - rather than the more technical users looking for information in an academic reference library[47].

The practice of relevance ranking has led to additional problems however as Web site maintainers have striven to have their pages appear higher up the list of hits from major search engines. One of the more underhand methods used to achieve this end is the process of *spamming*, the repetitive use of particular keywords in a document, purely as a means of making the document *appear* more relevant to a search engine. With as many as nine relevant hits in the first ten documents retrieved, an end-user searching for information on the Web may look no further. Appearing eleventh on the list, therefore, may confine a document to the same sort of oblivion as if it did not appear at all. For this reason, few search engines give the user any indication as to how their particular relevance ranking algorithm works, and this has led some services, such as *The Webmaster's Guide to Search Engines and Directories*[48], to speculate on ways of getting your Web site further up a search engine's *hit list.* They concluded that both putting a keyword in the <*TITLE*> HTML tag of a document, and putting the keyword closer to the front of a title (for example, by calling a document, *Loughborough Pubs: a Web User's Guide*, rather than, *A Web User's Guide to Pubs in Loughborough*) helped with most search engines, as did repeating keywords several times in the main body of text. They also advised that although spamming worked with certain search engines, most notably *Excite*, the practice could snowball until some documents featured nothing but repeated keywords, and this would annoy Internet users noted for their democratic self-regulation[49].

Finally, a more commercial side effect of the practice of relevance ranking has emerged with Open Text's announcement that it is to start selling *preferred status* to sites in its index in order to place them higher up results lists from searches. This has opened up a whole new debate on freedom of information and search engine credibility, whose issues are too many to enter into here[50].

2.2.4 Database Size

Before moving on to an evaluation of some of the current generation of Web databases, it is worth pointing out a discrepancy often encountered when comparing the size in terms of the number of documents indexed by a database. *Inktomi*, at the University of California at Berkeley, and *Excite*, both suggest that there are three methods a Web database may use to describe its size[51,52].

- The exact number of actual documents retrieved and subsequently indexed by the database,
- The number of unique URLs accessible from documents indexed by the database. This means that if the database finds a link pointing to another document in a URL that it has indexed, then the URL of that link will also be

counted, even if the link itself has not been indexed by the database,

- The number of non-distinct URLs in the database. The least accurate way of measuring the size of a database, this means that if a link to a document appears a hundred times then it will be counted as a hundred documents.

Throughout the following evaluation, unless stated otherwise, it can be assumed that the first of the above methods has been used to calculate the size of a database.

REFERENCES FOR CHAPTER TWO

1. **De Bra, P.M. & R.D.J. Post**, 1994. *Information retrieval in the World Wide Web.* In: *Computer Networks and ISDN Systems*, **27**, pp.183-192.
2. *Fish-Search form used on www.win.tue.nl.* URL: http://www.win.tue.nl/bin/fish-search/. 2 Sep 1996.
3. **Kent, Robert E., & Christian Neuss**, 1995. *Creating a Web analysis and visualisation environment.* In: *Computer Networks and ISDN Systems,* **28**, pp.109-117.
4. *The Internet Services List.* URL: http://www.spectracom.com/islist/. 2 Sep 1996.
5. *Yahoo!* URL: http://www.yahoo.com/. 2 Sep 1996.
6. **Poulter, Alan.** *Web search engines: a critical review.* In: *Program* (in press).
7. *Yahoo!* URL: http://www.yahoo.com/. 2 Sep 1996.
8. *Welcome to Lycos.* URL: http://www.lycos.com/. 2 Sep 1996.
9. *AltaVista: main page.* URL: http://www.altavista.digital.com/. 2 Sep 1996.
10. *A2Z home page.* URL: http://a2z.lycos.com/. 2 Sep 1996.
11. *CUI W3 Catalog.* URL: http://cuiwww.unige.ch/w3catalog/. 2 Sep 1996.
12. *Welcome to Open Text Corporation.* URL: http://www.opentext.com/. 2 Sep 1996.
13. *BUBL Information Service Web Server.* URL: http://www.bubl.bath.ac.uk/BUBL/cattree.html. 2 Sep 1996.
14. *The World-Wide Web Virtual Library: subject catalogue.* URL: http://www.w3.org/pub/ DataSources/bySubject/Overview.html. 2 Sep 1996.

15. *Social Science Information Gateway - SOSIG.* URL: http://sosig.ac.uk/. 2 Sep 1996.

16. *All-in-One Search Page.* URL: http://www.albany.net/allinone/. 2 Sep 1996.

17. *MetaCrawler searching.* URL: http://metacrawler.cs.washington.edu/. 2 Sep 1996.

18. *SEARCH.COM.* URL: http://www.search.com/. 2 Sep 1996.

19. *Search the Internet with the Internet Sleuth.* URL: http://www.isleuth.com/. 2 Sep 1996.

20. *Access to network resources projects.* URL: http://ukoln.bath.ac.uk/elib/lists/anr.html. 2 Sep 1996.

21. *The HUMBUL Gateway.* URL: http://sable.ox.ac.uk/departments/humanities/international. html. 2 Sep 1996.

22. *The ADAM project.* URL: http://adam.ac.uk/. 2 Sep 1996.

23. *OMNI welcome page.* URL: http://omni.ac.uk/. 2 Sep 1996.

24. *Amateur hardcore search engine.* URL: http://www.amateurs.com/searchex.htm. 2 Sep 1996.

25. *AC/DC: The ACademic DireCtory.* URL: http://acdc.hensa.ac.uk/. 2 Sep 1996.

26. *SavvySearch.* URL: http://www.cs.colostate.edu/~dreiling/smartform.html. 2 Sep 1996.

27. *Flipper home page.* URL: http://flp.cs.tu-berlin.de/flipper/. 2 Sep 1996.

28. *Recursos de Internet en Español y Portugués·* URL: http://www.ogilvy.com/spanish/hisplink.htm. 2 Sep 1996.

29. *QBIC home page.* URL: http:// wwwqbic.almaden.ibm.com/~qbic/qbic.html. 2 Sep 1996.

30. *Infoseek Guide.* URL: http://www.infoseek.com/. 2 Sep 1996.

31. *NlightN home page!* URL: http://www.nlightn.com/. 2 Sep 1996.

32. *infoMarket search page.* URL: http://www.infomkt.ibm.com/. 2 Sep 1996.

33. *Help/Frequently Asked Questions.* URL: http://www.nlightn.com/help/help.htm. 2 Sep 1996.

34. *Welcome to Magellan!* URL: http://www.mckinley.com/. 2 Sep 1996.

35. *Excite home.* URL: http://www.excite.com/. 2 Sep 1996.

36. *Excite and the McKinley Group Sign Letter of Intent to Merge.* In: *McKinley Press Information.* URL: http://www.mckinley.com/feature.cgi?pressroom2_bd. 2 Sep 1996.

37. **Courtois, Martin P.** *Cool tools for Web searching: an update.* In: *Online*, **20** (3), May/June 1996, p. 30.

38. *Building a Catalog of Internet-Accessible Materials.* URL: http://www.oclc.org/oclc/man/catproj/overview.htm. 2 Sep 1996.

39. Ibid.

40. *NetFirst information.* URL: http://www.oclc.org/oclc/netfirst/. 2 Sep 1996.

41. *CATRIONA.* URL: http://www.bubl.bath.ac.uk/BUBL/catriona.html. 2 Sep 1996.

42. *BUBL-LINK: Libraries of Networked Knowledge.* URL: http://catriona.lib.strath.ac.uk/. 2 Sep 1996.

43. For a good example of a meta tag functioning in practice see, *The META tag: controlling how your Web page is indexed by AltaVista.* URL: http://www.altavista.digital.com/cgi/bin/query?pg=ah&what=web#meta. 23 Aug 1996.

44. *Image Surfer category list.* URL: http://ipix.yahoo.com/. 2 Sep 1996.

45. *Firefly.* URL: http://www.ffly.com/. 2 Sep 1996.

46. *Lycos, Inc. - info.* URL: http://point.lycos.com/faq/. 2 Sep 1996.

47. **Tomaiuolo, Nicholas G. & Joan G. Packer.** *Quantitative analysis of five WWW search engines.* In: *Computers In Libraries*, **16** (6), June 1996. URL: http://neal.ctstateu.edu: 2001/htdocs/websearch.html. 2 Sep 1996.

48. *The webmaster's guide to search engines and directories.* URL: http://calafia.com/webmasters/. 2 Sep 1996.

49. *Maximized online search engine study: introduction.* URL: http://maxonline.com/searchstudy/. 2 Sep 1996.

50. *Engine sells results, draws fire.* URL: http://www.cnet.com/Content/News/Files/0,16,1635,00.html. 2 Sep 1996.

51. *Inktomi: counting documents.* URL: http://inktomi.berkeley.edu/counting.html. 18 July 1996.

52. *Counting URLs.* URL: http://www.excite.com/ice/counting.html. 2 Sep 1996.

Chapter Three - A Guide to Some Search Engines Available on the Web

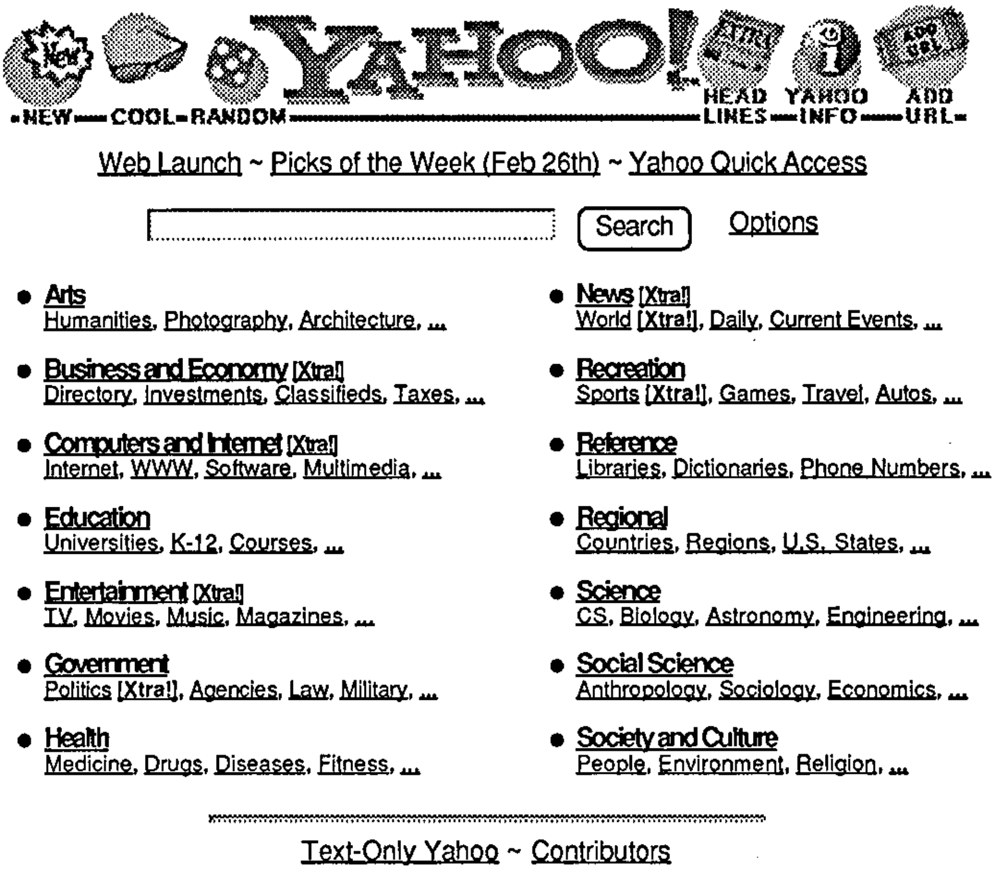
It would seem sensible, now, to look at a selection of some of the more popular and practical existing search engines, to see if some sort of strategy can be found for their use in information retrieval on the Internet. The following section details a number of Web search engines classified according to the distinctions made earlier, in section 2.2.2. Whether or not a given search engine can retrieve documents from protocols other than http, and media other than text, is also examined in each description. The aim of this exercise is to try and form an overall picture of where and when a particular search engine might be used in the process of information retrieval. It has not been possible here to provide a detailed comparison of search engines within the same class as outlined in 2.2.2 and the choice of engine for each description has been based on factors as arbitrary as personal preference and familiarity, and user popularity. At the end of each description, therefore, a brief consideration of other similarly constructed search engines is made.

3.1 A Manually-assembled Classified Index - Yahoo!

Yahoo![1], one of the most popular Web databases, was started as a hobby in April 1994 by two Ph.D. students in Electrical Engineering at Stanford University. In April 1996, Yahoo! Inc. went public, selling more than 2.5 million shares on the stock exchange at $13 apiece, and the directory's two co-founders, Jerry Yang and David Filo were, by July 1996, worth $132 million each. Despite this, Yahoo! makes virtually all of its income from on-line advertising and all of its services, in true Internet tradition, are still completely free to anyone with Internet access and a suitable Web browser[2,3,4].

3.1.1 What is Yahoo?

Yahoo! is a hierarchically-organised index of Web documents arranged logically under 14 main subject headings - Arts, Business and Economy, Computers and Internet,



Web Launch ~ Picks of the Week (Feb 26th) ~ Yahoo Quick Access

- **Arts**
  Humanities, Photography, Architecture, ...
- **Business and Economy** [Xtra!]
  Directory, Investments, Classifieds, Taxes, ...
- **Computers and Internet** [Xtra!]
  Internet, WWW, Software, Multimedia, ...
- **Education**
  Universities, K-12, Courses, ...
- **Entertainment** [Xtra!]
  TV, Movies, Music, Magazines, ...
- **Government**
  Politics [Xtra!], Agencies, Law, Military, ...
- **Health**
  Medicine, Drugs, Diseases, Fitness, ...
- **News** [Xtra!]
  World [Xtra!], Daily, Current Events, ...
- **Recreation**
  Sports [Xtra!], Games, Travel, Autos, ...
- **Reference**
  Libraries, Dictionaries, Phone Numbers, ...
- **Regional**
  Countries, Regions, U.S. States, ...
- **Science**
  CS, Biology, Astronomy, Engineering, ...
- **Social Science**
  Anthropology, Sociology, Economics, ...
- **Society and Culture**
  People, Environment, Religion, ...

Text-Only Yahoo ~ Contributors

Diagram 3.1 - Yahoo! subject headings

Education, Entertainment, Government, Health, News, Recreation, Reference, Regional, Science, Social Science, and Society and Culture - and over 350 sub-headings. Due to its position as *the* most popular Web database[5], most URLs in its index have been acquired by submission from Internet users, using the "Add URL," option on Yahoo's home page.

Because of the transitional nature of the Internet at present, it is difficult to estimate precisely how many Web documents Yahoo! indexes, and any figures will quickly become out of date. In the Autumn of 1995, *Online* magazine reported Yahoo! to have indexed about 66000 Web documents[6]. Yahoo! figured poorly in sample keyword query tests carried out for the above article, and the authors suggested that, "this poor performance is due primarily to Yahoo's indexing, which for most entries includes only the title and URL." According to this author's estimates, however, by July 1996, Yahoo! was indexing nearly 300000 documents and being added to weekly, and in a follow-up article, *Online* this time reported an improvement in Yahoo's search facility, provided firstly by the fact that it now indexes title, URL and "Comments", a short one or two line verbal description of the indexed document provided by Yahoo staff, and secondly by its collaboration with the Open Text automatically-assembled Web database (see later). *Online* reported that, "a search in Yahoo! produces a list of hits from the Yahoo! database, along with a link to sites provided by Open Text and greatly expands the number of sites retrieved" (Open Text claims its own database indexes over 19 million URLs, although it uses the number of non-distinct URLs within the database to calculate this figure)[7].

3.1.2 Features

The principal method of access to documents indexed in Yahoo!, however, is not by keyword searching, but by browsing. The Yahoo! Search page[8] contains its own search form, which allows both boolean AND and OR searching, and allows users to search for both complete words or for substrings within a word, but the tree-structure of headings and sub-headings is logically arranged and allows for easy browsing. Locating a given document does not normally prove difficult using this method, and a big advantage is that any document retrieved is likely to be relevant. There are also plenty of cross-references. The "@" symbol at the end of a sub-heading indicates that the sub-heading is available at different places within Yahoo! Clicking on the sub-heading then takes the searcher to the its original location within Yahoo!

Some areas where Yahoo! can be a particularly good source of information on the Web include the following:

- Children's pages. The Yahooligans! Web Guide for Kids features 8 subject categories - Around the World; School Bell; Art Soup; Science and Oddities; Computers, Games and Online; Sports and Recreation; Entertainment; and The Scoop - and over 100 sub-categories.

- Company Web sites. The http://www.yahoo.com/Business_and_Economy/Companies/ subdirectory contains the addresses of over 100000 company Web sites further subdivided into hundreds of categories and subcategories.

3.1.3 Other Internet Protocols

FTP Sites

Although Yahoo! does not have its own directory of Internet FTP sites, the following subdirectories contain links to other FTP search services:

- http://www.yahoo.com/Computers_and_Internet/Internet/Searching_the_Net/Archie/
- http://www.yahoo.com/Computers_and_Internet/Internet/FTP_Sites/

Gopher

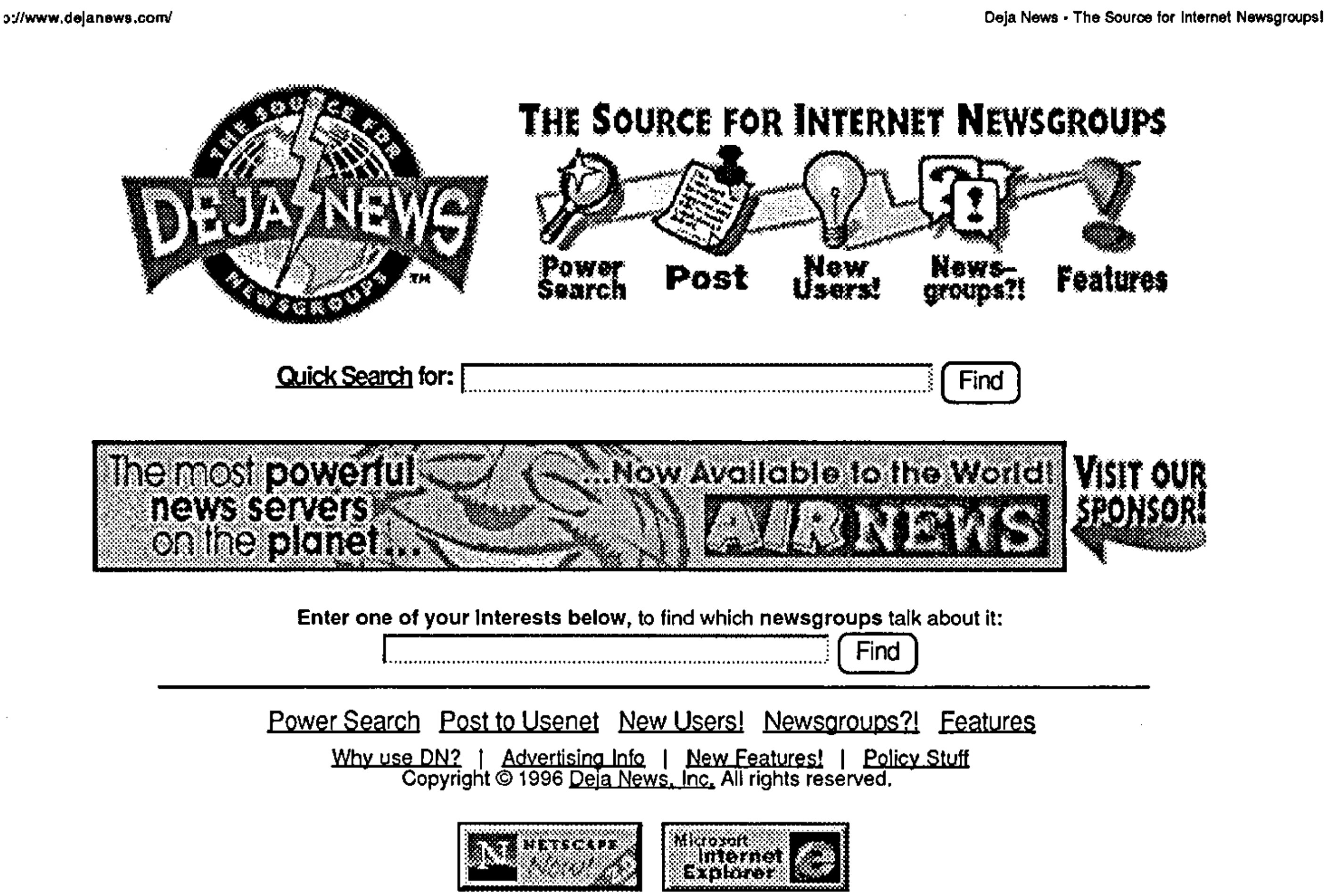
Again Yahoo! lacks its own directory of Gopher resources but contains links to other search services covering this area of the Internet. These links are contained within the following subdirectories:

- http://www.yahoo.com/Computers_and_Internet/Internet/Gopher/
- http://www.yahoo.com/Computers_and_Internet/Internet/Searching_the_Net/Jughead/
- http://www.yahoo.com/Computers_and_Internet/Internet/Searching_the_Net/Veronica/

Newsgroups

The Yahoo! search page allows one to search the *Deja News* directory[9] by keyword. In addition, Yahoo! contains links to other newsgroup search services within the following subdirectory:

- http://www.yahoo.com/News/Usenet/Searching_and_Filtering/



Enter one of your interests below, to find which newsgroups talk about it:

Copyright © 1996 Deja News, Inc. All rights reserved.

