WEBSITE INDEXING

enhancing access to information within websites

2nd edition

Glenda Browne
Jonathan Jermey

Adelaide, South Australia: Auslib Press, 2004

FOREWORD TO FIRST EDITION

When scrolls developed into books, it made a difference to people's ways of organising, indexing and retrieving information. And, since that time, with the arrival of each new information format – such as the periodical, the online database, and most recently the Internet – it has always been necessary to rethink the application of information principles, and to devise new techniques. All too often new practices have emerged in response to new technology but without sufficient regard to the lessons of the past.

Professional indexers work within a body of knowledge which is applicable to information in all its formats, and have a responsibility to provide guidance and to develop techniques in changed information environments.

Glenda Browne and Jonathan Jermey are indexers, information professionals with a solid grounding in indexing principles and broad practical experience of indexing in electronic formats. Through their knowledge and experience, which also includes librarianship, web management and computer training, they are well placed to teach the principles and practice of web indexing.

Website indexing will make a contribution to the improvement of the quality of Internet information systems and I commend it to all who are involved in providing access and indexes to web documents — whether as indexers, librarians, web managers or information architects.

Alan Walker
President, Australian Society of Indexers, 1997–2000

FOREWORD TO THE SECOND EDITION

The first edition of this book brought together the fields of traditional (back-of-the-book) indexing with the ever-growing world wide web. Much has changed however since this book was first published.

At the time of the first edition, the web was still seen as a market opportunity. Organisations that had a website (or intranet for that matter) were seen as ahead of the game, with an important leg-up on their competition. Few organisations were even considering the importance of helping users to find information, and the concept of applying indexing approaches to sites was an unfamiliar one.

This is no longer the case. Every major organisation has a website and intranet, and users are now demanding that they be easy to use. The price of not having a usable and effective site is that customers (or staff) will simply ignore it.

This new edition is therefore timely, as it expands on the fundamentals to address recent advances in navigation design, the semantic web, wikis, XML and much more.

This is no longer just a book for the uninitiated, instead it is a valuable resource for designers of information-rich sites, as well as information architects looking to broaden their understanding of indexing techniques.

I commend Glenda Browne and Jonathan Jermey for their efforts on this new edition, and unreservedly recommend *Website Indexing* to all.

James Robertson
Managing director, Step Two Designs

FOREWORD TO THE ELECTRONIC EDITION

In converting this book to electronic format I have followed the printed 2nd Edition as closely as possible. Some errors have been corrected but no attempt has been made to update the material or deal with web links that have changed or broken. The main change in layout has been in making sure as far as possible that URLs are displayed on one line. This makes them easier to copy and helps to ensure that they are properly interpreted by Microsoft Word, which was used to produce the MS, and Adobe Acrobat, by which it was converted to PDF format.

Naturally, any problems, errors or suggestions for improvement should be reported to the authors at webindexing@optusnet.com.au.

Jon Jermey
November 2004

ACKNOWLEDGEMENTS

We dedicate this book to our families, and especially to Bill and Jenny.

This book would not have happened without Maureen Henninger's innovative idea for a web indexing course, Alan Walker's suggestion that the course notes were worthy of conversion to a book, and Madeleine Davis' enthusiasm for the project.

We are also grateful to indexing colleagues in Australia, and to those from overseas who share ideas on the Index-L discussion group, for their input.

The book discusses the indexing program HTML Indexer. Thanks to David Brown for creating this tool, for our copy, and for his support to those who use it. This edition covers the program HTML/Prep in more detail than the first edition did – many thanks to David K. Ream of Leverage Technologies for a review copy and responses to our queries.

Much of the new information in this edition was first published in *Online Currents*. Thanks to Pamela Johnstone and Elizabeth Drynan from Enterprise Information Management for providing a forum for discussion of these issues. We are also indebted to all the practitioners and scholars who take the time to write about their experiences and make these publications freely available on the web.

Finally, thanks to Carole Best for design advice and to Madeleine Davis, Carol Browne, Leigh Browne and Derrick Browne for suggestions that enhanced the text.

PREFACE

The book is based on the premise that, while it is important to provide information on the web, it is equally important to provide methods for users to retrieve that information. There are many methods of information retrieval, and a good web manager will provide more than one access tool.

The first edition of this book developed out of a one-day course on Indexing Web Documents offered at the University of NSW. It has had an international readership, including a translation into Korean by Lee Young Ja. The course has changed over the years, as new software options developed and people explored different approaches. This book has changed too, and this edition has been significantly restructured, with much additional information. It now includes sections on recent developments in XML, topic maps, the semantic web, faceted classification, visualisation techniques and automated categorisation, as well as updated information and references in all sections. It no longer contains practical exercises.

This new edition has two major sections. The first examines **back-of-book-style indexing for the web**, including a general overview, advice on the creation of book-style indexes, and a discussion of software that can be used for indexing.

The second section has expanded to examine **information access on the web in general**, including site navigation, search engines, and the semantic web. This broad approach is important because access methods such as navigation and search are alternatives to back-of-book-style indexing, or are used in conjunction with back-of-book-style indexing. In addition, many people involved in the creation of book-style indexes for the web will also be involved in the creation of metadata and thesauri. Finally, since no-one will find the information *within* a website unless they can find the website itself, the book briefly covers ways of making websites more likely to be found by searchers using search engines or directories.

The web indexing courses have attracted professional indexers, librarians, technical writers, information architects, and web managers. The book is aimed at all these people, and anyone else who wants to create an index for their website.

The book is aimed at individuals, rather than those working in large teams of specialists. It is also practical, and many of the hints come from our own experience, or that of students we have taught. The tools and techniques discussed here can be applied to eBook indexing as well as to web indexing.

You can avoid typing in the hyperlinks in this book by accessing them through our website: www.webindexing.biz.. Webpages which are no longer active can sometimes be located via the Wayback Machine at www.archive.org.

By the time you have read this book, you will know whether you want to create an index for your website, and you will be ready to start doing so. The appendix lists resources and courses to further develop your skills and knowledge.

CONTENTS

FIGURES

1. OPTIONS IN ACCESS TO INFORMATION

My great-great-aunt Alice

My great-great-aunt, Alice Browne, wrote a novel called 'That Colony of God',[1] which was published in 1923. The title and plain cover didn't attract me until a few years ago when I was planning a trip to England, and started finding out more about my family over there. I finally read the book, and found it quite appealing – an early romance, with a serious discussion of spiritualism and its compatibility with the Christian church.

I then tried to find out more about Alice. A Google search found nothing obvious, but a search using the metasearch engine Ixquick (www.ixquick.com) led me to an index to the *Occult Review 1906-1928* (www.austheos.org.au/indices/OCCREV.HTM) which noted a review of her book. The entry was stark, but adequate:

> *OccRev y1924 v39 January p63 – review – That Colony of God – A Novel by Alice M Browne – PB Ince*

I wrote in the 'From the Literature' section of the *AusSI Newsletter* (www.aussi.org/anl/2002/03april/ftl.htm) about my attempts to obtain the book review on interlibrary loan. Meanwhile, Keith Richmond, owner of the Basilisk Bookshop in Melbourne, who has an interest in the 'strange and quirky' did a Google search to re-find the *Occult Review* index which he had used previously, and found my 'From the Literature' column as the sixth hit. He emailed me to say he had the issue with the review I needed, and would send me a copy if I liked.

Without the manually created index to the *Occult Review* I wouldn't have found the review of my great-great-aunt's book, but without Ixquick I wouldn't have found the index (even Google didn't help here), and without Google Keith wouldn't have found me. The moral of this is that just as television didn't replace radio, although it changed it significantly, electronic search hasn't replaced manual indexing. We need different information access methods for different needs and documents, and we even need a variety within each category of access.

The Internet, and specifically the World Wide Web (the web), has become a major source of information for many people. Web users can nearly always find a website on any topic they are interested in. But they may not be finding the best, the most accurate, or the most useful website they can. And once having reached a website, they often struggle to find the information they need.

There are a variety of methods for finding information on the web, just as there are in a traditional library. In a library people use catalogues to search for known items by title and author, and for unknown items by title keywords and subject headings. They can browse the shelves and use the classification scheme to find books on similar topics to ones they have already found. Once they find a potentially useful book they can use its table of contents to check the overall

structure and content, and they can use its index to find information about specific subjects.

Similarly, web users have access tools with different levels of granularity and different methods of organisation. This book focuses on the specialised skill of back-of-book-style indexing and then describes the range of information retrieval methods available or being developed including:

- Categorisation using taxonomies to allow hierarchical browsing
- Search engines, ideally supplemented by metadata and thesauri
- The semantic web, including RDF, ontologies, and topic maps.

The book then discusses the best ways to make a site findable through search engines and directories.

Using these tools we should be able to fulfil '…our shared vision of providing the charts for the user of the future, to navigate both the far reaches and the minutest details of the information universe.'[2]

2. INDEXES ON THE WEB

Granularity of indexes

Known-item searches

Indexes by document type

Indexes by creator

De facto indexes

Single sourcing

Back-of-book-style indexes are used for individual websites and for documents within websites as well as for non-textual material (for example, pictures) and multimedia presentations with text and other media combined.

Nearly all back-of-book-style indexes on the web have been created by people, not computers, although a growing trend is to automatically create the index from decentralised human-created metadata. Although software to automatically generate back-of-book-style indexes is available it generally produces poor results compared with those from human indexers.

The advantages of back-of-book-style indexes for information access on the web are many. They give direct access to specific subjects of interest, and can also be browsed for a quick overview of the coverage of a site. Links to all discussions of a subject are grouped, no matter what word has been used to describe the subject, and if there are many discussions of a subject they are distinguished by subheadings. Back-of-book-style indexes show main discussions of a topic and avoid passing mentions (that is, insignificant usages of the search word). Cross-references indicate alternative places to look for information.

Indexes need to be regularly updated to be of optimal use. They can be difficult to co-ordinate for websites with many contributors, and with continually changing information. Sometimes intellectually-created indexes do not cover the whole site (for example, a newsletter might not be indexed in detail), so a search engine is a useful supplement.

Bella Hass Weinberg ('Complexity in indexing systems – abandonment and failure: Implications for organizing the Internet'. ASIS 1996 Annual Conference Proceedings October 19-24, 1996, www.asis.org/annual-96/ElectronicProceedings/weinberg.html) has looked into options for flexible indexing of the Internet as a whole, and has suggested that a group of indexers using a controlled vocabulary could make an effective Internet index using a book index structure. The index would use specific headings with coined modifications (subheadings) and would be to the level of the document or series, not to the individual piece of information as in a book index. Problems with this approach include finding funding (or charging users), keeping the index up-to-date, and maintaining consistency on this scale.

Granularity of Indexes

> **Chunk:** smallest unit of content that is used independently and needs to be indexed individually.
>
> **Granularity:** level of detail at which information is viewed or described. The more granular the tool, the smaller the chunks of information it leads to. An index linking to specific paragraphs is more granular than a table of contents or site map linking to specific pages.