Diagram 3.2 - The DejaNews newsgroup search engine

E-mail addresses

Yahoo! allows two methods of searching for individual e-mail addresses.

- Firstly, by entering the Yahoo! search page, the searcher is given the option of searching either Yahoo!, Usenet, or e-mail addresses. A search using the e-mail address option will look up a keyword on the *Four11 Directory*[10]. Such a search is not as detailed as searching the Four11 Directory directly. Their own search form allows one to enter details for surname, christian name, city, domain, state and country.

- Secondly, Yahoo! provides its own directory of e-mail addresses, *Yahoo! People Search*[11]. This particular service indexes people in two separate directories: address/telephone number, and e-mail address. Each search form is slightly different. The address/telephone number directory allows the searcher to enter details for surname, christian name, city, state, and telephone number, while the e-mail address search form contains details for surname, christian name, and domain. Additionally, the option is given to search for a person by World Wide Web home page. This links to the Entertainment/People subdirectory of Yahoo![12] and allows a simple keyword search, as do all Yahoo! subdirectories, or a browsing menu by initial. Somewhat illogically, however, this initial does not reference their surname but their full name, and may therefore reference alphabetically the letters of their first name, or simply the initials of all of their christian names!

Telnet

Yahoo! does not appear to be quite so good for locating the addresses of telnet sites, although tucked away in the following directory is a link to one of the more important telnet directories, *Hytelnet*[13] as well as one or two other links:

- http://www.yahoo.com/Computers_and_Internet/Software/Communication/Telnet/

3.1.4 Other Media

Images

The *Yahoo! Image Surfer*[14] is currently one of the more comprehensive, if eclectic, tools available for retrieving images from the Web. At present, the service offers over 50 subject categories of image, although this figure is climbing rapidly and it would not be surprising if they were to be further subdivided into the standard Yahoo! subdirectory categories before very long. Clicking on one of these categories leads into the browser which displays thumbnail pictures of images ten at a time. Given that each category can contain hundreds of images, this browsing-centred procedure can prove very slow at retrieving the desired image. Indeed, the process is quite arbitrary and more specific image requests, for example, for a picture of Coventry Cathedral rather than of just *a* cathedral would still best be carried out by an ordinary text search for Coventry Cathedral in the hope that any page retrieved might contain a picture of it.

In addition to the Yahoo! Image Surfer, links to other image archives can be found in the following Yahoo! subdirectory:

- http://www.yahoo.com/Computers_and_Internet/Multimedia/Pictures/

Sound

Yahoo! does not have its own sound archive at the time of writing, but contains links to other sound archives in the following subdirectory:

- http://www.yahoo.com/Computers_and_Internet/Multimedia/Sound/

Video

Neither does Yahoo! have its own video archive but again links to other archives can be found in the subdirectory:

- http://www.yahoo.com/Computers_and_Internet/Multimedia/Video/

Software

Yahoo! has a large directory, currently containing over 1500 items, both of information relating to various software packages, and of directories of downloadable software. This can be found at the following address:

- http://www.yahoo.com/Computers_and_Internet/Software/

3.1.5 Other Manually-assembled databases and classified indices

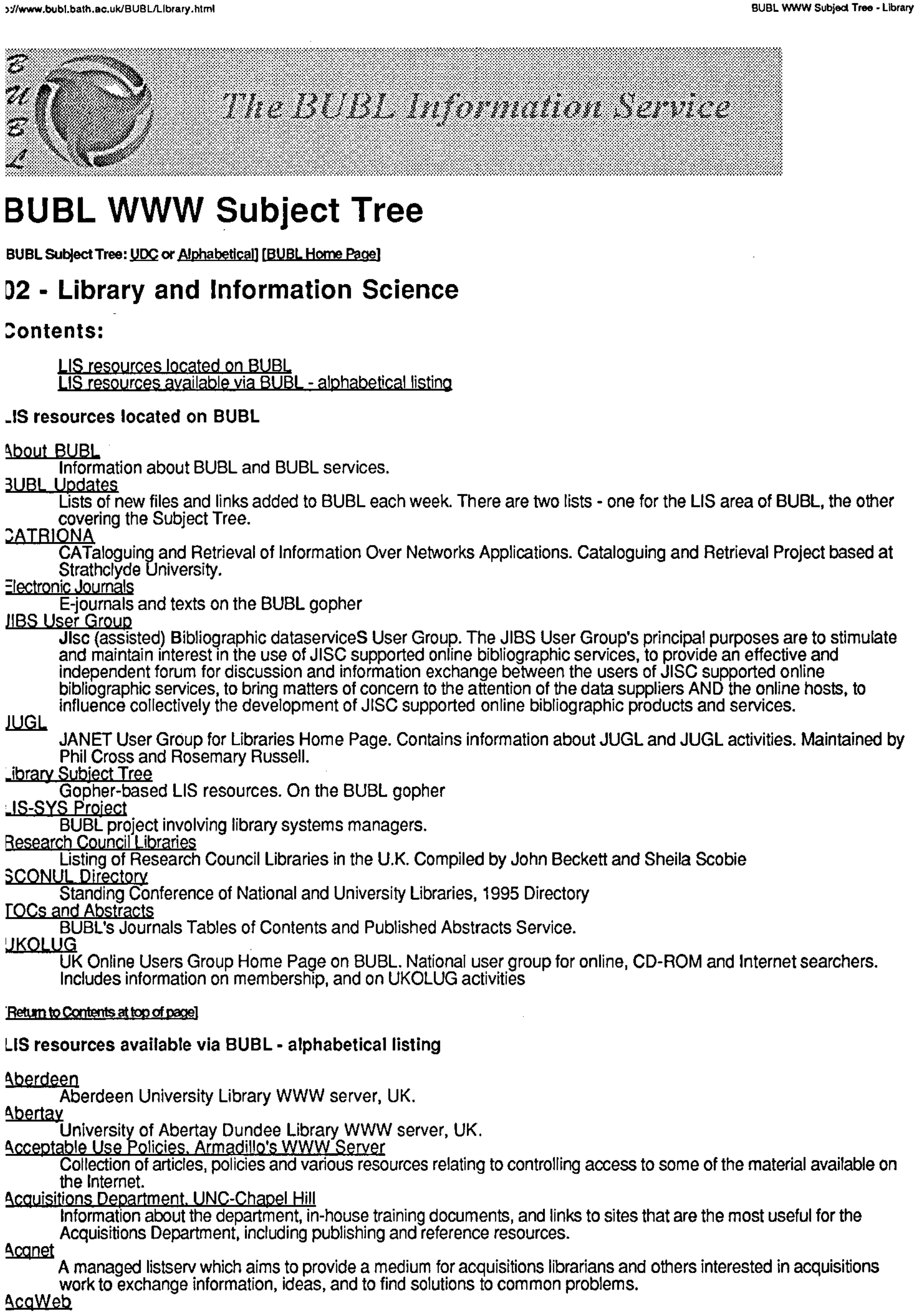
All classified indices on the Web are by their nature manually-assembled by the simple expedient that at present there is no easy way for a Web robot to classify a document into a specific subject category while it is being indexed. This situation looks set to remain unless and until standardisation of subject classification is arrived at the length and breadth of the Web.

Yahoo! uses its own unique subject classification, as do many other classified Web indices, and the logic of each scheme can vary considerably. It is worth looking at an index which has made concessions to standards. One example is the *Bulletin Board for Libraries (BUBL) Information Service*[15]. This has a number of features including links to other Web search tools, network training resources, news, and job vacancies, but the important service for the user looking for information on the Internet is the BUBL Subject Tree, a directory of documents which can be arranged either as an alphabetical list or in Universal Decimal Classification (UDC) order. The long alphabetical list is not as easy to use as Yahoo's directory and subdirectory structure and often requires a glance down the entire list before an appropriate subject can be found, and users familiar with UDC will obviously prefer to use this.

The emphasis of the BUBL Subject Tree is on documents of interest to the UK higher education community and the general interest Web searcher would be advised to look at Yahoo!, whose directory is much larger and covers a wider range of documents. BUBL definitely has its advantages though, mainly to UK users because of its emphasis on UK documents. Due to its origins it is also particularly strong in the areas of libraries and information science.

It has its own very basic keyword search facility which allows substring searching and the boolean operators AND and OR. This facility looks set to improve, however, as the index is being moved onto a new service based at the University of Strathclyde called LINK[16]. LINK is the practical result of the CATRIONA project, mentioned in the preceding chapter. It is a combined World Wide Web and Z39.50 service and is much more sophisticated than the BUBL Subject Tree, featuring records for the document indexed and a more advanced search facility.

Other popular classified indices include the *World Wide Web Virtual Library*[17], and the *Argus Clearinghouse*[18]. The latter describes itself as, "a central access point for value-added topical guides which identify, describe, and evaluate Internet-based information resources," and features guides to Internet resources on a given subject, submitted to Argus by independent users. Each guide is reviewed by Argus staff and given a rating of from one to five. The WWW Virtual Library works on a similar principle, in that links to Internet resources on a given subject area are provided by a voluntary independent body, but each independent body is assigned by the WWW Virtual Library project team, part of the World Wide Web Consortium, on the basis of being known



://www.bubl.bath.ac.uk/BUBL/Library.html

BUBL WWW Subject Tree - Library

The BUBL Information Service

BUBL WWW Subject Tree

BUBL Subject Tree: UDC or Alphabetical] [BUBL Home Page]

02 - Library and Information Science

Contents:

LIS resources located on BUBL
LIS resources available via BUBL - alphabetical listing

LIS resources located on BUBL

About BUBL
    Information about BUBL and BUBL services.
BUBL Updates
    Lists of new files and links added to BUBL each week. There are two lists - one for the LIS area of BUBL, the other covering the Subject Tree.
CATRIONA
    CATaloguing and Retrieval of Information Over Networks Applications. Cataloguing and Retrieval Project based at Strathclyde University.
Electronic Journals
    E-journals and texts on the BUBL gopher
JIBS User Group
    JIsc (assisted) Bibliographic dataserviceS User Group. The JIBS User Group's principal purposes are to stimulate and maintain interest in the use of JISC supported online bibliographic services, to provide an effective and independent forum for discussion and information exchange between the users of JISC supported online bibliographic services, to bring matters of concern to the attention of the data suppliers AND the online hosts, to influence collectively the development of JISC supported online bibliographic products and services.
JUGL
    JANET User Group for Libraries Home Page. Contains information about JUGL and JUGL activities. Maintained by Phil Cross and Rosemary Russell.
Library Subject Tree
    Gopher-based LIS resources. On the BUBL gopher
LIS-SYS Project
    BUBL project involving library systems managers.
Research Council Libraries
    Listing of Research Council Libraries in the U.K. Compiled by John Beckett and Sheila Scobie
SCONUL Directory
    Standing Conference of National and University Libraries, 1995 Directory
TOCs and Abstracts
    BUBL's Journals Tables of Contents and Published Abstracts Service.
UKOLUG
    UK Online Users Group Home Page on BUBL. National user group for online, CD-ROM and Internet searchers. Includes information on membership, and on UKOLUG activities

[Return to Contents at top of page]

LIS resources available via BUBL - alphabetical listing

Aberdeen
    Aberdeen University Library WWW server, UK.
Abertay
    University of Abertay Dundee Library WWW server, UK.
Acceptable Use Policies, Armadillo's WWW Server
    Collection of articles, policies and various resources relating to controlling access to some of the material available on the Internet.
Acquisitions Department, UNC-Chapel Hill
    Information about the department, in-house training documents, and links to sites that are the most useful for the Acquisitions Department, including publishing and reference resources.
Acqnet
    A managed listserv which aims to provide a medium for acquisitions librarians and others interested in acquisitions work to exchange information, ideas, and to find solutions to common problems.
AcqWeb

Diagram 3.3 - The BUBL Subject Tree in UDC classification

experts in that field. Again, these services are much more academically-inclined than Yahoo!

3.2 An Automatically-assembled Keyword Search Engine - AltaVista

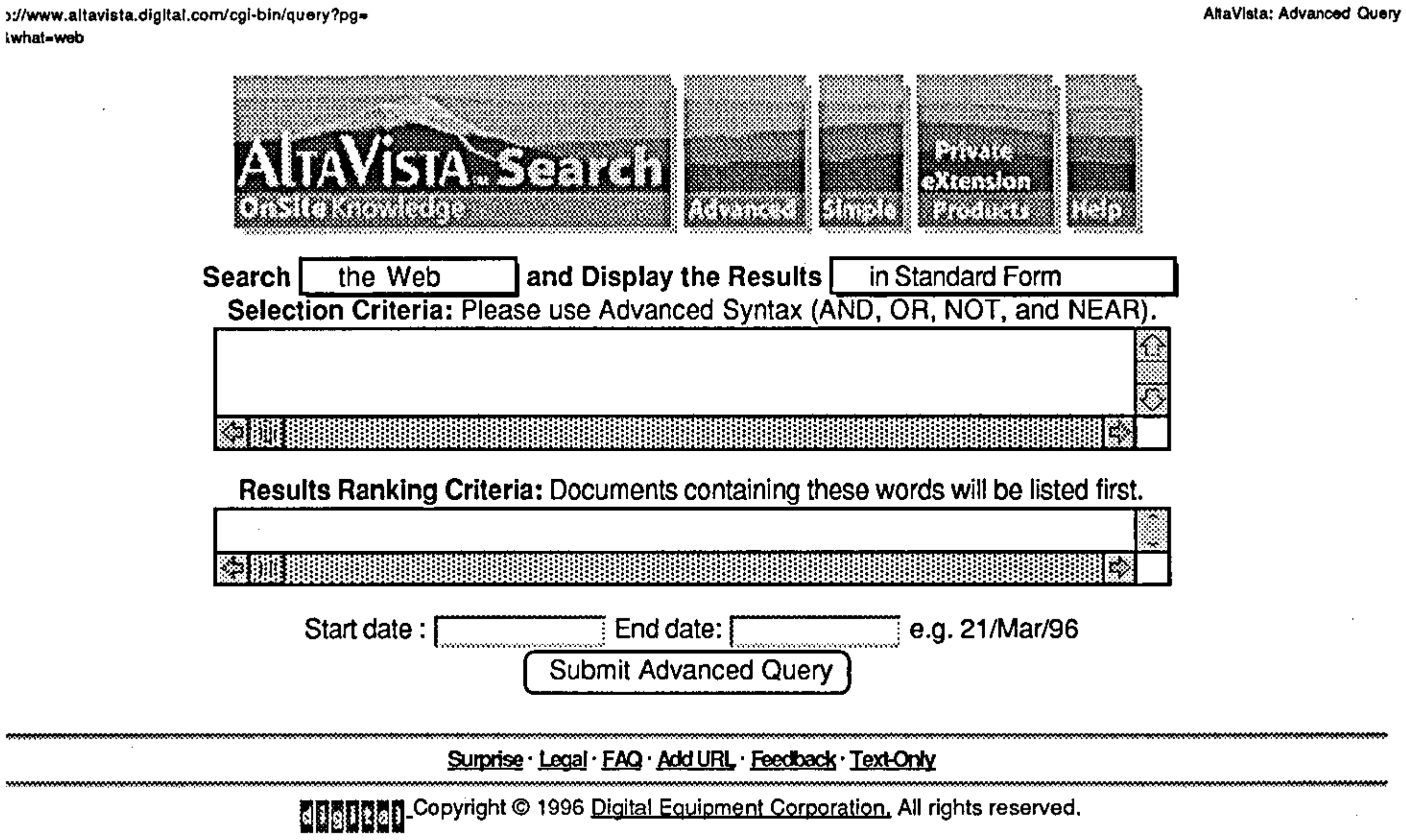
AltaVista[19] became publicly available on the Web in December 1995 as a result of a research project at Digital's Research Laboratories at Palo Alto, California. At that stage its database contained some 16 million documents, making it comfortably the largest search engine on the Web. By May 1996, it had indexed around 30 million documents and was still way ahead of the field in terms of database size. By August 1996 it was being caught up by other services, notably Excite, but by that stage it had gained a considerable number of new users over the first half of the year and was easily one of the most popular search engines on the Web.

3.2.1 Features

Besides its sheer size, one of the strongest features of AltaVista is its advanced keyword search language.

At the basic level, AltaVista has a *simple query* search which has the following features:

- a search for multiple keywords defaults to the OR boolean operator
- a plus sign ('+') before a keyword means that particular word *must* be present in any results obtained
- a minus sign ('-') before a keyword acts as a NOT boolean operator for that word
- phrases can be searched for by enclosing the words of a phrase in double quotation marks
- searches can be limited to particular fields of a Web page by using a field operator, '*field*:'. Fields available for this feature include *applet*, *host*, *image*, *link*, *text*, *title*, and *url* for Web documents, and *from*, *subject*, *newsgroups*, and *summary* for newsgroup articles
- truncation is not automatic but an asterisk can be used for a substring search
- search terms entered in wholly lower case result in a case-insensitive search



://www.altavista.digital.com/cgi-bin/query?pg=
what=web

AltaVista: Advanced Query

AltaVista Search
OnSite Knowledge
Advanced
Simple
Private eXtension Products
Help

Search the Web **and Display the Results** in Standard Form

Selection Criteria: Please use Advanced Syntax (AND, OR, NOT, and NEAR).

Results Ranking Criteria: Documents containing these words will be listed first.

Start date : ☐ End date: ☐ e.g. 21/Mar/96

Submit Advanced Query
Surprise · Legal · FAQ · Add URL · Feedback · Text-Only

digital Copyright © 1996 Digital Equipment Corporation. All rights reserved.

Diagram 3.4 - AltaVista's Advanced Query search page

- any results list produced is relevance ranked. Some studies have shown AltaVista's relevance ranking algorithm to be particularly accurate[20].

AltaVista also has a more comprehensive search facility called its *advanced query* search. Features here will be very familiar to those used to searching more conventional online databases and include:

- as with the simple query search, truncation by using an asterisk, and the limiting of searches by field
- nested boolean queries featuring the AND, OR and NOT operators
- the NEAR operator for finding words commonly associated with one another within ten words of each other in the text of a document
- a results ranking criteria option for selecting which keywords are more important in a search. Relevance ranking is not available in the advanced query search, and if this feature is not used then results will be listed in no particular order
- the ability to restrict a search by date

Comprehensive online documentation is available to help users with both the simple and advanced query searches.

3.2.2 Other Internet Protocols

FTP and Gopher Sites

AltaVista does not index FTP or Gopher sites.

Newsgroups

As of August 1996, AltaVista had indexed around 3 million articles from 14000 newsgroups. The searcher can specify whether he or she wishes to search the Web or Usenet. Usenet searching can also be limited by field.

E-mail Addresses

AltaVista does not specifically contain a directory of e-mail addresses, although performing a keyword search on a person's name may retrieve an article which contains an e-mail address. Given the size of AltaVista's database this could prove a fruitful option, especially since the advanced query search allows one to search for, "Joe

Bloggs", "Joe J. Bloggs", "Bloggs, Joe", or "Bloggs, Joe J." simply by searching on *Joe NEAR Bloggs.*

Telnet

AltaVista does not index Telnet sites

3.2.3 Other Media

Images

AltaVista can be used to retrieve images from a Web page if the name of the image is known or guessed correctly. This can be done by using the *image* field operator, ie. by entering, *image:"cathedral.jpeg" OR image:"cathedral.gif"*, there is a good chance of retrieving a page which features a picture of a cathedral, as long as the name of the image file is, "cathedral."

Sound and Video

Although there is no specific way of retrieving sound using AltaVista, again a little ingenuity may prove fruitful. For instance, by entering *"car.wav"* as a phrase, the user may find a Web page featuring a sound file of a car, as long as the name of the sound file is, "car." Similarly, searching on *"car.mpeg"* may find a movie file of a car.

Software

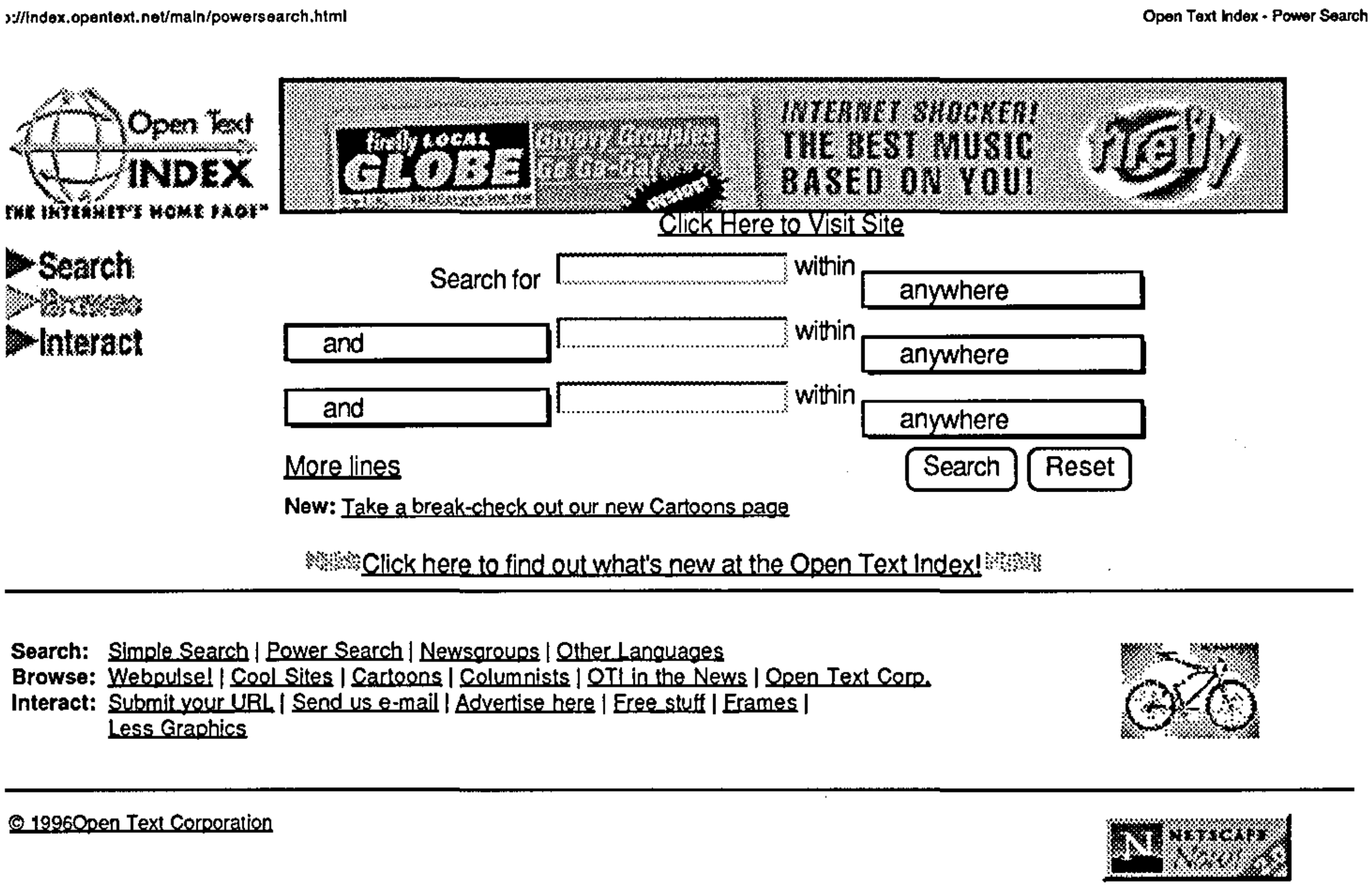
Using AltaVista can be a good way of retrieving software since usually the exact name of a piece of software is known to begin with. Performing a simple query search on the name of the software can very often lead to a direct link to a page where the software is downloadable.

3.2.4 Other Automatically-assembled databases and keyword search engines

Until very recently AltaVista was unrivalled in terms of size. Now its 30 million documents have been exceeded by *Excite*[21], which claims to index 50 million. Excite also has the advantage of providing reviews of over 60000 documents in its database. The down side, however, is that Excite has a much less sophisticated query language and the only advanced search features seem to be nested boolean logic, and the use of plus and minus signs to require and exclude words from a search respectively.

These two databases currently stand head and shoulders above other automatically-assembled databases in terms of size and it will clearly take a more detailed examination than this one to separate the two in terms of both breadth of recall and precision of results. Other databases cannot be disregarded however. By July 1996, *Lycos*[22] claimed to index 51 million documents[23] - although this figure is based on the number of unique URLs and its true size is likely to be much smaller, possibly only about one fifth of this figure[24] - and also has a smaller classified index of selected documents from its database, *Lycos A2Z*[25]. A recent study in the electronic journal *Ariadne*, conducted in February and March of 1996, suggested a number of features in Lycos' favour, such as automatic relevance ranking when using its advanced search feature, and adjacency ranking, which will give greater precedence to documents containing search keywords which are found close together within the text of a document. It also suggested, perhaps more importantly given that one of the prime advantages of electronic sources as opposed to paper ones is their currency, that the Lycos database is updated more frequently than AltaVista's. The study's final conclusion was that the user interface and search results were equally good on both search engines but that AltaVista had the best query language while Lycos was more up to date[26].

Another database which cannot be ignored is *Open Text*[27], which claims to index 19 million documents, although this figure is vastly inflated by the fact that its database size is measured in terms of the number of non-distinct URLs. Recently, Open Text has come under criticism for the fact that it is now charging Web page operators for the privilege of having their pages listed higher up on "relevance" ranked results lists. What keeps it in favour among online searchers however is the sophistication of its query language. Open Text's power search is particularly useful, allowing field searching, boolean and proximity operators, and comes with good online documentation containing plenty of example searches. A recent study of search engine query languages also pointed out that although pinpoint searching using complex nested boolean queries is possible using just one line of logic, "a searcher normally prefers to split concepts and operators into multiple search statements." Of the four systems tested - AltaVista, Infoseek, Lycos and Open Text - only Open Text allowed the user to do this[28]. Another well-established automatically-assembled Web database is *ALIWEB*[29], and the latest is *HotBot*[30], which claims to index 54 million documents.



://index.opentext.net/main/powersearch.html
Open Text Index - Power Search

Click Here to Visit Site

Search for [ ] within [anywhere]
[and] [ ] within [anywhere]
[and] [ ] within [anywhere]

More lines

New: Take a break-check out our new Cartoons page

Click here to find out what's new at the Open Text Index!

Search: Simple Search | Power Search | Newsgroups | Other Languages
Browse: Webpulse! | Cool Sites | Cartoons | Columnists | OTI in the News | Open Text Corp.
Interact: Submit your URL | Send us e-mail | Advertise here | Free stuff | Frames | Less Graphics

© 1996Open Text Corporation

Diagram 3.5 - Open Text's Power Search feature

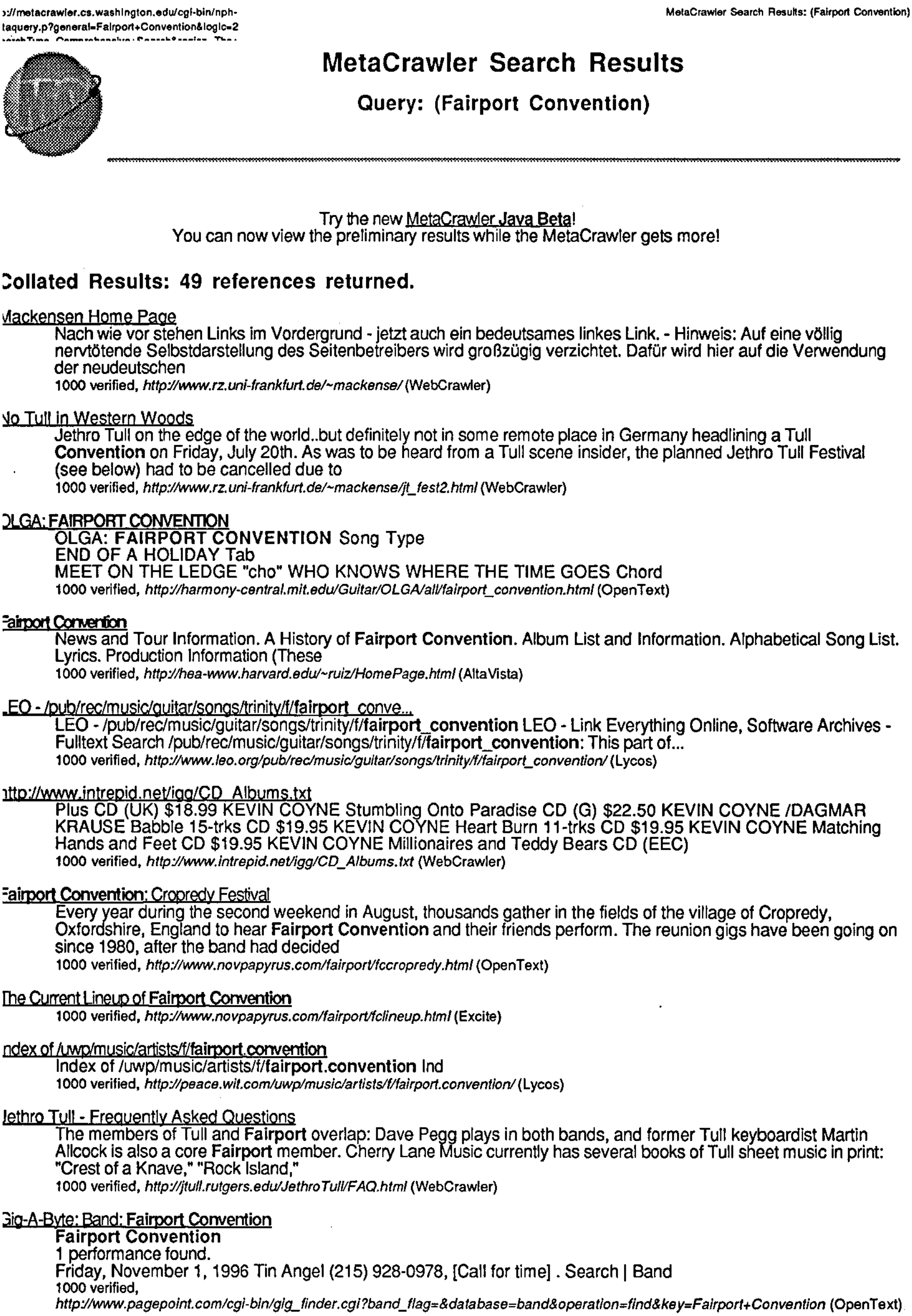
3.3 A Simultaneous Unified Search Index - MetaCrawler

MetaCrawler[31] is the result of a research project by Erik Selberg, a Ph.D. student, and Oren Etzioni, Assistant Professor of Computer Science and Engineering, at the University of Washington in Seattle. It is a Web Search service which has no database of Web documents itself but rather queries nine other Web databases - AltaVista, Excite, Galaxy, InfoSeek, Inktomi, Lycos, Open Text, WebCrawler, and Yahoo - simultaneously and collates the results into a single search page. The service is based on a number of premises.

Each Web search engine has a slightly different interface and search syntax from the next one. To query nine different search engines, a user would have to become accustomed to nine separate interfaces and, quite possibly, make nine unique search queries for an identical piece of information. This is clearly highly time-consuming, and MetaCrawler aims to avoid this by providing a single query syntax and interface which is automatically converted to the appropriate syntax for each of the nine search engines and multiple requests sent out.

Of course, this would only be necessary if each of the nine services was providing a different set of results. Research carried out between July and September 1995 showed that the best database in terms of the number of references returned for a single query and consequently followed was Lycos, which had around a 40% market share of followed references. This meant that Lycos was missing out on at least a further 60% of useful references on the Web, suggesting that it was necessary to query more than one search engine for one to assemble a comprehensive list of Web references for a given query. However, this research was carried out before AltaVista, Excite and Inktomi were added to MetaCrawler and would certainly need to be updated now that AltaVista and Excite have raced ahead of the field in terms of pushing closer towards supplying a comprehensive database of documents available on the Web.

This kind of service is certainly very useful to the general user searching for information on the Internet, but a further side effect of the research project is to enable a systematic and objective evaluation and comparison of the nine search engines. Criteria such as the number of references returned, the number of relevant references returned, the number of "dead" references (ie. links to pages which no longer exist) and speed of reply were able to be assessed[32].



://metacrawler.cs.washington.edu/cgi-bin/nph-
taquery.p?general=Fairport+Convention&logic=2

MetaCrawler Search Results: (Fairport Convention)

MetaCrawler Search Results

Query: (Fairport Convention)

Try the new MetaCrawler Java Beta!
You can now view the preliminary results while the MetaCrawler gets more!

Collated Results: 49 references returned.

Mackensen Home Page
Nach wie vor stehen Links im Vordergrund - jetzt auch ein bedeutsames linkes Link. - Hinweis: Auf eine völlig nervtötende Selbstdarstellung des Seitenbetreibers wird großzügig verzichtet. Dafür wird hier auf die Verwendung der neudeutschen
1000 verified, *http://www.rz.uni-frankfurt.de/~mackense/* (WebCrawler)

No Tull in Western Woods
Jethro Tull on the edge of the world..but definitely not in some remote place in Germany headlining a Tull **Convention** on Friday, July 20th. As was to be heard from a Tull scene insider, the planned Jethro Tull Festival (see below) had to be cancelled due to
1000 verified, *http://www.rz.uni-frankfurt.de/~mackense/jt_fest2.html* (WebCrawler)

OLGA: **FAIRPORT CONVENTION**
OLGA: **FAIRPORT CONVENTION** Song Type
END OF A HOLIDAY Tab
MEET ON THE LEDGE "cho" WHO KNOWS WHERE THE TIME GOES Chord
1000 verified, *http://harmony-central.mit.edu/Guitar/OLGA/all/fairport_convention.html* (OpenText)

Fairport Convention
News and Tour Information. A History of **Fairport Convention.** Album List and Information. Alphabetical Song List. Lyrics. Production Information (These
1000 verified, *http://hea-www.harvard.edu/~ruiz/HomePage.html* (AltaVista)

LEO - /pub/rec/music/guitar/songs/trinity/f/**fairport**_conve...
LEO - /pub/rec/music/guitar/songs/trinity/f/**fairport_convention** LEO - Link Everything Online, Software Archives - Fulltext Search /pub/rec/music/guitar/songs/trinity/f/**fairport_convention**: This part of...
1000 verified, *http://www.leo.org/pub/rec/music/guitar/songs/trinity/f/fairport_convention/* (Lycos)

http://www.intrepid.net/igg/CD_Albums.txt
Plus CD (UK) $18.99 KEVIN COYNE Stumbling Onto Paradise CD (G) $22.50 KEVIN COYNE /DAGMAR KRAUSE Babble 15-trks CD $19.95 KEVIN COYNE Heart Burn 11-trks CD $19.95 KEVIN COYNE Matching Hands and Feet CD $19.95 KEVIN COYNE Millionaires and Teddy Bears CD (EEC)
1000 verified, *http://www.intrepid.net/igg/CD_Albums.txt* (WebCrawler)

Fairport Convention: Cropredy Festival
Every year during the second weekend in August, thousands gather in the fields of the village of Cropredy, Oxfordshire, England to hear **Fairport Convention** and their friends perform. The reunion gigs have been going on since 1980, after the band had decided
1000 verified, *http://www.novpapyrus.com/fairport/fccropredy.html* (OpenText)

The Current Lineup of **Fairport Convention**
1000 verified, *http://www.novpapyrus.com/fairport/fclineup.html* (Excite)

ndex of /uwp/music/artists/f/**fairport.convention**
Index of /uwp/music/artists/f/**fairport.convention** Ind
1000 verified, *http://peace.wit.com/uwp/music/artists/f/fairport.convention/* (Lycos)

Jethro Tull - Frequently Asked Questions
The members of Tull and **Fairport** overlap: Dave Pegg plays in both bands, and former Tull keyboardist Martin Allcock is also a core **Fairport** member. Cherry Lane Music currently has several books of Tull sheet music in print: "Crest of a Knave," "Rock Island,"
1000 verified, *http://jtull.rutgers.edu/JethroTull/FAQ.html* (WebCrawler)

Gig-A-Byte: Band: **Fairport Convention**
Fairport Convention
1 performance found.
Friday, November 1, 1996 Tin Angel (215) 928-0978, [Call for time] . Search | Band
1000 verified,
http://www.pagepoint.com/cgi-bin/gig_finder.cgi?band_flag=&database=band&operation=find&key=Fairport+Convention (OpenText)

Diagram 3.6 - MetaCrawler's search results format

3.3.1 Features

MetaCrawler has a reasonably sensible search query syntax. A searcher entering a string of words into the search box has the option of searching for the words as a phrase, or linking them by the AND or OR boolean operators. MetaCrawler does not allow nested boolean queries, but other options are available, such as the use of plus and minus signs to require or reject a given word from any results, and brackets around words to indicate that they are to be searched for as a phrase. Additionally, a search can be restricted by geographical region, such as country or continent, or by Internet domain.

One very useful feature MetaCrawler has is the ability to *verify* links and check that they are still in existence before retrieving them. Its designers have estimated that around 15% of all hits retrieved by MetaCrawler no longer exist, so removing them from the results list can save the searcher a certain amount of aggravation. The down side of this feature is that verifying a list of results can take over five times as long as merely collating it.