Back-of-book-style indexes have great potential for providing a familiar format for access to information within websites. They can be particularly important when a site is large or complex, although this is also when they are much more difficult to implement. A website index complements the broad subject access that is available through the structural links in the site and the keyword access that might be available through a site-specific search engine.

Indexes have more entries than a table of contents so give access to more granular chunks of information.

Known-item searches

It has been claimed that an index suits searches where users know exactly what they are looking for, but that directory browsing is appropriate for more ambiguous and exploratory searches. To an indexer this limited view is extraordinary, as a good index provides a variety of synonyms as entry points, and guides users between terms using cross-references. Of course the user has to know enough to find *some* entry point into the index. Browsing requires the same or more knowledge, as the user has to know the broader categories that their vague interest belongs in. In addition, their view of the appropriate category may be different to the view presented by the taxonomy being used. For example, in the Yahoo directory (dir.yahoo.com) you look for birds under animals, but if you are interested in the activity 'Birding' you have to know to follow the trail 'Recreation > Outdoors > Birding'. To find a section on 'Topic maps' the trail is 'Business and Economy > Business to Business > Information > Records Management > Software'. This is not obvious – it requires learning and practice. This is not a criticism of the Yahoo directory structure – it is inherent in the nature of categorisation that it is not obvious to one person where another person would group a topic.

Indexes by document type

Most of the websites indexes discussed in this book are indexes to whole websites. The same principles apply, however, to indexing subsites, intranets, and online books and journals. The main differences in these cases are due to the size of the content being indexed, the users, frequency of updating, and the expected audience. Indexes to books require consideration of the way text will be provided (for example, will you access a whole chapter at a time?) and the level of

granularity of indexing (for example, will you link to individual paragraphs?). Journal indexers need to decide how much citation information to provide (for example, will the link state the volume, issue and year of the article being indexed?).

Website index examples

Jean Weber maintains indexes to the Technical Editors' Eyrie site (www.jeanweber.com/Indexer.htm) and the travel site, Avalook at Australia (www.avalook.com.au/Indexer.htm).

Site index

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

A

accessibility of Web sites
- design considerations
- editing

accessibility, editing for
accessibility, visual, example of problems
accessible website, SeniorNet and IBM pilot
acronyms, when to spell out
AddALL book search site
Adminder for tracking advertising results
Adobe Acrobat v4
- Distiller compression settings
- editing using
- PDF file creation
- setting up

Adobe Acrobat, books about
Adobe FrameMaker, books about
Adobe Photoshop
advertising policy, newsletter
advertising
- tracking results with Adminder
- your website

Alertbox, Jakob Nielsen
alternative text for images on Web sites
alternatives to the paragraph
Amazon.com, books about

Figure 1. Website index to Technical Editors' Eyrie (Reproduced with the permission of Jean Weber)

The Australian, British, and United States societies of indexers all maintain website indexes. The Australian Society of Indexers (AusSI) site is discussed in detail in 'HTML Indexer' in the 'Software' section. The American Society of Indexers (ASI, www.asindexing.org/site/asindx.htm) uses HTML Indexer to maintain their website index, and the British Society of Indexers (SI, www.socind.demon.co.uk/site/sitdex.htm or select 'Site index' at www.indexers.org.uk) uses Macrex and HTML/Prep.

The W3C (World Wide Web Consortium) maintains a website index at www.w3.org/Help/siteindex. W3C technology keywords are also listed in a mini-index on the home page.

Books

The UNIX manual (unixhelp.ed.ac.uk/index/index.html) and its index have been online for many years. Another online book with an index is Orders of Magnitude (www.informationuniverse.com/ordersmag/orders.htm).

The Murdoch University Handbook (wwwcomm.murdoch.edu.au/handbook/index.html) has an online index, with a sample entry shown below:

> Admission
>
> Deferred
> Honours
> International School Leavers
> Interstate School Leavers
> Mid-Year
> Non-School Leavers
> Postgraduate Coursework Degrees…

Simon and Schuster published a book, *Burn Rate*, without an index. Apparently the author wanted to encourage people to read the book from cover to cover, rather than look up specific topics in the index, so the book was published on paper and the index placed online (previously at www.simonsays.com/burnrate/players.html). The index is no longer there, however, showing one of the problems with this approach.

Online indexes are being used for marketing purposes. O'Reilly & Associates include indexes in their online catalogue to advertise their books (for example, www.oreilly.com/catalog/infotecture2/inx.html) and the Amazon Look Inside the Book and Search Inside the Book programs (www.amazon.com) have provided access to book indexes.

Brother Tom Murphy's site (www.brtom.org/ind.html) links to a number of indexes to works of fiction, some of which were created by students as part of their study of the literature. [3] The text is not online, so the indexes refer to page numbers within a named edition. In the case of Catcher in the Rye the index is to the 'regular edition', but a list of equivalent page number ranges for the 'anniversary edition' is given (www.geocities.com/exploring_citr/bookindex.htm). [4] This index has internal

links for cross-references within the index, and has some external links to supplementary information.

Figure 2 shows an index to the Milan Jacovich detective series from the Leverage Technologies website (click on 'index' at www.levtechinc.com/Milan/MilanHN.htm).[5] In this index the different books in the series are referred to by a two-letter abbreviation of the title.