Once a search is initiated, the user waits while the results from each search engine are retrieved one by one. MetaCrawler sets a cap on the number of hits returned by each service. For instance where, say, AltaVista may return 40000 "relevant" hits for a given query, MetaCrawler will only take account of the first ten, which are assumed to be the most relevant, and ignore all of the others. Yet this procedure is still obviously going to be much slower than initiating a search on a single search engine, although the speed of service is rapidly improving and certainly is not so slow as to be a significant factor in the choice of search engine. A dramatic improvement will be the introduction of the *JavaCrawler*[33], a local copy of the MetaCrawler, loaded automatically from the MetaCrawler home page with a Java-compatible browser, which will sit on the user's desktop, thus handling the load there, rather than on a central server dealing with multiple requests from around the world.

Finally, after searching, retrieving, collating and (optionally) verifying the results from the nine search services, MetaCrawler confidence ranks them before presenting them to the user. This is done by summing the scores given by the various services before displaying them in what it calls a *voted* order.

3.3.2 Other Internet Protocols

As it has no database of its own, MetaCrawler's ability to find information stored in areas of the Internet other than the World Wide Web is dependent on the effectiveness of the nine services it searches at indexing these sources. It is suggested that MetaCrawler is slightly less effective when used for this purpose. This is for two reasons. Due to its nature, MetaCrawler's query interface seems to be a compromise of some of the interfaces of the nine services it searches, and is therefore less advanced than the more sophisticated interfaces, such as those of AltaVista and Open Text. Secondly, being a keyword only search interface, it loses the power of a classified index as a browsing-orientated search system and, as we have seen already, Yahoo's subject categories are logically structured to make retrieval of information from other parts of the Internet relatively simple.

3.3.3 Other Media

Again, as with other Internet protocols, MetaCrawler is dependent on the nine services it searches for its ability to retrieve media other than text. For the reasons outlined above, it is suggested that MetaCrawler is not the best service to use for this type of search.

3.3.4 Other Simultaneous Unified Search Indices

At the time of writing, the only Web search service comparable to MetaCrawler in terms of design and extensiveness is *SavvySearch*[34]. SavvySearch queries Aliweb, Alta Vista, CSTR, Deja News, Excite, Galaxy, Four11, FTPSearch95, Infoseek, Inktomi, InReference, Internet Movie Database, LinkStar, LookUP!, Lycos, Magellan, NlightN, OKRA, Open Text, Pathfinder, Point Search, shareware.com, SIFT - Stanford Information Filtering Tool, Tribal Voice, WebCrawler, WhoWhere?, Yahoo, and Yellow Pages. Because of the mix of search engines, in containing FTP, newsgroup and e-mail resources, it would be a better service than MetaCrawler for finding information stored in areas of the Internet other than the Web. It also allows one to limit a search by *Sources and Types of Information*, such as news, software, and images, and is available in 19 different languages, although the services it searches are the same (primary

English) databases in each language. The query language is no less advanced than MetaCrawler's either, but the list of results is not confidence ranked like MetaCrawler's and merely lists results by the search engine which found them, neglecting to remove duplicates or verify links, although an *integrate results* option, which takes a little time, is available once results have been collected.

It is also worth mentioning the *Internet Softbot*[35], an earlier version of MetaCrawler developed by Oren Etzioni, who also helped design the latter.

3.4 A Multi-engine Search Page - The All-in-One Search Page

Based on the principle of gathering together the search forms from a variety of existing search engines and making them available from a single page, the *All-in-One Search Page*[36] has now become so large that it is better described as a collection of multi-engine search pages all available from one site. It is maintained by William Cross, registered with the AlbanyNet access provider in the US, and is kept up to date by user submissions recommending their own favourite search engines. Its popularity is such that it had nearly 8 million accesses in its first year of operation, from 1 June 1995.

3.4.1 Features

The All-in-One Search Page groups its search engines into one of eleven subject categories - World Wide Web, General Internet, Specialized Interest, Software, People, News/Weather, Publications/Literature, Technical Reports, Documentation, Desk Reference, and Other Interesting Searches/Services. The front page of the service presents a main menu containing the subject categories, and clicking on a heading expands it to display the search forms of services encompassed in that category while leaving the rest of the main menu intact, rather like Peter Brown's *Guide* hypertext format[37]. Most search engines available from the All-in-One Search Page have a single sentence of text to describe them.

One of the drawbacks of using a multi-engine search page to access the database of a particular Web search service is the prospect of losing the advanced search options of services such as AltaVista. In partial remedy, the All-in-One Search Page also provides a direct link to each search page that it indexes.

3.4.2 Other Internet Protocols

The *General Internet* subject category contains search forms for many of the more useful Gopher resources on the Internet, such as *Jughead* and *Veronica.* Telnet sites are also available through this category via Hytelnet, as are many newsgroup search services such as DejaNews, Excite and AltaVista. FTP sites are also searchable through the *tile.net Internet References* service. Services for locating e-mail addresses are also available through the *People* subject category.

3.4.3 Other Media

Images, Sound and Video

The *Other Interesting Searches/Services* subject category contains some useful facilities for locating multimedia files, although it can take some browsing around before a suitable search engine is located.

Software

The All-in-One Search Page has a whole subject category devoted to software archives available on the Internet.

3.4.4 Other Multi-engine search pages

There are quite possibly thousands of multi-engine search pages available on the Web if one looks hard enough. A popular custom for individual users when assembling their Web home page is to paste search forms for favourite search engines onto them, and it is quite possible that some of the most comprehensive are also some of the least publicised. One need only take a look through newsgroups such as `comp.infosystems.www.announce` and almost daily one can find another individual advertising his or her own multi-engine search page. The concept is very much the same for all of them. Some of the larger ones, like the All-in-One Search Page, categorise search engines by subject. An extensive list of multi-engine search pages, of varying quality can be found in the Yahoo! All-in-One Search Pages subdirectory[38].

3.5 A Search Engine of Search Engines - C/Net's "Search.Com"

The next step up the evolutionary scale from the multi-engine search page is the search engine of search engines. Where the All-in-One Search Page, which is becoming so large that it is almost a search engine of search engines itself, presents a list of search forms for various search facilities around the Internet, *Search.Com*[39] presents 23 different subject categories, from *Arts* to the *Web*, which in turn point to many different subject-specific search engines available. It even has its own search form which, rather than finding a specific document in a given subject area, instead recommends a particular search engine dedicated to that subject. In an expanding area, with more and more information being indexed by more and more search engines, Search.Com presents just one more way of making the choice of which one to use.

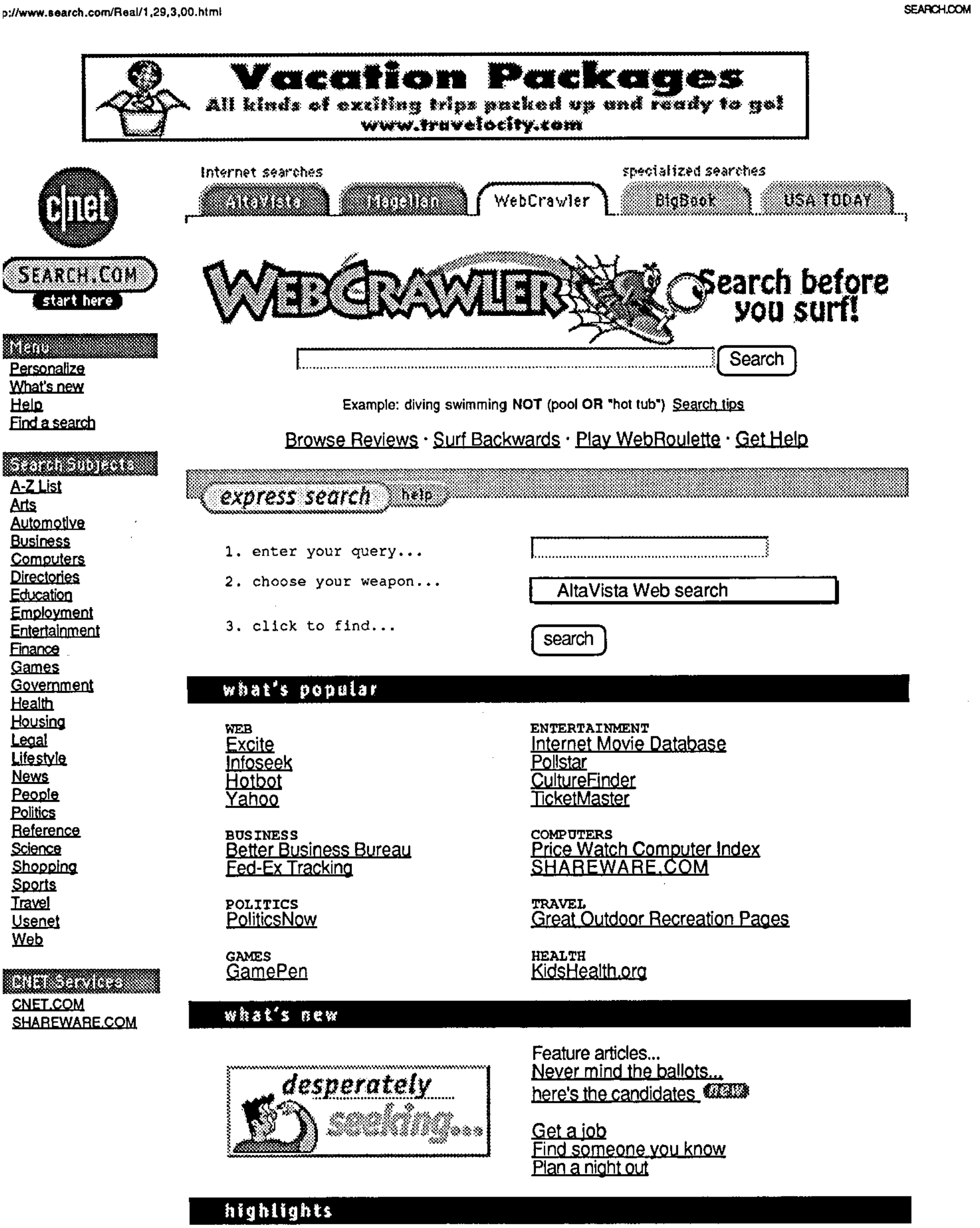
3.5.1 Features

Search.Com offers three ways of choosing the appropriate search tool.

- by browsing through the subject options presented at the left-hand edge of the screen,
- if the name of a search engine is known then a single alphabetical list of all the search engines available from Search.Com can be called up,
- by using Search.Com's search form. If the search is kept very general, such as to a broad subject area, then a dedicated subject-specific search engine may be found.

Additionally, five of the larger Web search engines' forms are available directly from the Search.Com home page. These seem to change quite frequently as new strategic alliances are formed, but at the time of writing these search engines are AltaVista, Magellan, WebCrawler, BigBook, and USA Today.

A *what's new* section keeps the user up to date with Search.Com's latest developments, and a *what's popular* section presents us with links to some of the Internet's more sought after search facilities. Both of these features are available from the Search.Com home page.



Example: diving swimming **NOT** (pool **OR** "hot tub") Search tips

Browse Reviews · Surf Backwards · Play WebRoulette · Get Help

Menu
- Personalize
- What's new
- Help
- Find a search

Search Subjects
- A-Z List
- Arts
- Automotive
- Business
- Computers
- Directories
- Education
- Employment
- Entertainment
- Finance
- Games
- Government
- Health
- Housing
- Legal
- Lifestyle
- News
- People
- Politics
- Reference
- Science
- Shopping
- Sports
- Travel
- Usenet
- Web

CNET Services
- CNET.COM
- SHAREWARE.COM

express search help

1. enter your query...
2. choose your weapon... AltaVista Web search
3. click to find... search

what's popular

WEB
Excite
Infoseek
Hotbot
Yahoo

ENTERTAINMENT
Internet Movie Database
Pollstar
CultureFinder
TicketMaster

BUSINESS
Better Business Bureau
Fed-Ex Tracking

COMPUTERS
Price Watch Computer Index
SHAREWARE.COM

POLITICS
PoliticsNow

TRAVEL
Great Outdoor Recreation Pages

GAMES
GamePen

HEALTH
KidsHealth.org

what's new

Feature articles...
Never mind the ballots...
here's the candidates NEW

Get a job
Find someone you know
Plan a night out

highlights

Diagram 3.7 - The Search.Com home page

Finally, Search.Com allows one to personalise its home page. The user can select a number of search engines from Search.Com's database and paste their search forms onto one page so that whenever Search.Com's site is accessed the personalised search page becomes available. This feature also allows the user to add his or her own favourite links to the page.

3.5.2 Other Internet Protocols

FTP

Available from the *Computers* subject category is a search form for *tile.net FTP*, a useful FTP search service.

E-mail Addresses

The *Four11 Directory Service* is available from the *highlights* section on the Search.Com home page. In addition, the *Web* subject category contains search forms for a number of other useful e-mail address services, such as *InterNIC Whois* and *Yahoo! People Search*. The *Directories* category is another good source of search facilities.

Newsgroups

Search.Com has a whole subject category dedicated to Usenet search facilities. This includes search forms for all the major Usenet search tools, such as DejaNews, AltaVista, Excite and InfoSeek.

Gopher and Telnet

Search.Com is not as good for these sources and does not seem to offer any dedicated databases for these two particular areas of the Internet.

3.5.3 Other Media

Images, Sound and Video

Finding the appropriate search engines can be a problem. Of seven hits obtained when searching for multimedia on Search.Com's search form, only *Search Microsoft* and *Shareware.Com* were true multimedia search resources. Neither are dedicated to image, sound and video files.

Software

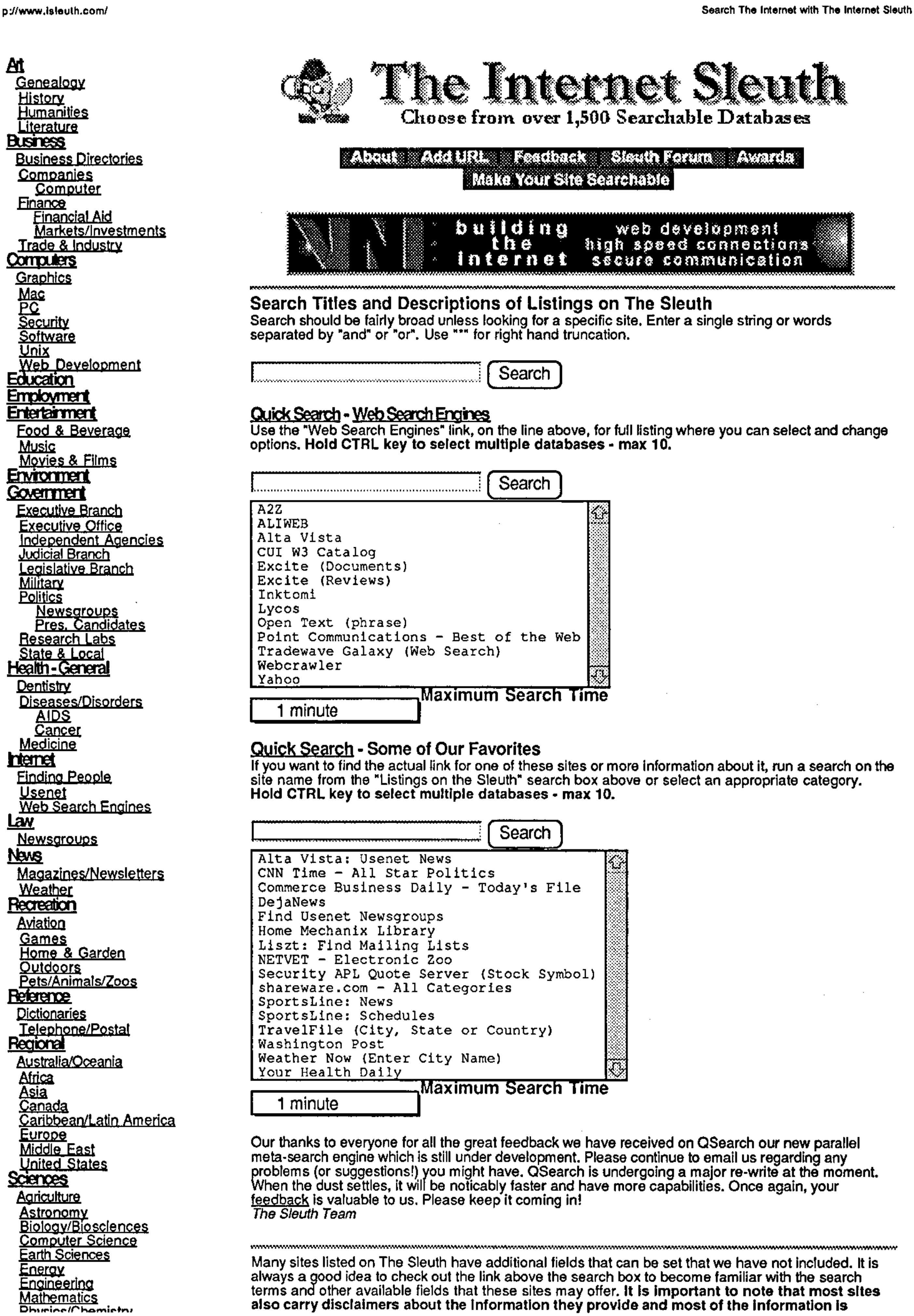
C/Net's sister service to Search.Com is called *Shareware.Com*[40] and is one of the Web's biggest software archives. Links to Shareware.Com are available both from the Search.Com home page and the *Computers* subject category. Also in this section is a search form for the *SBA Shareware Library*[41], another expansive software directory available on the Web.

3.5.4 Other search engines of search engines

The closest Web search tool in design to Search.Com is the *Internet Sleuth*[42]. These two services are very similar, but there are important differences. The Internet Sleuth's subject headings are more detailed and much easier to browse than Search.Com's. Its combined database of over 1500 separate search engines, devoted to all areas of the Internet and including various multimedia search facilities make it considerably bigger than Search.Com and, on balance, the better option for most subjects. Its popularity is much smaller than Search.Com's however, and this could be because fewer people are aware of its existence. This may well be due to C/Net's considerable Internet clout, currently lying 18th on Web21's hot100 list of most popular Web sites by the number of accesses. The Internet Sleuth is not even in the top 100[43].

3.6 Subject-specific Search Engines

A quick look at services such as Search.Com and the Internet Sleuth will give a good idea of the sheer number of search engines available for surprisingly narrow fields of interest. Some of the better ones are so comprehensive as to be the only stop for particular pieces of information. The *Internet Movie Database*[44], for instance, started life as a voluntary service run by a hobbyist requesting submissions from visitors. It is now *the* place took look for information on films, directors, actors and actresses. If all one wishes to know is the date of birth of a certain actor, searching AltaVista or Lycos would simply make no sense at all. A quick query on the Internet Movie Database would give the answer. Likewise, if one has been browsing a computer magazine and wishes to know the meaning of the acronym WYSIWYG, a simple keyword search in *FOLDOC* - the *Free Online Dictionary of Computing*[45] would give a full definition and



p://www.isleuth.com/

Search The Internet with The Internet Sleuth

- Art
  - Genealogy
  - History
  - Humanities
  - Literature
- Business
  - Business Directories
  - Companies
    - Computer
  - Finance
    - Financial Aid
    - Markets/Investments
  - Trade & Industry
- Computers
  - Graphics
  - Mac
  - PC
  - Security
  - Software
  - Unix
  - Web Development
- Education
- Employment
- Entertainment
  - Food & Beverage
  - Music
  - Movies & Films
- Environment
- Government
  - Executive Branch
  - Executive Office
  - Independent Agencies
  - Judicial Branch
  - Legislative Branch
  - Military
  - Politics
    - Newsgroups
    - Pres. Candidates
  - Research Labs
  - State & Local
- Health - General
  - Dentistry
  - Diseases/Disorders
    - AIDS
    - Cancer
  - Medicine
- Internet
  - Finding People
  - Usenet
  - Web Search Engines
- Law
  - Newsgroups
- News
  - Magazines/Newsletters
  - Weather
- Recreation
  - Aviation
  - Games
  - Home & Garden
  - Outdoors
  - Pets/Animals/Zoos
- Reference
  - Dictionaries
  - Telephone/Postal
- Regional
  - Australia/Oceania
  - Africa
  - Asia
  - Canada
  - Caribbean/Latin America
  - Europe
  - Middle East
  - United States
- Sciences
  - Agriculture
  - Astronomy
  - Biology/Biosciences
  - Computer Science
  - Earth Sciences
  - Energy
  - Engineering
  - Mathematics
  - Physics/Chemistry

The Internet Sleuth

Choose from over 1,500 Searchable Databases

Search Titles and Descriptions of Listings on The Sleuth

Search should be fairly broad unless looking for a specific site. Enter a single string or words separated by "and" or "or". Use "*" for right hand truncation.