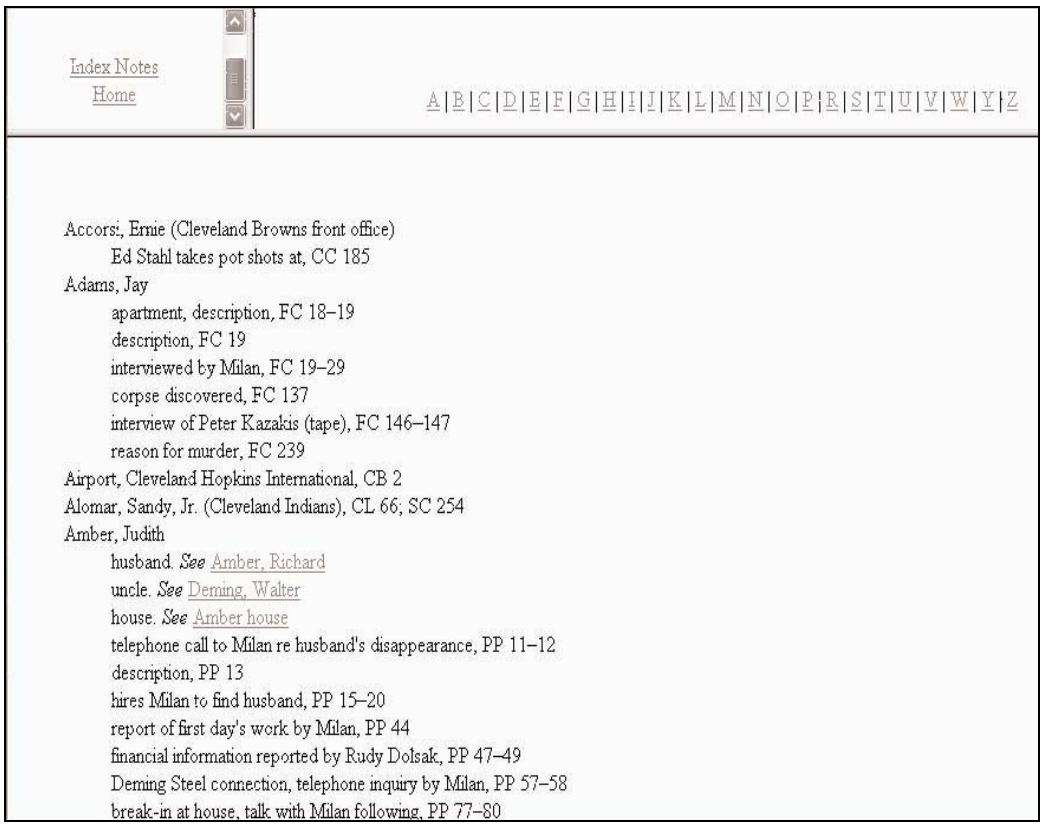

Figure 2. Framed, unlinked index to the Milan Jacovich detective story series created using HTML/Prep (Reproduced with the permission of David Ream)

Ebooks

Ebooks are standalone documents intended for on-screen reading on a PC or a handheld device, either a dedicated 'reader' or a general purpose Personal Digital Assistant (PDA).

The closest approaches to a standard eBook format are currently the OpenEBook (OEB) format, supported by Microsoft and other major corporations, and the MobiPocket format. Both essentially consist of a website compiled into a single document file, with each webpage corresponding to a chapter in the eBook. Additional information and encryption may be added to control marketing and distribution. Any hyperlinks created between the pages on the site before compilation are retained in the eBook.

For indexing purposes this means that an existing website index can be retained when the site is compiled into an eBook. For new eBooks, an index can be created

using the methods described in this book and added to the eBook at the compilation stage.

Because eBooks are inherently more stable than websites, eBook indexing resembles traditional book indexing in that it involves a series of one-off jobs rather than an ongoing effort for maintenance and review.

Online journals

Website indexes are a very practical format for provision of annual and cumulative journal indexes. See, for example, the section 'NSW Public Health Bulletin index' under 'Single sourcing'.

The index to the Los Alamos National Laboratory Research newsletter (lib-www.lanl.gov/libinfo/news/newsindx.htm) has been online since at least 1994.

Other online journal indexes include:

- Veterinary Neurology and Neurosurgery Index (www.neurovet.org/Indexer.htm)
- Psychosomatic Medicine: Journal of the American Psychosomatic Society (www.psychosomatic.org/journal_index200.html)
- Darien Newspaper Index (www.darien.lib.ct.us/news/default.htm)
- *Online Currents*: *Australia's magazine for users of online services, CD-ROMs and the Internet* (www.onlinecurrents.com.au/2003Index.html) – a print journal with an unlinked web-based index to show potential readers what the journal contains, and to allow subscribers to search the index online
- Rochester History Index (www.rochester.lib.ny.us/~rochhist/mainlist.html), which links to PDF issues of the historical journal, eg:

> A.C.Way and Sons
> location, 3(2):14 (Apr 1941)
> product line, 3(2):14 (Apr 1941)

Graphical content

Although search engines now offer image searches, these still tend to rely on associated text. Carefully constructed indexes can give useful access to online images.

The 'Rudiments of wisdom encyclopaedia' index leads to cartoons by Tim Hunkin (www.rudimentsofwisdom.com/index_atoz.htm). Sample entries include:

> AmericanWords
> Anaesthetics
> Angels
> Angling
> Animal eyes
> Animated film

Multimedia collections of slides, video, audio and transcripts recording the Online and On Disc conference have been indexed (www.cadre.com.au/showcase/olod99 – requires RealPlayer G2 plug-in). This concept originated in a project for the AVCC (Australian Vice-Chancellor's Committee) Symposium on Australian Electronic Publishing (Sydney, Australia, May 1996) and is discussed by Bob Jansen and others at www.turtlelane.com.au/TLS/EP98/EP98.html.