Quick Search - Web Search Engines

Use the "Web Search Engines" link, on the line above, for full listing where you can select and change options. **Hold CTRL key to select multiple databases - max 10.**

```
A2Z
ALIWEB
Alta Vista
CUI W3 Catalog
Excite (Documents)
Excite (Reviews)
Inktomi
Lycos
Open Text (phrase)
Point Communications - Best of the Web
Tradewave Galaxy (Web Search)
Webcrawler
Yahoo
```

Maximum Search Time: 1 minute

Quick Search - Some of Our Favorites

If you want to find the actual link for one of these sites or more information about it, run a search on the site name from the "Listings on the Sleuth" search box above or select an appropriate category. **Hold CTRL key to select multiple databases - max 10.**

```
Alta Vista: Usenet News
CNN Time - All Star Politics
Commerce Business Daily - Today's File
DejaNews
Find Usenet Newsgroups
Home Mechanix Library
Liszt: Find Mailing Lists
NETVET - Electronic Zoo
Security APL Quote Server (Stock Symbol)
shareware.com - All Categories
SportsLine: News
SportsLine: Schedules
TravelFile (City, State or Country)
Washington Post
Weather Now (Enter City Name)
Your Health Daily
```

Maximum Search Time: 1 minute

Our thanks to everyone for all the great feedback we have received on QSearch our new parallel meta-search engine which is still under development. Please continue to email us regarding any problems (or suggestions!) you might have. QSearch is undergoing a major re-write at the moment. When the dust settles, it will be noticably faster and have more capabilities. Once again, your feedback is valuable to us. Please keep it coming in!
The Sleuth Team

Many sites listed on The Sleuth have additional fields that can be set that we have not included. It is always a good idea to check out the link above the search box to become familiar with the search terms and other available fields that these sites may offer. **It is important to note that most sites also carry disclaimers about the information they provide and most of the information is**

Diagram 3.8 - The Internet Sleuth home page

relevant links.

An important area where subject-specific search engines have become perhaps the best tool is for finding company information and corporate Web sites. Some of the better directories in this field are the *World Wide Yellow Pages*[46] and the *Global On-Line Directory*[47].

Likewise, in the case of multimedia information, media-specific search engines can produce a result much more easily than any ordinary keyword search engine. The *Query By Image Content* service[48] matches images by colour content. Searching for a green object on a red background is easy with QBIC, but practically impossible using any other search engine available on the Web.

Being aware of the existence of these tools is a problem of course, and so many exist that even the most clued-up of experts is not going to be aware of all of them. Search engines of search engines may be the answer to this, as long as they have the necessary databases available and easy to find. The Internet Movie Database is well publicised and can be located from all major search engines. At the time of writing, however, FOLDOC did not seem to be available on the Internet Sleuth, although it was easily located under the *Computers* subject category on Search.Com.

3.7 A Geographically-specific Search Engine - AC/DC

The *Academic Directory*, or *AC/DC*[49] as it is known, indexes only UK academic sites within the Internet domain `ac.uk`. Its creators cite two reasons for the necessity for such a search facility.

- Most of the major search engines are based in the US and, although they index UK documents, are more concerned with US documents, which constitute the vast majority of material available on the Internet. Much of this is not greatly relevant to UK users who therefore have to browse through sometimes long results lists looking for relevant documents.

- Much bandwidth on the US-UK connection is used by UK users making use of search engines based the other side of the Atlantic. There is clearly a need, therefore, for a comprehensive database of Web documents based in the UK[50].

AC/DC was initially set up as an experimental Web project using the *Harvest* gatherer and broker[51] with UK academic institutions invited to set up their own indices and contribute them to the AC/DC project. By June 1996, AC/DC had indexed 175000 Web, Gopher and news postings across over a thousand `ac.uk` sites. However, the issue of whether or not it is intended to continue running AC/DC and expand the project into a long-term service is still unclear.

3.7.1 Features

AC/DC is available to search either the Web or Usenet. An advanced query form allows the following:

- optional case-insensitive matching
- phrase searching
- substring searching
- Boolean AND and OR operators
- the ability to restrict a search by field. Fields used are particular to the Harvest broker (as are all of the above query options) and include abstract, author, keywords, title, and URL, as well as many more technical options dependent on the Harvest gatherers.

3.7.2 Other Internet Protocols

AC/DC indexes Web, Gopher and newsgroup documents.

3.7.3 Other Media

AC/DC is designed to index text files available from servers in the UK academic community. Although clever use of the search query interface may result in sound, image, video or software files being located, the service was never designed with multimedia in mind and would not be a good starting point for retrieval of any medium other than text.

3.7.4 Other Geographically-specific search engines

At present there are no large database services of the comprehensiveness of some of the US services such as Lycos and AltaVista specifically designed for the UK market, although services such as the *UK Index*[52], the *UK Directory*[53] and the *UK Internet World Wide Web*[54] go some way towards remedying this, albeit on a much smaller scale. This may be set to change, but any progress looks likely to come from the US itself. Yahoo! has already made available two geographically-specific versions of its index in Japan[55] and Canada[56] and are looking to release a UK version. Rumours have passed around some UK newsgroups, such as `uk.jips`, that AltaVista are looking to do the same. Other countries have their own indices however. The *Austrian Internet Directory*[57], *Flipper*[58], *GoGREECE*[59], *Recursos de Internet en Español y Portugués*[60], *SwissSearch*[61], and *WebRider*[62] all serve various European countries. Systematic comparison of each is problematic however given the vastly different spread of Internet-based information available in each country.

3.8 A Fee-based Service - Infoseek Professional

This report concentrates on the free search services available on the Web, and a full review of fee-based services is therefore beyond its bounds. It would make sense to look at some of the services available for those who are prepared to pay however, to see if the extra expense is worthwhile. One of the more popular fee-based services, *Infoseek Professional*[63], is considered briefly below.

3.8.1 Infoseek Professional

Infoseek has its own free service, the *Infoseek Guide*[64], with a large database of URLs and many of the usual advanced search features such as phrase search, the use of plus and minus signs and case-insensitive searching. Although superior in many ways to several of the large automatically-assembled search engines, it does however possess sufficient similarities not to warrant further discussion here.

Infoseek's fee-based service, *Infoseek Professional* is of more interest here. For $4.95 a month, an individual can sign up and receive 50 transactions (corporate membership plans are also available). Transactions do not apply to the database of Web documents, which remains free, but includes the following:

- Usenet News (Internet newsgroup articles)
- Cineman (entertainment reviews)
- Wire Services (latest business news)
- Computer Periodicals (computer industry news)
- Health and Medicine (medical news and research)
- Corporate Information (key facts on companies)

In addition, the following services are available on a pay-per-search basis (prices vary depending on the service):

- Hoover Company Profiles (specific private and public company data)
- MDX Health (consumer health database)
- Infoworld (desktop computing issues)
- CorpTech Directory (data on emerging technology product manufacturers)
- Microcomputer Abstracts (personal computing developments)
- CSA Biomedical Database (summaries of published materials in biomedicine)
- CSA Computer and Engineering Database (summaries of published materials in computing and engineering)
- CSA Worldwide Market Research Database (market research report summaries)
- SoftBase (reviews, evaluations, and comparisons of software products)
- EDGE Newsletters (information on AT&T and workgroup computing)

3.8.2 Other Fee-based services

Other popular fee-based services are *NlightN*[65], *IBM infoMarket*[66] and the *Electric Library*[67]. None have quite as good a free service as Infoseek, but each follows the same principle as Infoseek Professional in using the Internet as a mode of access to existing commercial online databases. Such is the depth and authority of these databases that they are clearly a worthwhile expense for those regularly requiring more specific information on good authority, and for this type of searcher, it will be a long time before the free Web search engines, or indeed the Web as an information source, catch up.

3.9 A Search Engine with Document Reviews - Magellan

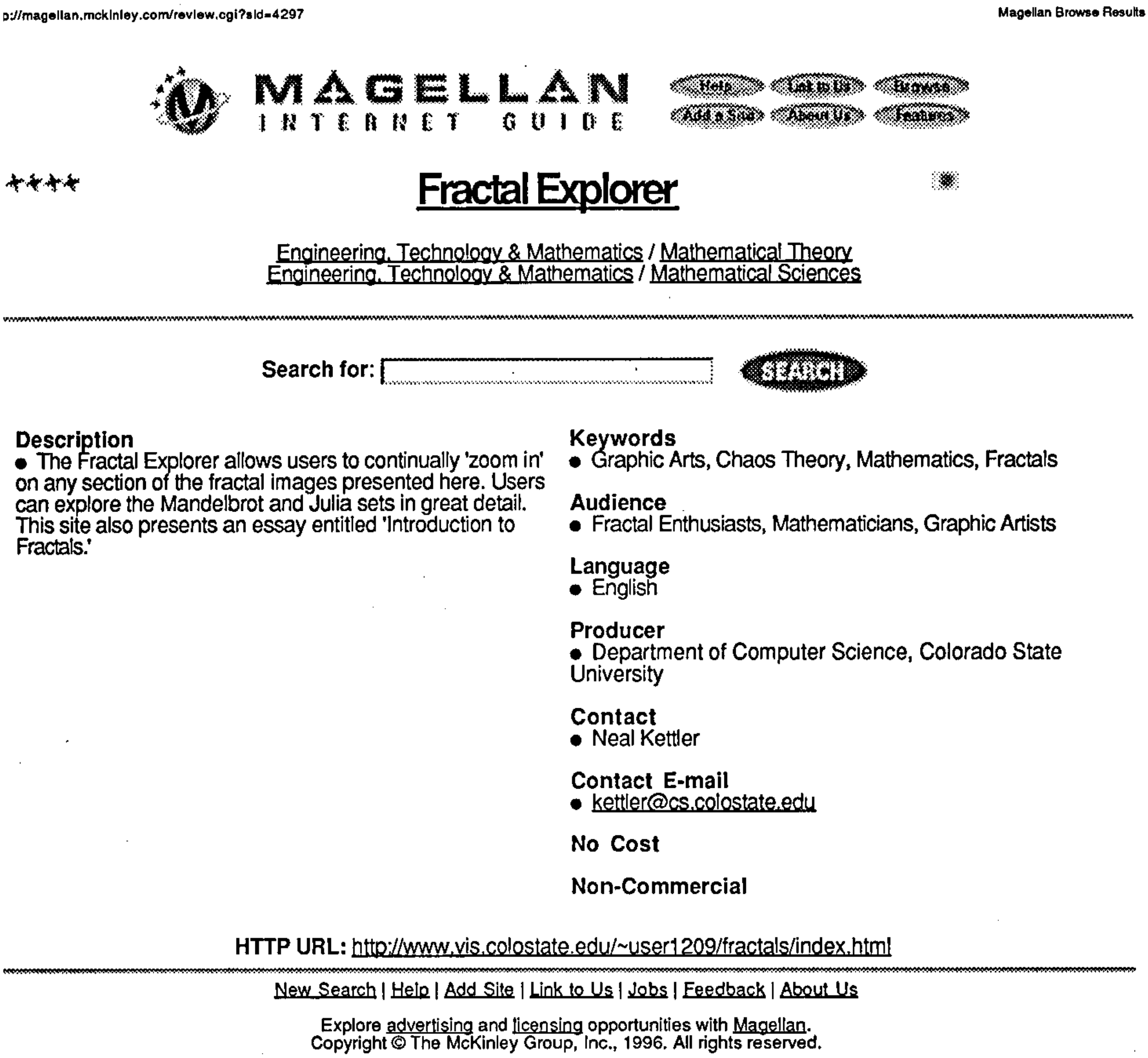
The *Magellan Internet Guide*[68], run by The McKinley Group Inc., is an example of a new generation of search engine emerging in the first few months of 1996, which has responded to the debate surrounding quality and authority of information found on the Internet by providing reviews of Web documents in its database. Like Yahoo!, Magellan contains a hierarchically-organised directory of Web documents arranged under 23 subject headings - Arts, Business and Economics, Communications, Computing, Daily Living, Education, Employment, Entertainment and Pop Culture,, Environment, Food, Health, Humanities, Internet, KidZone, Law, Mathematics and Technology, Music, News, Politics, Science, Spirituality, Sports, and Travel - and over 500 subheadings.

Magellan's editorial content also helps the searcher considerably. Each reviewed article on a list retrieved contains two or three lines about the document, including a link to the document itself and to Magellan's comprehensive review of it, a rating of one to four stars based on marks out of ten in three areas (depth, ease of exploration, and "net appeal"), and a little "green light" symbol if the document is deemed by Magellan staff to be free from adult-orientated content.

Martin Courtois stated that in January 1996, Magellan had about 30000 reviews of documents in its database[69]. By June, according to McKinley's own figures, this had grown to around 40000 sites[70]. In addition to this, Magellan has a huge database of, as yet, unreviewed and unedited sites which, according to the same sources, stood at 1 million entries in January and had grown to some 15 million by June.

3.9.1 Features

Each Magellan review contains information about the audience the site is aimed at, the



p://magellan.mckinley.com/review.cgi?sid=4297 Magellan Browse Results

MAGELLAN
INTERNET GUIDE

★★★★

Fractal Explorer

Engineering, Technology & Mathematics / Mathematical Theory
Engineering, Technology & Mathematics / Mathematical Sciences

Search for: ☐

Description
- The Fractal Explorer allows users to continually 'zoom in' on any section of the fractal images presented here. Users can explore the Mandelbrot and Julia sets in great detail. This site also presents an essay entitled 'Introduction to Fractals.'

Keywords
- Graphic Arts, Chaos Theory, Mathematics, Fractals

Audience
- Fractal Enthusiasts, Mathematicians, Graphic Artists

Language
- English

Producer
- Department of Computer Science, Colorado State University

Contact
- Neal Kettler

Contact E-mail
- kettler@cs.colostate.edu

No Cost

Non-Commercial

HTTP URL: http://www.vis.colostate.edu/~user1209/fractals/index.html

New Search | Help | Add Site | Link to Us | Jobs | Feedback | About Us

Explore advertising and licensing opportunities with Magellan.
Copyright © The McKinley Group, Inc., 1996. All rights reserved.

Diagram 3.9 - An example of a Magellan review

producer of the site including a contact address for e-mail, language the site is written in, keywords, and additional information regarding the site such as whether it is commercial or moderated, as well as the review itself which includes a rating of from one to four stars.

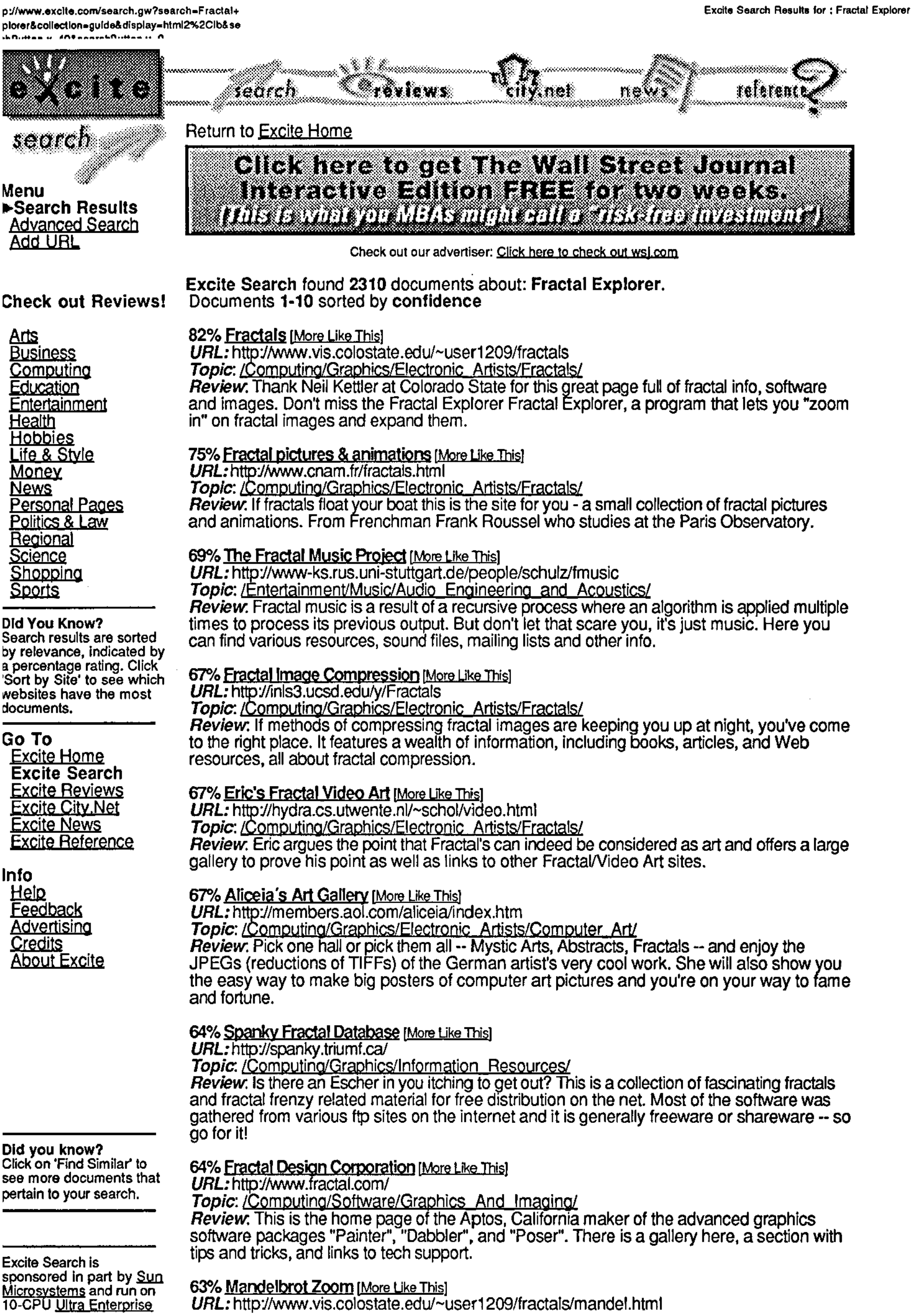
In addition to Magellan's browsing-centred subject tree, which accesses the database of reviewed items, the search engine also has a simple keyword search interface available on every screen in Magellan. This search facility defaults to the boolean OR command for any keywords entered, but allows one to fine tune the search by placing a minus sign ('-') in front of a keyword to exclude documents which contain that keyword, or a plus sign ('+') to make sure all documents returned contain a particular keyword. Magellan relevance ranks its output and, unusually, its documentation stipulates the six criteria it uses for doing so. These are:

- whether or not the article has been reviewed by Magellan (reviewed documents are usually considered more relevant,
- the number of a searcher's keywords that appear in the document,
- whether the keywords can be found in the document's title,
- whether or not the searcher's keywords appear in Magellan's review or its own set of keywords for the document,
- whether the keywords can be found in the document's URL,
- how many times each keyword appears in the body of the document.