See also 'Index display' in the section 'Structure and style of website indexes' for information on the use of maps as indexes.

Indexes by creator

Hand-crafted indexes on the web are created by professional indexers, librarians, website designers and enthusiasts. Most of the detailed indexes discussed in this book were created by professional indexers. The section below briefly mentions alternatives to this approach.

Distributed authoring

> **Distributed authoring:** content creation by people distributed throughout an organisation, not by a centralised group of web specialists or writers. With distributed authoring there is often an expectation that subject metadata will also be created by authors. This is **distributed indexing**.

Indexes based on metadata can be automatically generated by computers, but their quality is dependent on the quality of metadata which has been added. This metadata may be created by professional indexers or website designers, but as distributed authoring becomes more common, metadata will be created by the people who create the content, at the time of content creation. This has the advantage of immediacy, but can lead to problems with consistency and user-centred access.

James Cook University uses distributed metadata creation in the automatic generation of its alphabetical index (www.jcu.edu.au/atoz). It describes the process for contributors (www.jcu.edu.au/office/itr/webmanagers/atozletter.html). The index is automatically generated from contents of the A-Z metadata tag. If a webpage doesn't have metadata it will not be included in the index. This places responsibility for index inclusion in the hands of the webpage creators. The site also provides a separate forms index which is indexed following the same principles.

Enthusiast-created indexes

***Dragon* magazine** is a monthly publication for fans of the Dungeons & Dragons game. There are a number of enthusiast-created indexes to *Dragon* available on the web. These range from Tables of Contents (called indexes) to detailed A-Z indexes. Most use general categories such as 'Player resources', 'Characters' and 'Magic etc' as general groupings, and some have true specific indexing as well. The number of enthusiast-created indexes in this area shows the need players and

readers have felt for structured access to the content of the magazine. It also shows the willingness of fans to volunteer to create indexes, and the great variety of approaches that can be taken. Some of the Dragon indexes are listed below:

- Dragon index (crpp0001.uqtr.uquebec.ca/www_wanderer/Index/Dragonindex.html)
- Dragondex: a complete index to Dragon Magazine (www.aeolia.net/dragondex). This complex index uses subheadings and provides details of title, author, issue, page number, and system (the version of Dungeons & Dragons it applies to)
- Tholos: Dragon Magazine Article Index (home.earthlink.net/~dmindex). This is a framed index. D&D topics (for example, 'Adventures', 'Campaign Design') are listed at the left hand side of the page; results (title, issue and page number) are presented in a table
- Vardania: Dragon Magazine index (www.math.utah.edu/~calfeld/adnd/dragon.html . Entries are grouped by category and also have keywords
- Zup's Dragon Magazine index (www.geocities.com/thezup/dragon/dmindex.html).

De facto indexes

Sometimes tools such as glossaries stand in as de facto indexes. If a site has a glossary that links to locations throughout a site, it can perform much of the job of an index. In this case the glossary can be enhanced with references to related terms to get the best possible value from it. I have observed people using a glossary as a de facto index in an online help system which also had an index – the glossary served as a simple entry tool for people with general enquiries, and was complemented by the index for more in-depth or complex queries.

The UKOnline A to Z of government (select 'A-Z of government at www.ukonline.gov.uk/Home/Homepage/fs/en) appears to function as a mixed index and glossary. Each agency has a name, a brief description of its role, and a link, eg:

Letter selected A

Adjudicators Office

Investigates complaints about the Inland Revenue.

Adjudicators Office

Adult Learning Inspectorate

Reports on the quality of education and training received by adult learners and young people in England.

Adult Learning Inspectorate

Single sourcing

> **Single sourcing:** using one content repository to generate documents in different formats. The content only needs to be written and maintained in one place, but can be output in formats such as HTML and RTF (rich text format) as required. Also known as multi-purposing. Repurposing refers to the sequential output of content in different formats using different software tools.

Ideally an indexing system will allow an indexer to create content once and output it in a number of formats. This is known as single sourcing. HTML Indexer, for instance, creates HTML back-of-book-style indexes, and can also output HTML Help and JavaHelp indexes. The indexing is stored as metadata in the webpages themselves, enabling multiple output formats.

NRMA online help single sourcing

NRMA Insurance (a division of IAG) uses a proprietary authoring package based on SGML/XML that allows output in RTF, HTML and online help formats. This means that the writing and indexing are only done once, but the information can be published in print form, on the web, and in an online help system. The indexing is optimised for the online help version, as it is the most important output format (Robertson 'Online help publishing solution for NRMA Insurance Limited', www.steptwo.com.au/papers/nrma/index.html).

NSW Public Health Bulletin index

The *NSW Public Health Bulletin* has been published since 1990. Recent issues are available online in Adobe Acrobat PDF format and HTML. The annual *Bulletin* index is published in print, and is available online in the PDF version of the print document. There is also a cumulative index online in reverse date order (www.health.nsw.gov.au/ public-health/phb/Subject_Index_for_2002web.htm) that links directly to the articles online. In the case of early issues that are only available in PDF format it links to the issue as a whole.

In 2003 the *Bulletin* was accepted for inclusion in the PubMed bibliographic database (www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed), and indexing in XML format is submitted monthly.

Because the index is required in three different formats, indexing is done using an MS-Access database (Figure 3) that can output text, XML (Figure 4) and HTML (Figure 5). The database includes all the fields that are required by any of the output formats, and the appropriate content is selected for each version.