Finally, from the point of view of an Internet "surfer", Magellan's capacity to display *related topics* is another feature which could prove useful. In many cases a given search result will supply a list of subheadings from its subject tree which are deemed to be relevant to that particular search. Clicking on the topic will link to that part of Magellan's database.

3.9.2 Other Internet Protocols

The *Magellan People Finder*[71] is available from the Magellan home page, and provides telephone and fax directories for the USA and Canada. Other Internet resources do not appear to be indexed by Magellan.



p://www.excite.com/search.gw?search=Fractal+
plorer&collection=guide&display=html2%2Clb&se

Excite Search Results for : Fractal Explorer

Return to Excite Home

Excite Search found **2310** documents about: **Fractal Explorer.**
Documents **1-10** sorted by **confidence**

82% Fractals [More Like This]
URL: http://www.vis.colostate.edu/~user1209/fractals
Topic: /Computing/Graphics/Electronic_Artists/Fractals/
Review: Thank Neil Kettler at Colorado State for this great page full of fractal info, software and images. Don't miss the Fractal Explorer Fractal Explorer, a program that lets you "zoom in" on fractal images and expand them.

75% Fractal pictures & animations [More Like This]
URL: http://www.cnam.fr/fractals.html
Topic: /Computing/Graphics/Electronic_Artists/Fractals/
Review: If fractals float your boat this is the site for you - a small collection of fractal pictures and animations. From Frenchman Frank Roussel who studies at the Paris Observatory.

69% The Fractal Music Project [More Like This]
URL: http://www-ks.rus.uni-stuttgart.de/people/schulz/fmusic
Topic: /Entertainment/Music/Audio_Engineering_and_Acoustics/
Review: Fractal music is a result of a recursive process where an algorithm is applied multiple times to process its previous output. But don't let that scare you, it's just music. Here you can find various resources, sound files, mailing lists and other info.

67% Fractal Image Compression [More Like This]
URL: http://inls3.ucsd.edu/y/Fractals
Topic: /Computing/Graphics/Electronic_Artists/Fractals/
Review: If methods of compressing fractal images are keeping you up at night, you've come to the right place. It features a wealth of information, including books, articles, and Web resources, all about fractal compression.

67% Eric's Fractal Video Art [More Like This]
URL: http://hydra.cs.utwente.nl/~schol/video.html
Topic: /Computing/Graphics/Electronic_Artists/Fractals/
Review: Eric argues the point that Fractal's can indeed be considered as art and offers a large gallery to prove his point as well as links to other Fractal/Video Art sites.

67% Aliceia's Art Gallery [More Like This]
URL: http://members.aol.com/aliceia/index.htm
Topic: /Computing/Graphics/Electronic_Artists/Computer_Art/
Review: Pick one hall or pick them all -- Mystic Arts, Abstracts, Fractals -- and enjoy the JPEGs (reductions of TIFFs) of the German artist's very cool work. She will also show you the easy way to make big posters of computer art pictures and you're on your way to fame and fortune.

64% Spanky Fractal Database [More Like This]
URL: http://spanky.triumf.ca/
Topic: /Computing/Graphics/Information_Resources/
Review: Is there an Escher in you itching to get out? This is a collection of fascinating fractals and fractal frenzy related material for free distribution on the net. Most of the software was gathered from various ftp sites on the internet and it is generally freeware or shareware -- so go for it!

64% Fractal Design Corporation [More Like This]
URL: http://www.fractal.com/
Topic: /Computing/Software/Graphics_And_Imaging/
Review: This is the home page of the Aptos, California maker of the advanced graphics software packages "Painter", "Dabbler", and "Poser". There is a gallery here, a section with tips and tricks, and links to tech support.

63% Mandelbrot Zoom [More Like This]
URL: http://www.vis.colostate.edu/~user1209/fractals/mandel.html

Menu
- ▶**Search Results**
- Advanced Search
- Add URL

Check out Reviews!

- Arts
- Business
- Computing
- Education
- Entertainment
- Health
- Hobbies
- Life & Style
- Money
- News
- Personal Pages
- Politics & Law
- Regional
- Science
- Shopping
- Sports

Did You Know?
Search results are sorted by relevance, indicated by a percentage rating. Click 'Sort by Site' to see which websites have the most documents.

Go To
- Excite Home
- **Excite Search**
- Excite Reviews
- Excite City.Net
- Excite News
- Excite Reference

Info
- Help
- Feedback
- Advertising
- Credits
- About Excite

Did you know?
Click on 'Find Similar' to see more documents that pertain to your search.

Excite Search is sponsored in part by Sun Microsystems and run on 10-CPU Ultra Enterprise

Diagram 3.10 - An example of an Excite review

3.9.3 Other Media

Magellan's tendency to put reviewed sites in order of rating scores rather than content makes it a particular frustrating source for those looking for specific pieces of information. Although the unreviewed database is very large, pinpointing details such as multimedia can be very arduous. Until a better system is arrived at for listing results, possibly by using a more sophisticated query language, Magellan would not be the best site to explore for those seeking multimedia material.

3.9.4 Other search engines with document reviews

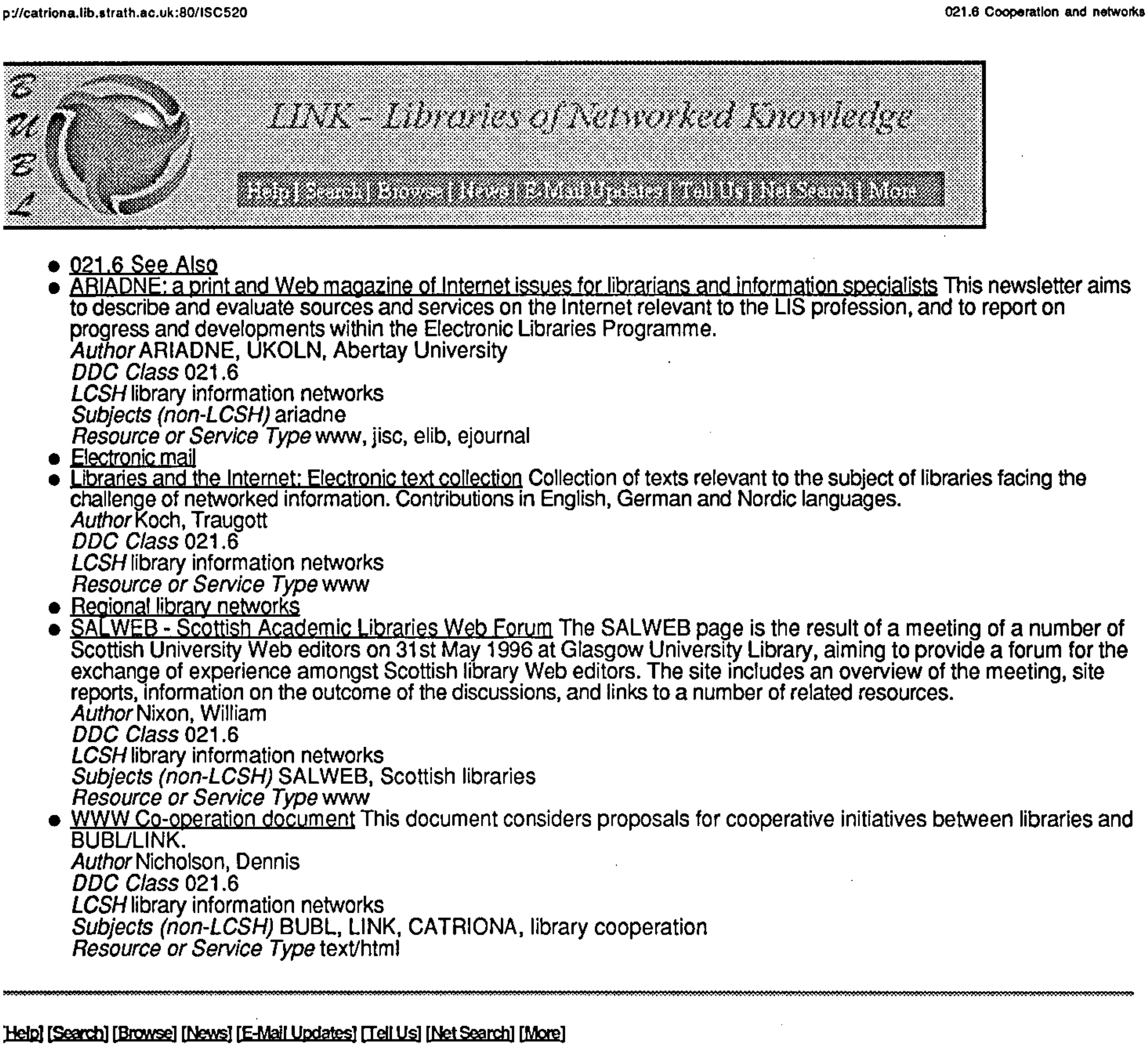
Excite is very similar to Magellan in that it has a large database of Web documents, a subset of which contains items which have been reviewed. Magellan certainly contains the more comprehensive reviews - Excite's tend to be short and quirky, while Magellan's reviews often provide useful relevant information not to be found in the document itself - but Excite's query language certainly makes it considerably easier to find precise details.

These are not the only two services offering reviews. *WebCrawler*'s merging with the *GNN Whole Internet Catalog* means that it too now offers a limited database of reviewed sites[72]. Likewise, Lycos has assembled its own database of reviews, *Lycos A2Z*[73], and also owns *Point*[74], another popular Web review service. *The iGuide Net Reviews*[75] is another site worth looking at, as are the graphics-intensive *JumpCity*[76], and the impressive-looking *CyberHound*[77].

3.10 A Search Engine with Document Records - LINK

As mentioned earlier in this report, LINK (Libraries of Networked Knowledge) is the practical result of the British Library-funded CATRIONA project, which aimed to create a catalogue of Internet resources based on the Z39.50 standard for information retrieval and the MARC catalogue record. The project demonstrated that,

> *"there were already commercially available Z39.50 OPAC clients which,*



p://catriona.lib.strath.ac.uk:80/ISC520 021.6 Cooperation and networks

BUBL

LINK - Libraries of Networked Knowledge

Help | Search | Browse | News | E-Mail Updates | Tell Us | Net Search | More

- 021.6 See Also
- ARIADNE: a print and Web magazine of Internet issues for librarians and information specialists This newsletter aims to describe and evaluate sources and services on the Internet relevant to the LIS profession, and to report on progress and developments within the Electronic Libraries Programme.
  Author ARIADNE, UKOLN, Abertay University
  DDC Class 021.6
  LCSH library information networks
  Subjects (non-LCSH) ariadne
  Resource or Service Type www, jisc, elib, ejournal
- Electronic mail
- Libraries and the Internet: Electronic text collection Collection of texts relevant to the subject of libraries facing the challenge of networked information. Contributions in English, German and Nordic languages.
  Author Koch, Traugott
  DDC Class 021.6
  LCSH library information networks
  Resource or Service Type www
- Regional library networks
- SALWEB - Scottish Academic Libraries Web Forum The SALWEB page is the result of a meeting of a number of Scottish University Web editors on 31st May 1996 at Glasgow University Library, aiming to provide a forum for the exchange of experience amongst Scottish library Web editors. The site includes an overview of the meeting, site reports, information on the outcome of the discussions, and links to a number of related resources.
  Author Nixon, William
  DDC Class 021.6
  LCSH library information networks
  Subjects (non-LCSH) SALWEB, Scottish libraries
  Resource or Service Type www
- WWW Co-operation document This document considers proposals for cooperative initiatives between libraries and BUBL/LINK.
  Author Nicholson, Dennis
  DDC Class 021.6
  LCSH library information networks
  Subjects (non-LCSH) BUBL, LINK, CATRIONA, library cooperation
  Resource or Service Type text/html

[Help] [Search] [Browse] [News] [E-Mail Updates] [Tell Us] [Net Search] [More]

Diagram 3.11 - A LINK catalogue record

> *having searched for and retrieved MARC records describing electronic resources and containing URLs, could automatically load a client like Netscape, pass it the URL and so deliver the electronic resource to the desktop via the catalogue."*[78]

What this meant, in simple terms, was that CATRIONA's aim of building an application program capable of retrieving, classifying and cataloguing Internet documents was already a reality. There *were* applications capable of doing this, even then, as the options were being explored. The software used to build LINK was AmeriTech's *NetPublisher*[79], and it was capable of fulfilling all of the criteria necessary to create the system envisaged by CATRIONA. LINK is still in its development stage at the moment, but a prototype is accessible and the service is expected to be fully operational by Autumn 1996.

3.10.1 Features

There are two principle features central to the design of LINK which distinguish it from most other Internet search services.

- It is accessible via both Web and Z39.50 clients,
- All of the resources indexed by it have their own catalogue records.

Each catalogue record in LINK contains the following fields, which are searchable:

- Item name
- Item abstract
- Item content type
- Author
- Library of Congress Subject Heading (LCSH)
- Subjects (non-authority controlled subject headings ie. free text)
- LINKhelp
- Resource or Service Category
- Date Added
- Date Expires

LINK is browsable via a classified index available in alphabetical order or by Dewey

Decimal Classification, or searchable via a search query form which allows the following options:

- Boolean queries using the AND, OR and NOT operators
- field searching within any of the above fields
- left and right hand truncation

Unfortunately phrase searching is not available on LINK, but even without it the search interface is clearly very powerful.

3.10.2 Other Internet Protocols

FTP and Gopher resources are fully indexed and can be searched for via the *Resource or Service Category* catalogue field by entering *"ftp?"* or *"gopher?"* respectively (right hand truncation). Newsgroups do not appear to be catered for, but listservs are searchable by typing *"listserv?"* in the same catalogue field. Telnet sites which are library OPAC catalogues can be searched using *"opac?"*. Although LINK does not have its own e-mail directory, presumably searching the *Author* catalogue field may lead to Web resources containing an e-mail address. All in all, the LINK database contains documents found in a wide spread of Internet protocols.

3.10.3 Other Media

LINK is not so strong when used for locating multimedia material. Indeed, in keeping with the BUBL Information Service, it is designed with textual information of relevance to the UK higher education community in mind. It is not, therefore, a recommended source for those searching for non-text files.

3.10.4 Other search engines with document records

Although LINK looks to be the most comprehensive resource which uses the concept of catalogue records, designed for a UK audience, there are equivalents elsewhere, most notably the American *NetFirst* search engine[80]. Like LINK, it grew out of a library-initiated research project, in this case, "Building a Catalogue of Internet-Accessible

Materials," run by the OCLC[81]. Because they each catalogue a different set of resources, the two services are not really comparable, but in both cases the general principle is the same: to build a database of Internet resources based on more traditional library cataloguing techniques.

At the moment, both services are tiny compared with automatically-assembled databases like AltaVista and Excite, and this is almost certain to remain the case given the ethics behind the two services. AltaVista and Excite will remain the services to go to for those wishing to seek out the more subjective areas of the Internet, such as advertisements and hobbyists home pages, but it is just possible that services like LINK and NetFirst will overtake them in the number of benefits available to those looking for academic documents and information relating to higher education.

3.11 The Future?

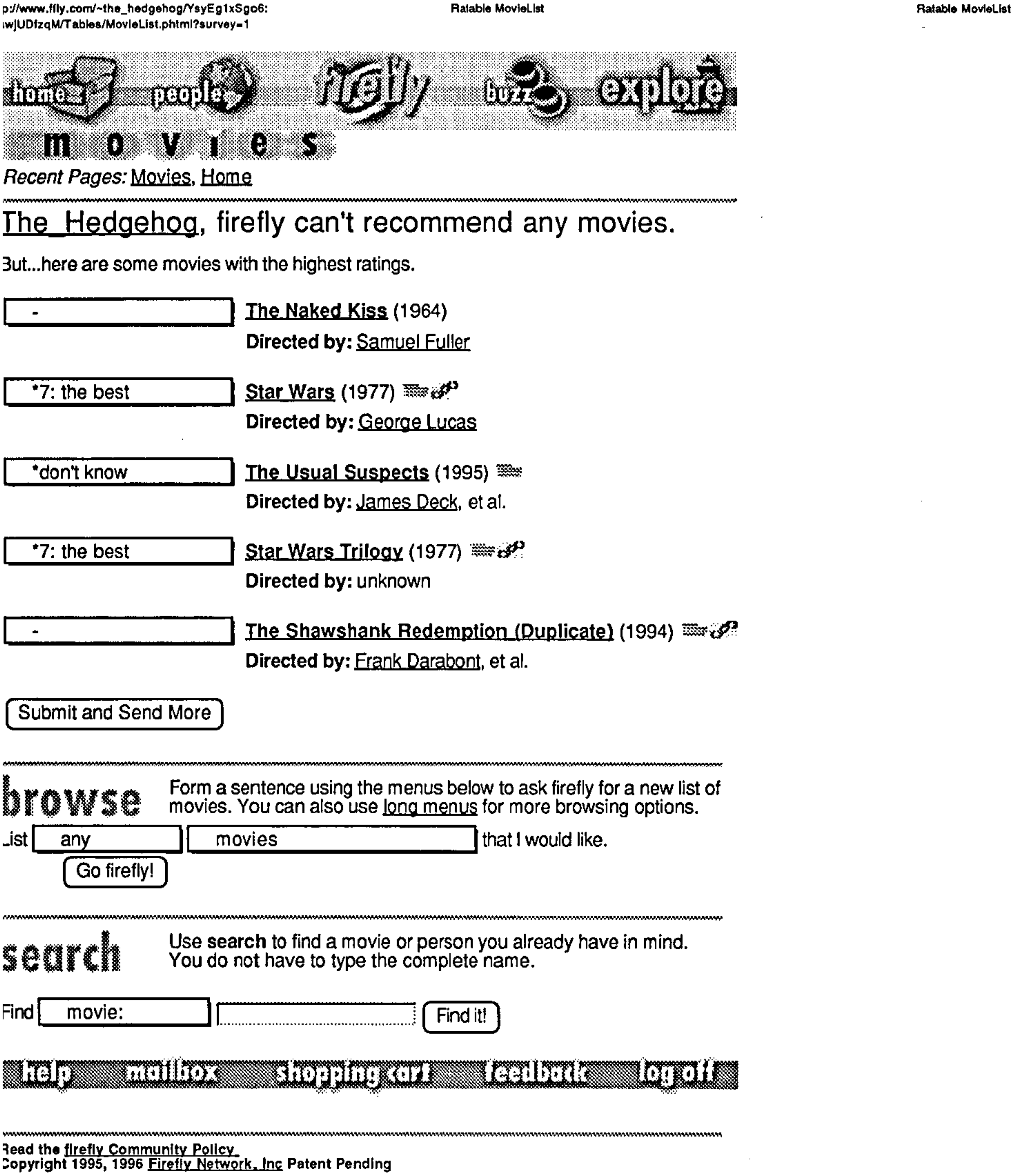
Firefly, WebCompass and Other User-Configurable Search Tools

As the Web develops and information providers strive to devise the "perfect" search tool, more unusual services are appearing on the Web at regular intervals. Two of these services, *Firefly*[82] and Quarterdeck's *WebCompass*[83], are described below. Firefly uses a primitive artificial intelligence to customise its behaviour to each individual user, while WebCompass, again allowing a certain degree of customisation, brings search engines off the Internet and onto the desktop.

3.11.1 Firefly

Firefly is a leisure-based search engine, if search engine is the right way to describe it, which attempts to build a user-profile of each searcher in terms of their tastes in music and movies and then recommend artists, albums and movies which it thinks they may like.