Figure 3. MS-Access database entry form for PHB article information

```
<!DOCTYPE ArticleSet PUBLIC "-//NLM//DTD PubMed 2.0//EN" "PubMed.dtd">
<ArticleSet>
  <Article>
    <Journal>
      <PublisherName>NSW Department of Health</PublisherName>
      <JournalTitle>N S W Public Health Bull</JournalTitle>
      <Issn>1034-7674</Issn>
      <Volume>14</Volume>
      <Issue>6</Issue>
      <PubDate>
        <Year>2003</Year>
        <Month>Jun</Month>
      </PubDate>
    </Journal>
      <ArticleTitle>An outbreak of Norwalk-like virus gastroenteritis in an
aged-care residential hostel</ArticleTitle>
      <FirstPage>105</FirstPage>
      <LastPage>109</LastPage>
      <Language>EN</Language>
    <AuthorList>
```

Figure 4. XML output from the database

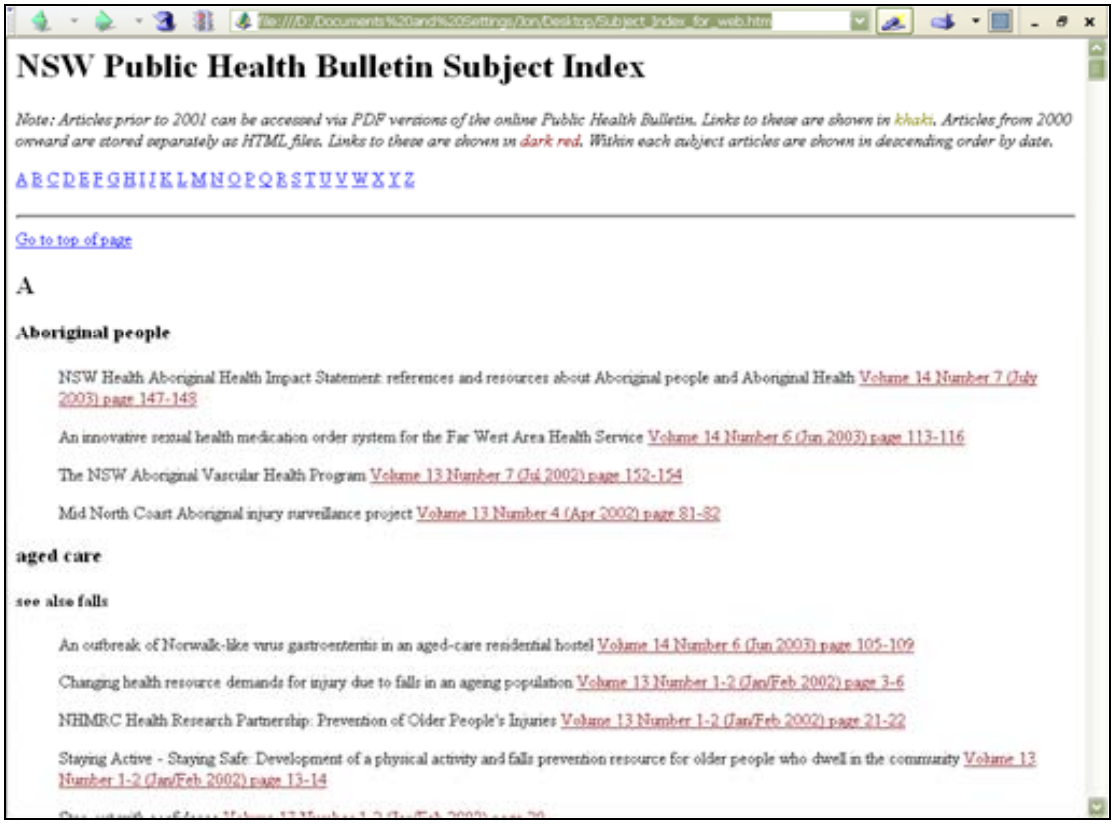

Figure 5. HTML output from the database as displayed in a browser

3. INDEXING POLICIES

To create a back-of-book-style index for a website, the indexer needs to decide what will be included in the project, and the way the index will be presented. This section of the book covers decision-making and policy development, and is followed by discussion of structure and style, and then the format of entry terms and links.

Project definition

To create an effective website index it is crucial to know the purpose of the index and the potential users. Any information about user needs from earlier research should be provided to the indexer.

The structure of the website should be finalised before the index is created (just as with a book you wait for the final page proofs). The web manager will then have to notify the indexer of updates to the site so that the index can be updated.

Whenever you start an indexing project you have to clearly define the work required. When you index a book the editor sends you page proofs on paper, or via email. Indexers need to discuss:

- the purpose of the index, and the potential users
- the size of the project – page or word count, minus nonindexable material
- the material to index – for example, whether appendixes and exercises are included
- the sort of entries to include – for example, names
- the depth of indexing needed – often described as space available, for example, four pages for the index, or as the number of index entries per page
- whether multiple indexes are required – for example, subject, author, recipe

- the format in which the index will be presented – for example, RTF
- the time you will take and what you will charge – often a set quote per job; sometimes as an hourly rate.

Indexers of web documents are sent a group of documents in HTML format, often with links between them. They then need to address the same questions as above. Find out:

- the purpose of the index, and the potential users
- the size of the project – for example, number of webpages; number of megabytes of text; number of words
- the material to index – for example, ephemeral material (news with a brief lifespan); external links
- the type of entries to include – for example, names, level one headings, departments
- the depth of indexing to be done – for example, index entries per project; per file; per heading
- the size of files or sections that index entries will lead to – for example, index entries could lead to anchors at each paragraph
- whether multiple 'indexes' are required – for example, subject; author; chronological
- the format in which the index should be supplied – for example, as an HTML index, with metadata for updating included in the webpages
- who is responsible for loading it on the web
- the updating schedule and arrangements
- data archiving schedules, as the project might need frequent updating
- the time and money available.

If the project will be evaluated it is important to take some pre-implementation measures of search success so there will be a baseline for comparison. Positive metrics can be important for continued funding of the index. [6]

Users and usability

Indexing the web has to start with business objectives and user needs. While book indexers are passionate about taking an unbiased approach and allocating equal weight to all information within a document, web and intranet indexers usually need to give a higher priority to information that is newer, is more commonly sought, is important for legal compliance (for example, privacy law) or is more central to business goals. Content, context (business goals) and user needs all need to be considered.

Usability: efficiency with which a user can perform required tasks with a product, for example, a website. Usability can be measured objectively via performance errors and productivity, and subjectively via user preferences and interface characteristics. Web design features that affect usability include navigation design and content layout.

User: also known as a visitor, participant, actor, searcher, employee, customer, and client.

Usability on the web refers to the interfaces that allow people to achieve their goals. To do this, it is important to design with an understanding of users, what they want to achieve, and how they go about this.

Many interfaces are designed with a typical or average user in mind. The best also offer alternatives for people who wish to take a different approach. For example, search engines may default to simple searching, requiring users to do nothing more than type words into a box, but often offer an advanced search for people with more specific needs and the time and knowledge to search effectively. The appropriate interface might also depend on the environment – for example, business people often want quick searches that give them something useful, while students and academics may be willing to spend time to find *all* of the relevant information on a topic.

It has been found by User Interface Engineering (1999, www.uie.com/branding.htm) that usability of a commercial site is important in presenting a positive brand image for the company involved. This suggests that usability has both community and commercial benefits.

See also 'Book index usability research' in the 'Terms, references, and locators' section.

Jakob Nielsen's usability heuristics

Usability heuristics are generally applicable design guidelines that provide a structure for analysing system usability. Even if you do not perform a thorough user test it can be invaluable to ask a few people to use an index, and to observe their approaches and any errors they make. It can also be interesting to see which access method people choose to use. This may differ before and after they have been introduced to website features such as indexes.

You can evaluate an index using Jakob Nielsen's usability heuristics (www.useit.com/papers/heuristic/heuristic_list.html), which are listed below with notes on their relevance to indexes:

- Visibility of system status – if you have an index, make sure there is a link to it from the home page
- Match between system and the real world – use the language of users
- User control and freedom – offer a choice of index and other access tools; let users move through the index following references

- Consistency and standards – index according to a nationally agreed standard; index all similar pages to a similar level of detail
- Error prevention – users sometimes get confused about the difference between *see* and *see also* references, so rewording one of these (for example, using *search using* instead of *see*) might help
- Recognition rather than recall – because indexes are browsable they allow users to recognise and select an entry, rather then choosing ('recalling') a term to search
- Flexibility and efficiency of use – large indexes will load more quickly if split into letter groups; smaller indexes are more efficient when kept in one file as this makes them readily browsable
- Aesthetic and minimalist design – use minimal capitalisation, so that when capitals must be used they stand out; avoid images that serve no purpose
- Help users recognise, diagnose and recover from errors – cross-references guide users from one location to another possibly more useful one; allow users to backtrack as needed
- Help and documentation – include an introduction explaining general index features and those specific to the individual index.

If you change index features after user feedback make sure you do not introduce more problems when you make the changes.

Lori Lathrop's checklist

Another approach is to use a checklist specifically listing index-related usability issues. Lori Lathrop has published a list of 19 statements for ranking when evaluating a particular index (1999, www.indexingskills.com/usabhtml.html). They include:

- Navigating the index is easy
- The index reflects the terminology of the document
- The size of the index seems appropriate for the document
- The index successfully gathers information together that is scattered throughout the text
- The index entries are concise
- The index includes entries for important concepts
- The index appears to be balanced, with equal treatment for topics of similar importance
- The index contains entries useful to novices as well as to expert users
- The structure of the index is visually attractive.

Watch what they do, don't listen to what they say

When researching user responses to various information access methods, it is important to pay attention primarily to what users *do*, rather than to what they *say*, as users' verbal responses do not always reflect the effectiveness of the tool or approach they are using. It is not uncommon for users to state that they prefer a system in which their performance was lower than in the system they didn't like. (See, for example, Nielsen, 2001, www.useit.com/alertbox/20010805.html).

Paradox of the active user

Sometimes the discrepancy between user preference and user performance is caused by the 'paradox of the active user' (Carroll and Rosson, 1987, www.cs.vt.edu/~rosson/papers/paradox.pdf). [7] Carroll and Rosson identify the problem of 'production bias', in which the paramount goal of a user is throughput, and any time spent learning a system is considered wasted. This means users might prefer a search in which they actively click through many pages, but find nothing, to a slower browsing experience where they feel less in control but actually find more useful information. They also describe 'assimilation bias', which means that people apply what they already know to interpret new situations. This can be helpful when there are true similarities, but it can blind users to features of the new situation they are facing.