It is a subscriber only service, although membership is free. Its members, described throughout as the *Firefly community*, are encouraged to converse with each other and contribute ratings and reviews to a potentially unlimited number of bands, albums and



p://www.ffly.com/~the_hedgehog/YsyEg1xSgo6:
iwjUDfzqM/Tables/MovieList.phtml?survey=1

Ratable MovieList

Ratable MovieList

Recent Pages: Movies, Home

The_Hedgehog, firefly can't recommend any movies.

But...here are some movies with the highest ratings.

-	**The Naked Kiss** (1964) **Directed by:** Samuel Fuller
*7: the best	**Star Wars** (1977) **Directed by:** George Lucas
*don't know	**The Usual Suspects** (1995) **Directed by:** James Deck, et al.
*7: the best	**Star Wars Trilogy** (1977) **Directed by:** unknown
-	**The Shawshank Redemption (Duplicate)** (1994) **Directed by:** Frank Darabont, et al.

Submit and Send More

browse Form a sentence using the menus below to ask firefly for a new list of movies. You can also use long menus for more browsing options.

List any movies that I would like.

Go firefly!

search Use **search** to find a movie or person you already have in mind. You do not have to type the complete name.

Find movie: Find it!

Read the firefly Community Policy
Copyright 1995, 1996 Firefly Network, Inc Patent Pending

Diagram 3.12 - Firefly's movie ratings

films. The more each member contributes to the Firefly database, the better picture Firefly develops of the individual's leisure tastes. Once a certain number of reviews and ratings have been supplied, Firefly then starts to recommend music and movies it believes the member may like, based on other members tastes and recommendations. Because of the personal nature of Firefly's service - each individual member is allowed his or her own "home page" on the Firefly database, and members are allowed access to each others ratings and reviews - members are encouraged to get to know each other and find people with similar tastes to obtain recommendations from.

This is a novel idea with many benefits for anyone prepared to put the time and effort into making use of it. There are clear reservations however. How accurate are Firefly's own recommendations? Recommendations start trickling through after one has supplied around twenty reviews or ratings, but is this sufficient information to enable Firefly to build an accurate picture of an individual's tastes, and if not then how many ratings would it truly require? The user-contact aspect is an interesting one too and an obvious analogy to draw here is that Firefly is a miniature version of Usenet. Posting to newsgroups has long been a popular way of contacting those with similar tastes all over the world. The big difference, of course, is that in the case of newsgroups, no computer is monitoring every recommendation and trying to build a user-profile.

Clearly there are flaws with Firefly then, but it will be interesting to see how the service develops. Naturally the more members it acquires, the more comprehensive its database will become. Another service worth mentioning here is the *Similarities Engine*[84]. This database was assembled from a survey of around 18000 Internet users which asked them to list their favourite music and artists and suggest other artists whom they regard as similar. The result is a database of bands, musicians and composers from all areas of the musical spectrum. Each record in the database, representing a single artist, contains a list of artists (with hypertext links) suggested to be similar, with a degree of *confidence*, from "very high" to "low", for each based on the number of people citing them as similar.

3.11.2 WebCompass

WebCompass describes itself as a knowledge manager, a phrase it goes on to define as, "a software application that automates the process of extracting and managing the knowledge users seek from online information search resources."[85].

WebCompass is a software application, developed by Quarterdeck, which can best be described as a personalised MetaCrawler. It allows one to choose the search engines with which to perform a search, from a list of over a hundred, retrieves and collates the results, creates a brief description including keywords for each resource, and relevance ranks them. Results can then be placed into numerous topic directories as created by the user, and stored for future reference. This feature has the consequent effect of allowing a user to search their previous searches.

This service, along with *JavaCrawler* as mentioned in 3.3.1, moves search engines into the realm of software applications rather than Internet resources to be accessed online - another interesting development, of which the latest is AutoNomy's *Agentware* product. Agentware is not due for release until October 1996, although beta versions are currently downloadable[86]. This uses artificially intelligent Web robots to trawl the Web for specific information tailored to particular preferences as stated by the searcher. It can perform this function even when its home computer is switched off, and indexes the pages retrieved for future reference.

REFERENCES FOR CHAPTER THREE

1. *Yahoo!* URL: http://www.yahoo.com/. 2 Sep 1996.
2. **Lofthouse Gareth.** *Downloading the dollar.* In: *The Web*, **7**, July/August 1996, pp.30-33.
3. *Yahoo! History.* URL: http://www.yahoo.com/docs/pr/history.html. 2 Sep 1996.
4. *Yahoo! Press Releases.* URL: http://www.yahoo.com/docs/pr/releaseindex.html. 2 Sep 1996.
5. **Web21.** *100 hot websites.* URL: http://www.hot100.com/. 2 Sep 1996.
6. **Courtois, Martin P., William M. Baer & Marcella Stark.** *Cool tools for searching the Web: a performance evaluation.* In: *Online*, **19** (6), November/December 1995, pp. 14-32.
7. **Courtois, Martin P.** *Cool tools for Web searching: an update.* In: *Online*, **20** (3), May/June 1996, pp. 29-36.
8. *Yahoo search.* URL: http://www.yahoo.com/bin/search. 2 Sep 1996.

9. *DejaNews - the source for Internet newsgroups.* URL: http://www.dejanews.com/. 2 Sep 1996.

10. *Four11 Directory Services.* URL: http://www.four11.com/. 2 Sep 1996.

11. *Yahoo! People Search.* URL: http://www.yahoo.com/search/people/. 2 Sep 1996.

12. *Yahoo! entertainment: people.* URL: http://www.yahoo.com/Entertainment/People/. 2 Sep 1996.

13. *Hytelnet information page.* URL: http://www.lights.com/hytelnet/. 2 Sep 1996.

14. *Image Surfer category list.* URL: http://ipix.yahoo.com/. 2 Sep 1996.

15. *BUBL Information Service Web Server.* URL: http://www.bubl.bath.ac.uk/BUBL/cattree.html. 2 Sep 1996.

16. *BUBL-LINK: Libraries of Networked Knowledge.* URL: http://catriona.lib.strath.ac.uk/. 2 Sep 1996.

17. *The World-Wide Web Virtual Library: subject catalogue.* URL: http://www.w3.org/pub/DataSources/bySubject/Overview.html. 2 Sep 1996.

18. *The Argus Clearinghouse.* URL: http://www.clearinghouse.net/. 2 Sep 1996.

19. *AltaVista: main page.* URL: http://www.altavista.digital.com/. 2 Sep 1996.

20. **Tomaiuolo, Nicholas G. & Joan G. Packer.** *Quantitative analysis of five WWW search engines.* In: *Computers In Libraries,* **16** (6), June 1996. URL: http://neal.ctstateu.edu:2001/htdocs/websearch.html. 2 Sep 1996.

21. *Excite home.* URL: http://www.excite.com/. 2 Sep 1996.

22. *Welcome to Lycos.* URL: http://www.lycos.com/. 2 Sep 1996.

23. *Lycos database of URLs jumps to 51 million.* In: *For immediate release.* URL: http:// www.lycos.com/press/51million.html. 2 Sep 1996.

24. *Inktomi: counting documents.* URL: http://inktomi.berkeley.edu/counting.html. 18 July 1996.

25. *A2Z home page.* URL: http://a2z.lycos.com/. 2 Sep 1996.

26. **Stanley, Tracey.** *AltaVista vs. Lycos.* In: *Ariadne,* Issue 2. URL: http://ukoln.bath.ac.uk/ariadne/issue2/engines/. 2 Sep 1996.

27. *Welcome to Open Text Corporation.* URL: http://www.opentext.com/. 2 Sep 1996.

28. **Zorn, Peggy,** et al. *Advanced Web searching: tricks of the trade.* In: *Online,* **20** (3), May/June 1996, p.27.

29. *ALIWEB search form.* URL: http://web.nexor.co.uk/public/aliweb/search/doc/form.html. 2 Sep 1996.

30. *HotBot.* URL: http://www.hotbot.com/. 6 Sep 1996.

31. *MetaCrawler searching.* URL: http://metacrawler.cs.washington.edu/. 2 Sep 1996.

32. **Selberg, Erik & Oren Etzioni.** *Multi-service search and comparison using the MetaCrawler.* In: *Proceedings of the 4th International WWW Conference.* URL: http:// metacrawler.cs.washington.edu:8080/papers/www4/html/Overview.html. 2 Sep 1996.

33. *MetaCrawler Java Beta.* URL: http://beta.metacrawler.com/. 2 Sep 96.

34. *SavvySearch.* URL: http://www.cs.colostate.edu/~dreiling/smartform.html. 2 Sep 1996.

35. *The Internet Softbot.* URL: http://www.cs.washington.edu/research/softbots/. 2 Sep 1996.

36. *All-in-One Search Page.* URL: http://www.albany.net/allinone/. 2 Sep 1996.

37. **Brown, Peter.** *Turning ideas into products: the Guide system.* In: *Proceedings of Hypertext '87.* University of North Carolina, pp. 33-40.

38. *Yahoo! - Computers and Internet:Internet:World Wide Web:Searching the Web:All-in-One Search Pages.* URL: http://www.yahoo.com/Computers_and_Internet/ Internet/World_Wide_Web/Searching_the_Web/All_in_One_Search_Pages/. 2 Sep 1996.

39. *SEARCH.COM.* URL: http://www.search.com/. 2 Sep 1996.

40. *SHAREWARE.COM - the way to find shareware on the Internet.* URL: http://www. shareware.com/. 2 Sep 1996.

41. *SBA shareware library.* URL: http://www.sbaonline.sba.gov/shareware/index.html. 2 Sep 1996.

42. *Search the Internet with the Internet Sleuth.* URL: http://www.isleuth.com/. 2 Sep 1996.

43. **Web21.** *100 hot websites.* URL: http://www.hot100.com/. 2 Sep 1996.

44. *Internet Movie Database search.* URL: http://uk.imdb.com/search.html. 2 Sep 1996.

45. *Computing dictionary.* URL: http://wombat.doc.ic.ac.uk/. 2 Sep 1996.

46. *World Wide Yellow Pages.* URL: http://www.yellow.com/. 2 Sep 1996.

47. *The Global On-Line Directory.* URL: http://www.gold.net/gold/search2.html. 2 Sep 1996.

48. *QBIC home page.* URL: http:// wwwqbic.almaden.ibm.com/~qbic/qbic.html. 2 Sep 1996.

49. *AC/DC: The ACademic DireCtory.* URL: http://acdc.hensa.ac.uk/. 2 Sep 1996.

50. **Beckett, Dave & Neil G Smith.** The ACademic DireCtory - AC/DC. In: Ariadne, Issue 3. URL: http://ukoln.bath.ac.uk/ariadne/issue3/acdc/. 2 Sep 1996.

51. *Frequently Asked Questions (and Answers) about Harvest.* URL: http://newbruno. cs.colorado.edu/harvest/FAQ.html. 2 Sep 1996.

52. *The comprehensive index of UK Internet sites.* URL: http://www.ukindex.co.uk/. 2 Sep 1996.
53. *UK directory - welcome!* URL: http://www.ukdirectory.com/. 2 Sep 1996.
54. *UK Internet WWW.* URL: http://www.internetweb.co.uk/. 5 Jun 1996.
55. *Yahoo! JAPAN.* URL: http://www.yahoo.co.jp/. 2 Sep 1996.
56. *Yahoo! Canada.* URL: http://www.yahoo.ca/. 2 Sep 1996.
57. *AID - Austrian Internet Directory.* URL: http://www.aid.co.at/aid/. 2 Sep 1996.
58. *Flipper home page.* URL: http://flp.cs.tu-berlin.de/flipper/. 2 Sep 1996.
59. *GoGREECE.com: the Internet guide to Greece.* URL: http://www.gogreece.com/. 2 Sep 1996.
60. *Recursos de Internet en Español y Portugués.* URL: http://www.ogilvy.com/spanish/hisplink.htm. 2 Sep 1996.
61. *Swiss Search.* URL: http://www.search.ch/. 2 Sep 1996.
62. *WebRider - Belgian search engine.* URL: http://www.webrider.be/. 2 Sep 1996.
63. *Infoseek Professional home page.* URL: http://professional.infoseek.com/. 2 Sep 1996.
64. *Infoseek Guide.* URL: http://www.infoseek.com/. 2 Sep 1996.
65. *NlightN home page!* URL: http://www.nlightn.com/. 2 Sep 1996.
66. *infoMarket search page.* URL: http://www.infomkt.ibm.com/. 2 Sep 1996.
67. *Welcome to the Electric Library.* URL: http://www.elibrary.com/. 2 Sep 1996.
68. *Welcome to Magellan!* URL: http://www.mckinley.com/. 2 Sep 1996.
69. **Courtois, Martin P.** *Cool tools for Web searching: an update.* In: *Online*, **20** (3), May/June 1996, p. 30.
70. *McKinley Press Information.* URL: http://www.mckinley.com/feature.cgi?pressroom2_bd. 2 Sep 1996.
71. *Magellan People Finder.* URL: http://www.infospace.com/mage/index.html. 2 Sep 1996.
72. *WebCrawler searching.* URL: http://webcrawler.com/. 2 Sep 1996
73. *A2Z home page.* URL: http://a2z.lycos.com/. 2 Sep 1996.
74. *Point: it's what you're searching for.* URL: http://point.lycos.com/. 2 Sep 1996
75. *Net reviews - search.* URL: http://www.iguide.com/search/insites.sml. 2 Sep 1996.
76. *Welcome to JumpCity.* URL: http://www.jumpcity.com/. 2 Sep 1996.
77. *CyberHound.* URL: http://www.thomson.com/cyberhound/default.html. 2 Sep 1996.
78. **Nicholson, Dennis & Joanne Gold.** *LINK: a new beginning for BUBL.* In: *Ariadne*, Issue 3. URL: http://ukoln.bath.ac.uk/ariadne/issue3/bubl/. 2 Sep 1996.

79. *Ameritech Library Services' NetPublisher demonstration server.* URL: http://netpub.notis.com/. 2 Sep 1996.

80. *NetFirst information.* URL: http://www.oclc.org/oclc/netfirst/. 2 Sep 1996.

81. *Building a Catalog of Internet-Accessible Materials.* URL: http://www.oclc.org/oclc/man/catproj/overview.htm. 2 Sep 1996.

82. *Firefly.* URL: http://www.ffly.com/. 2 Sep 1996.

83. *WebCompass 2.0; it's coming.* URL: http://arachnid.qdeck.com/qdeck/products/webcompass/. 2 Sep 1996.

84. *The Similarities Engine.* URL: http://www.ari.net/se/. 2 Sep 1996.

85. *WebCompass 2.0; White Paper - WebCompass 2.0 Overview.* In: *WebCompass 2.0; it's coming.* URL: http://arachnid.qdeck.com/qdeck/products/webcompass/wc20.over.html. 2 Sep 1996.

86. *Autonomy Corporation.* URL: http://www.agentware.com/. 2 Sep 1996.

Chapter Four - Is there a Strategy? Conclusions and Findings

It has become clear during the course of this document, that if an expert search strategy exists for the World Wide Web, then it is going to rely heavily on the use of search engines. It has also become clear that given the diverse range of information available, especially given the Web's capacity as a multimedia resource, that any strategy will have to be very general and dependent upon the information sought. Precise pinpointing of which particular service to use is next to impossible. There are, however, reasonable guidelines, and these have emerged over the course of this project.

4.1 Locating Web Sources

4.1.1 Find a Good Starting Point: Meta-search Facilities

Given the large number of search engines available, it is difficult to know which is the best for the particular piece of information being sought. It is suggested, firstly, to use one of the meta-search engines, such as MetaCrawler or SavvySearch, as described in 3.3, for the initial query. These cover the best (ie. largest, most popular and comprehensive) databases available and searching all of them simultaneously will give the searcher a general idea of how easily the information is going to be found.

Use MetaCrawler (3.3) for Web resources since it covers the larger, more general databases. SavvySearch (3.3.4) uses a broader mix of search services and it is suggested that this be used if FTP and gopher resources also need to be searched.

If it is true to say that information obtained from another country would be irrelevant (for instance, if one is looking for legal information), then use MetaCrawler's ability to limit a search by region or domain.

4.1.2 Limit the Search: Advanced Query Syntax

Having used the meta-search facilities, the searcher will have a good idea of the abundance of the sort of information he or she is looking for on the Web, and how easy it will be to find. It is now time for them to add precision to their search by acquainting themselves with the advanced query syntax of Web search engines described, in particular, in 3.2.

By using some of the advanced search options, result lists can be narrowed considerably. This process is slightly more arduous than it needs to be due to the fact that every major service uses slightly different syntax, but it is certainly worth learning the quirks of each of the major services.

It is recommended that AltaVista (3.2), Excite (3.2.4 and 3.9.4), HotBot (3.2.4), Infoseek (2.2.1.5), Lycos (3.2.4), Magellan (3.9), MetaCrawler (3.3), Open Text (3.2.4), SavvySearch (3.3.4), WebCrawler (3.9.4), and Yahoo (3.1) all be learned, and any others beside these would also be of benefit. This may seem like hard work, but usually the differences between each are only minor.

4.1.3 Try a More Specialised Database

Having tried a meta-search engine and advanced queries on some of the larger databases, as described above (4.1.1 and 4.1.2), one should see if there exists a more specialised vertical database (2.2.1.4 and 3.6).

To find a more specialised database, use one of the, "search engines of search engines," Search.Com (3.5) and the Internet Sleuth (3.5.4).

If Search.Com and the Internet Sleuth fail to produce a relevant database, then return to the major search engines (4.1.2) and check the results list to see if any of the hits returned point to other information sources rather than specific information *per se*. If this still fails then broaden the search slightly

As mentioned in 4.1.1, if it is true to say that information obtained from another country would be irrelevant then look at a geographically-specific search engine instead of one of the major global search services. Presently there are few services available comprehensive enough to be used for this purpose, but for the time being, the best UK services are described in 3.7.4.

4.1.4 Consider Whether to Use a Subscription Service

If none of the above leads to relevant sources then it is definitely worth looking at some of the subscription services such as Infoseek Professional and NlightN (3.8) to see if any of the databases offered match requirements. The question of whether they are preferable to existing commercial online databases is an issue, but unfortunately this has not been covered in sufficient depth here for any recommendations to be made in this respect.

4.2 Locating Other Internet Sources

In the case of Internet protocols other than http, as described in 1.3, it is still better to use dedicated databases rather than the major search engines. The suggested resources for each particular protocol follow.

4.2.1 FTP and Gopher

Archie (1.3.1) and Veronica (1.3.3) may seem archaic by Web standards but they are still essential Internet resources for their respective protocols, and it is suggested that they be used. Some, mainly academic, services such as BUBL (3.1.5) and LINK (3.10) index documents on subject regardless of Internet protocol in any case.

It is suggested that retrieving FTP and Gopher resources may no longer be an issue in the future as the more useful documents and files will be converted to http.

4.2.2 Newsgroups

It is recommended that DejaNews (3.1.3), AltaVista (3.2), Excite (3.2.4) and Infoseek (2.2.1.5), be used when searching newsgroups.

4.2.3 E-mail addresses

There are many dedicated services for locating people. Perhaps the largest is the Four11 Directory (3.1.3), although this is far from comprehensive. It is suggested that Search.Com (3.5) or the Internet Sleuth (3.5.4) be used to locate and search the many dedicated "people" databases. At the time of writing it will probably be necessary to search many before a particular e-mail address is found.

4.2.4 Telnet

Hytelnet (3.1.3) is the outstanding resource for locating telnet sites, and particularly library catalogues.

4.3 Locating Multimedia Sources

When searching for images, try the Yahoo! Image Surfer (3.1.4) or Query By Image Content (2.2.1.4 and 3.6) initially.

If this fails, or the search needs to be more specific, use a keyword search on a major search engine (4.1.2) and scan the results list for pages that seem likely to contain images.

Features such as AltaVista's *"image:"* field search (3.2.2) are novel but not very effective since the exact file name of an image must be known, but are worth using as a last resort. A final option would be to add *"AND img"* to a Boolean search, although this will only be effective if the search engine indexes hidden HTML tags.

Video and sound archives are less abundant on the Internet at present. Although search options are available, Yahoo's subject categories (3.1.4) are still probably the best place to find those that exist.

Many software archives are available and always worth being aware of, but locating software files is not such a problem given the tendency to be familiar with exact file names. It is suggested that the searcher try a keyword search on a major search engine (4.1.2), which usually proves effective for locating software.

4.4 Assessing Quality: Reviews and Records

It is assumed that the searcher has now found his or her information source as a result of following one of the procedures described above (4.1, 4.2 and 4.3). It is now time to confirm the quality and authority of the document retrieved. This is done by using a search engine or directory which provides reviews of documents, such as Magellan (3.9) and Excite (3.9.4).

It may be the case that the source has already been located in one of these databases. If this is the case then the user will already have a good idea about the content of the article. Presently, however, few of these are as good as Yahoo! (3.1) for actually *finding* information. It is suggested, therefore, that documents be found using the above procedures (4.1, 4.2 and 4.3) and quality then assessed by keyword searching the review databases once a document title is already known. The more of these that are used then the better idea one will receive regarding content. Some of the better services have been described in 3.9.4.

If this procedure fails, then a final resort is to use newsgroups. Somewhere on the Internet is an expert already acquainted with the better sources. If the searcher has failed to gain any information regarding the quality of a source from the existing services, it is suggested that a polite question be posted to an appropriate newsgroup. There are resources available on the Web for locating newsgroups, but a keyword query on a similar topic in a newsgroup search engine (see 4.2.2) should point the searcher in the direction of a relevant newsgroup.

It is difficult, at present, to make recommendations about databases that use catalogue records (2.2.1.6 and 3.10), as Web services are still relatively underdeveloped in this respect. In a year's time the picture will be much clearer but it looks to be the case that these services will tend to concentrate on academic documents. At the moment, existing online databases are better for this, but it is definitely an area the searcher should keep his or her eye on.

4.5 Customising a Search Page

After using the previously described strategy a few times, the searcher will have a good idea of the Web search sources that he or she is using most frequently. It has already been seen that multi-engine search pages such as the All-in-One Search Page (3.4) can be a very useful time-saving device in the pursuit of information retrieval on the Internet. However, no single one of these can be recommended as the best, since the requirements of each individual information professional may be very different. It is also fair to say that most of the existing multi-engine search pages have only simple keyword search forms containing none of the advanced search features available on many search engines, and this is not a practice to be encouraged by the expert searcher (see 4.1.2). Far better than use one of the existing ones is to customise one's own based on the services most frequently used. It is not even necessary to have space on a network for this purpose, since any HTML document can be stored on the hard disk and viewed on a Web browser, usually from an option in the *file* menu.

4.6 Staying Aware of New Developments

Keeping abreast of each service is also essential. It is tempting to become accustomed to one service only and use it all the time because the search interface seems easier to use or a series of successful results have been obtained. Most search engine providers update their service regularly and make improvements, and keeping on top of these developments is a must. A poor interface one week may become the best interface the following week, so keeping an open mind about which is the favoured service is very important.

An important criterion in the choice of service can very often be currency, and in many cases the right database to use is the one that has been updated yesterday. It is very difficult to tell which of the major search services are updated most frequently. Although some research has been done on the currency of Web databases (see 3.2.4) on present evidence it seems likely that there will become a core of databases that are updated on a daily basis.

4.7 Final Comments

Standardisation of subject classification across the Web would be extreme helpful as far as the information professional is concerned, but this does not seem to be very likely. Increased use of meta tags (2.2.1.6), both by Webmasters and search engines is a suitable alternative, although an issue here is whether a Webmaster is qualified to apply his or her own subject classification to each document. Many would say that this process should be carried out by a neutral party (see 2.2.3). In addition, while meta tags are useful when the subject content of a document is clear, they are not as easy to apply in the case of individual home pages where content may be wide-ranging. Yahoo! side-steps the issue of classifying the latter by indexing pages by an individual's name (3.1.3).

Finally, it is always worth keeping an eye on future developments. New search technology is appearing all the time, and these guidelines will only suffice for the present. As new ideas are formulated for the purpose of information retrieval on the Internet, the information professional must be prepared to look at each one, assess its potential for making their job easier, and update their search strategy accordingly.

BIBLIOGRAPHY

AC/DC: The ACademic DireCtory. URL: http://acdc.hensa.ac.uk/. 2 Sep 1996.

Access to network resources projects. URL: http://ukoln.bath.ac.uk/elib/lists/anr.html. 2 Sep 1996.

The ADAM project. URL: http://adam.ac.uk/. 2 Sep 1996.

AID - Austrian Internet Directory. URL: http://www.aid.co.at/aid/. 2 Sep 1996.

ALIWEB search form. URL: http://web.nexor.co.uk/public/aliweb/search/doc/form.html. 2 Sep 1996.

All-in-One Search Page. URL: http://www.albany.net/allinone/. 2 Sep 1996.

AltaVista: main page. URL: http://www.altavista.digital.com/. 2 Sep 1996.

Amateur hardcore search engine. URL: http://www.amateurs.com/searchex.htm. 2 Sep 1996.

Ameritech Library Services' NetPublisher demonstration server. URL: http://netpub.notis.com/. 2 Sep 1996.

The Argus Clearinghouse. URL: http://www.clearinghouse.net/. 2 Sep 1996.

Autonomy Corporation. URL: http://www.agentware.com/. 2 Sep 1996.

A2Z home page. URL: http://a2z.lycos.com/. 2 Sep 1996.

Beckett, Dave & Neil G Smith. The ACademic DireCtory - AC/DC. In: Ariadne, Issue 3. URL: http://ukoln.bath.ac.uk/ariadne/issue3/acdc/. 2 Sep 1996.

Berners-Lee, Tim, 1989. *Information Management: A Proposal.* URL: http://www.w3. org/hypertext/WWW/History/1989/proposal.html. 2 Sep 1996.

Brown, Peter. *Turning ideas into products: the Guide system.* In: *Proceedings of Hypertext '87.* University of North Carolina, pp. 33-40.

BUBL Information Service Web Server. URL: http://www.bubl.bath.ac.uk/BUBL/home.html. 2 Sep 1996.

BUBL-LINK: Libraries of Networked Knowledge. URL: http://catriona.lib.strath.ac.uk/. 2 Sep 1996.

Building a Catalog of Internet-Accessible Materials. URL: http://www.oclc.org/oclc/man/catproj/overview.htm. 2 Sep 1996.

CATRIONA. URL: http://www.bubl.bath.ac.uk/BUBL/catriona.html. 2 Sep 1996.

The comprehensive index of UK Internet sites. URL: http://www.ukindex.co.uk/. 2 Sep 1996.

Computing dictionary. URL: http://wombat.doc.ic.ac.uk/. 2 Sep 1996.

Counting URLs. URL: http://www.excite.com/ice/counting.html. 2 Sep 1996.

Courtois, Martin P., William M. Baer & Marcella Stark. *Cool tools for searching the Web: a performance evaluation.* In: *Online*, **19** (6), November/December 1995, pp. 14-32.

Courtois, Martin P. *Cool tools for Web searching: an update.* In: *Online,* **20** (3), May/June 1996, p. 30.

CUI W3 Catalog. URL: http://cuiwww.unige.ch/w3catalog/. 2 Sep 1996.

CyberHound. URL: http://www.thomson.com/cyberhound/default.html. 2 Sep 1996.

De Bra, P.M. & R.D.J. Post, 1994. *Information retrieval in the World Wide Web.* In: *Computer Networks and ISDN Systems*, **27**, pp.183-192.

DejaNews - the source for Internet newsgroups. URL: http://www.dejanews.com/. 2

Sep 1996.

EFFweb - the Electronic Frontier Foundation. URL: http://www.eff.org/. 2 Sep 1996.

Elm,W. & D.Woods, 1985. *Getting lost: a case study in interface design.* In: *Proceedings of the Human Factors Society 29th Annual Meeting.* Santa Monica, CA: Human Factors Society, pp.927-931.

Engine sells results, draws fire. URL: http://www.cnet.com/Content/News/Files/0,16,1635,00.html. 2 Sep 1996.

Excite and the McKinley Group Sign Letter of Intent to Merge. In: *McKinley Press Information.* URL: http://www.mckinley.com/feature.cgi?pressroom2_bd. 2 Sep 1996.

Excite home. URL: http://www.excite.com/. 2 Sep 1996.

Firefly. URL: http://www.ffly.com/. 2 Sep 1996.

First WWW conference. URL: http://www1.cern.ch/WWW94/Welcome.html. 2 Sep 1996.

Fish-Search form used on www.win.tue.nl. URL: http://www.win.tue.nl/bin/fishsearch/. 2 Sep 1996.

Flipper home page. URL: http://flp.cs.tu-berlin.de/flipper/. 2 Sep 1996.

Four11 Directory Services. URL: http://www.four11.com/. 2 Sep 1996.

Frequently Asked Questions (and Answers) about Harvest. URL: http://newbruno.cs.colorado.edu/harvest/FAQ.html. 2 Sep 1996.

The Global On-Line Directory. URL: http://www.gold.net/gold/search2.html. 2 Sep 1996.

GoGREECE.com: the Internet guide to Greece. URL: http://www.gogreece.com/. 2 Sep 1996.

Help/Frequently Asked Questions. URL: http://www.nlightn.com/help/help.htm. 2 Sep 1996.

Highways Agency home page. URL: http://www.open.gov.uk/hiagency/highhome.htm. 2 Sep 1996.

HotBot. URL: http://www.hotbot.com/. 6 Sep 1996.

The HUMBUL Gateway. URL: http://sable.ox.ac.uk/departments/humanities/international.html. 2 Sep 1996.

Hytelnet information page. URL: http://www.lights.com/hytelnet/. 2 Sep 1996.

Image Surfer category list. URL: http://ipix.yahoo.com/. 2 Sep 1996.

infoMarket search page. URL: http://www.infomkt.ibm.com/. 2 Sep 1996.

Infoseek Guide. URL: http://www.infoseek.com/. 2 Sep 1996.

Infoseek Professional home page. URL: http://professional.infoseek.com/. 2 Sep 1996.

Infoville Schoolhouse: WAIS. URL: http://canyon.ucsd.edu/infoville/schoolhouse/wais.html. 2 Sep 1996.

Inktomi: counting documents. URL: http://inktomi.berkeley.edu/counting.html. 18 July 1996.

Internet Movie Database search. URL: http://uk.imdb.com/search.html. 2 Sep 1996.

The Internet Services List. URL: http://www.spectracom.com/islist/. 2 Sep 1996.

The Internet Softbot. URL: http://www.cs.washington.edu/research/softbots/. 2 Sep 1996.

Internet Society home page. URL: http://www.isoc.org/. 2 Sep 1996.

Kent, Robert E., & Christian Neuss, 1995. *Creating a Web analysis and visualisation environment.* In: *Computer Networks and ISDN Systems,* **28**, pp.109-117.

Lofthouse Gareth. *Downloading the dollar.* In: *The Web*, **7**, July/August 1996, pp.30-33.

Lycos database of URLs jumps to 51 million. In: *For immediate release.* URL: http:// www.lycos.com/press/51million.html. 2 Sep 1996.

Lycos, Inc. - info. URL: http://point.lycos.com/faq/. 2 Sep 1996.

Magellan People Finder. URL: http://www.infospace.com/mage/index.html. 2 Sep 1996.

Maximized online search engine study: introduction. URL: http://maxonline.com/ searchstudy/. 2 Sep 1996.

McKinley Press Information. URL: http://www.mckinley.com/feature.cgi?pressroom 2_bd. 2 Sep 1996.

McKnight, Cliff, Andrew Dillon & John Richardson, 1991. *Hypertext in context.* Cambridge: CUP, p.69.

McKnight, Cliff, Andrew Dillon & John Richardson, 1993. *Space - the final chapter, or why physical representations are not semantic intentions.* In: *McKnight,C, A.Dillon & J.Richardson, 1993. Hypertext: a psychological perspective.* New York: Ellis Horwood, p.170.

MetaCrawler Java Beta. URL: http://beta.metacrawler.com/. 2 Sep 96.

MetaCrawler searching. URL: http://metacrawler.cs.washington.edu/. 2 Sep 1996.

The META tag: controlling how your Web page is indexed by AltaVista. URL: http://www.altavista.digital.com/cgi/bin/query?pg=ah&what=web#meta. 23 Aug 1996.

Nelson, Ted, 1986. *A technical overview of the Xanadu electronic storage and* *publishing system* [Videocassette]. Texas: Fredericksburg.

NetFirst information. URL: http://www.oclc.org/oclc/netfirst/. 2 Sep 1996.

Net reviews - search. URL: http://www.iguide.com/search/insites.sml. 2 Sep 1996.

The New York Times on the Web. URL: http://www.nytimes.com/. 2 Sep 1996.

Nicholson, Dennis & Joanne Gold. *LINK: a new beginning for BUBL.* In: *Ariadne*, Issue 3. URL: http://ukoln.bath.ac.uk/ariadne/issue3/bubl/. 2 Sep 1996.

NlightN home page! URL: http://www.nlightn.com/. 2 Sep 1996.

OMNI welcome page. URL: http://omni.ac.uk/. 2 Sep 1996.

People involved with the World Wide Web Consortium. URL: http://www.w3.org/pub/WWW/People/W3Cpeople.html#Berners-Lee. 2 Sep 1996.

People who have contributed to the World Wide Web project. URL: http://www.w3.org/pub/WWW/People.html. 2 Sep 1996.

Point: it's what you're searching for. URL: http://point.lycos.com/. 2 Sep 1996

Poulter, Alan. *Web search engines: a critical review.* In: *Program* (in press).

QBIC home page. URL: http:// wwwqbic.almaden.ibm.com/~qbic/qbic.html. 2 Sep 1996.

Rankin, Bob. *Beginners' [guide to the Internet].* In: *The Web*, 6. May/June 1996, pp.52-53. Macclesfield: IDG Media. URL: http://www.wcentral.co.uk/listings/beginners/index.html, 21 May 1996.

Recursos de Internet en Español y Portugués· URL: http://www.ogilvy.com/spanish/hisplink.htm. 2 Sep 1996.

Reuters - the business of information. URL: http://www.reuters.com/. 2 Sep 1996.

SavvySearch. URL: http://www.cs.colostate.edu/~dreiling/smartform.html. 2 Sep 1996.

SBA shareware library. URL: http://www.sbaonline.sba.gov/shareware/index.html. 2 Sep 1996.

SEARCH.COM. URL: http://www.search.com/. 2 Sep 1996.

Search the Internet with the Internet Sleuth. URL: http://www.isleuth.com/. 2 Sep 1996.

Segal, Ben M. 1995. *A short history of Internet protocols at CERN.* URL: http://wwwcn. cern.ch/pdp/ns/ben/TCPHIST.html. 2 Sep 1996.

Selberg, Erik & Oren Etzioni. *Multi-service search and comparison using the MetaCrawler.* In: *Proceedings of the 4th International WWW Conference.* URL: http://metacrawler.cs.washington.edu:8080/papers/www4/html/Overview.html. 2 Sep 1996.

SHAREWARE.COM - the way to find shareware on the Internet. URL: http://www. shareware.com/. 2 Sep 1996.

The Similarities Engine. URL: http://www.ari.net/se/. 2 Sep 1996.

Social Science Information Gateway - SOSIG. URL: http://sosig.ac.uk/. 2 Sep 1996.

Stanley, Tracey. *AltaVista vs. Lycos.* In: *Ariadne*, Issue 2. URL: http://ukoln. bath.ac.uk/ariadne/issue2/engines/. 2 Sep 1996.

Swiss Search. URL: http://www.search.ch/. 2 Sep 1996.

tile.net. URL: http://tile.net/. 2 Sep 1996.

Tomaiuolo, Nicholas G. & Joan G. Packer. *Quantitative analysis of five WWW search engines.* In: *Computers In Libraries*, **16** (6), June 1996. URL: http://neal. ctstateu.edu:2001/htdocs/websearch.html. 2 Sep 1996.

Tseng, Gwyneth, Alan Poulter & Debra Hiom, 1996. *The library and information professional's guide to the Internet.* London: Library Association. p.91.

UK directory - welcome! URL: http://www.ukdirectory.com/. 2 Sep 1996.

UK Internet WWW. URL: http://www.internetweb.co.uk/. 5 Jun 1996.

UKOLN: UK Office for Library and Information Networking. URL: http://ukoln.bath.ac.uk/about.html.

The URL-minder! Your own personal Web robot. URL: http://www.netmind.com/URL-minder/URL-minder.html. 2 Sep 1996.

WebCompass 2.0; it's coming. URL: http://arachnid.qdeck.com/qdeck/products/webcompass/. 2 Sep 1996.

WebCompass 2.0; White Paper - WebCompass 2.0 Overview. In: *WebCompass 2.0; it's coming.* URL: http://arachnid.qdeck.com/qdeck/products/webcompass/wc20.over.html. 2 Sep 1996.

WebCrawler searching. URL: http://webcrawler.com/. 2 Sep 1996

The webmaster's guide to search engines and directories. URL: http://calafia.com/webmasters/. 2 Sep 1996.

WebRider - Belgian search engine. URL: http://www.webrider.be/. 2 Sep 1996.

WebWatcher home page. URL: http://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/webwatcher/. 2 Sep 1996.

Web21. *100 hot websites.* URL: http://www.hot100.com/. 2 Sep 1996.

Welcome to JumpCity. URL: http://www.jumpcity.com/. 2 Sep 1996.

Welcome to Lycos. URL: http://www.lycos.com/. 2 Sep 1996.

Welcome to Magellan! URL: http://www.mckinley.com/. 2 Sep 1996.

Welcome to Open Text Corporation. URL: http://www.opentext.com/. 2 Sep 1996.

Welcome to the Electric Library. URL: http://www.elibrary.com/. 2 Sep 1996.

WWW Consortium. *A little history of the World Wide Web.* URL: http://www.w3.org/pub/WWW/History.html. 2 Sep 1996.

World Wide Web FAQ. URL: http://info.ox.ac.uk/help/wwwfaq/index.html. 2 Sep 1996.

The World-Wide Web Virtual Library: subject catalogue. URL: http://www.w3.org/pub/ DataSources/bySubject/Overview.html. 2 Sep 1996.

World Wide Yellow Pages. URL: http://www.yellow.com/. 2 Sep 1996.

Yahoo! URL: http://www.yahoo.com/. 2 Sep 1996.

Yahoo! Canada. URL: http://www.yahoo.ca/. 2 Sep 1996.

Yahoo! - Computers and Internet:Internet:World Wide Web:Searching the Web:All-in-One Search Pages. URL: http://www.yahoo.com/Computers_and_Internet/Internet/World_Wide_Web/Searching_the_Web/All_in_One_Search_Pages/. 2 Sep 1996.

Yahoo! entertainment: people. URL: http://www.yahoo.com/Entertainment/People/. 2 Sep 1996.

Yahoo! History. URL: http://www.yahoo.com/docs/pr/history.html. 2 Sep 1996.

Yahoo! JAPAN. URL: http://www.yahoo.co.jp/. 2 Sep 1996.

Yahoo! People Search. URL: http://www.yahoo.com/search/people/. 2 Sep 1996.

Yahoo! Press Releases. URL: http://www.yahoo.com/docs/pr/releaseindex.html. 2 Sep 1996.

Yahoo search. URL: http://www.yahoo.com/bin/search. 2 Sep 1996.

Zorn, Peggy, et al. *Advanced Web searching: tricks of the trade.* In: *Online,* **20** (3), May/June 1996, p.27.

APPENDIX: Index of Search Engines Described