The authors suggest a number of approaches that might mitigate these problems, but acknowledge that they are trade-offs that might introduce problems of their own. One suggestion to mitigate the assimilation paradox is to 'Make or describe the system as truly similar to something familiar'. Since indexes are a common information retrieval method in traditional systems, this suggests that they could be of value in providing an old and familiar tool that works well in the new environment.

Skills needed

Web indexers need traditional indexing skills such as the ability to analyse the subject of documents, to describe that subject in appropriate language for an index, to think of alternative access points (that is, other ways of describing a topic), and to create references to or from headings. In addition to traditional skills, web indexers need to understand the structure of information on the web.

Rosenfeld and Morville state that 'While a site index can be a heck of a lot of work, it is typically created and maintained manually, and can therefore be maintained by anyone who knows HTML'.[8] This statement may be based on their limited experience of website indexes, as none of the examples in their book have more than one level of indexing (that is, none have subheadings) and few of them provide other features such as cross-references. Many of them do not even appear to have analysed the text. The AOL site index (www.aol.com/info/siteindex.html)[9], for example, has an entry 'About Find and Search' filed under 'A'. This entry leads to a page called 'Welcome to AOL Customer Support', which is also linked to from an entry 'Search Terms You

Should Know', although the entry 'Search Help' leads elsewhere. There is no equivalent entry anywhere at 'Find'. This example indicates that indexing beyond the provision of simple word lists requires knowledge of indexing principles as well as HTML.

Size

It is easy to skim through a pile of printed pages and work out how much indexable material is there. You can see the size of the pile, and the size of the font used, and you can get a good idea of the amount of white space and number of pictures.

It is much harder to estimate the amount of indexable material on a website. To see it all you have to open separate files. Even then, you cannot always see a whole webpage on one screen, and therefore have to scroll.

Look at the number of megabytes and sample a range of files to get a rough idea. Try and get a word count from a word processing program – most of these will now read HTML files. Ask the web manager to provide as much information as possible about the files, including the number of authors who have contributed webpages. The more authors, the more variation you can expect, and the more sampling you should do.

Materials indexed

A website index should reflect the priorities of the organisation that has created the website. If the site being indexed provides information about a professional organisation then this information should be the focus of the index, with other material, such as details of related organisations, being indexed more broadly.

Decide whether you will index subjects, names of people, organisations, publications, and so on. If you decide to index some names it can be more efficient to index all names than to spend time deciding which ones are significant.

Ephemeral material

A decision has to be made about the indexing of ephemeral material, for example, news with a brief lifespan, and notices of coming events. If possible, index ephemeral material under broad headings that will not change – for example, *courses* – but not under the names of specific events – for example, *database indexing course 22–26 March*.

External links

Some websites, particularly intranets, aim to be self-contained. The indexes to these sites would have no external links. Other websites make extensive use of external links. If you are indexing such a site you have to decide whether you will include external links in your index, and if so, how many.

For example, a bridge players' website might contain a page with links to websites of bridge clubs and bridge players in the local area. The index could include a link to take users directly to each of these external websites. On the other hand, indexing is simpler if the index links only to the page on your website that contains the external links. This allows users to view links in context before following them.

If you do include external links in your index you should indicate in the index entry that this is an external link so that people are not taken to another site without expecting it. Someone will also have to check the external links from time to time. This includes checking that the content on the external site is still relevant as well as checking that the link still works.

Index depth

Be realistic. Remember that the index has to be maintained as well as created. Discuss the expected size of the index with the client. Describe it in terms of the length of the final index, the number of entries per webpage, or the number of entries per paragraph.

Try indexing all webpages at a general level before you do more detail. This will help determine what is practical and appropriate for a site.

On the web you do not have the space limitations that you do with print indexes and you can make your index long and detailed. On the other hand, a shorter index is easier to browse. If you have long lists of subheadings these can be particularly difficult to read, and the main heading they belong to often moves off the top of the page, removing context for the subheadings. Very long indexes may need to be broken up over several webpages.

Target sites for index entries

Anchor (Bookmark): an HTML anchor makes the location in the file at which it is inserted available as a target for a link. It is written in the format *<A NAME=AnchorName>…</A>*.

Pageless index: electronic index in which index entries link directly to the text they refer to rather than listing page numbers for the user to find.

Ascertain the size of files in your project, and the point within the document that you will index to (for example, whole document, section, paragraph). To index to paragraph level each paragraph must contain an anchor. If it doesn't, the web manager or indexer should add the anchors.

In a book, for instance, each chapter might be a single webpage, with anchors leading to each paragraph. In this case, when someone selects an index entry that leads to any part of that chapter, the whole chapter will download before they are taken to the anchor at the paragraph of interest.

To decrease download time, the book could be separated into smaller files. To simplify indexing, links could be to chapters or to section headings rather than to paragraphs.

Once the anchors are in, the index will remain valid even if text is added. If text is removed, the relevant entries will have to be removed from the index. Pageless indexing means that you can index chapters independently, but the index will still have to be edited as a whole when it is complete.

In general, index entries should lead to the webpage with the information content (destination page) rather than to navigation pages which exist simply to lead users to other pages. There are some exceptions to this rule, for example, when there is an index entry leading to the information covered by the navigation page as a whole.

Subsite indexes

There may be more than one back-of-book-style index on a website where there is distinctly different content that needs indexing. For example, a large site might have separate documents such as policy and procedure manuals with detailed indexes of their own, as well as a whole-site index in which the manual has just one entry.

The AusSI website, for instance, has a site index (www.aussi.org/indexer.htm) and also an index to the *AusSI Newsletter* (www.aussi.org/anl/AusSINews.htm).

The Hewlett-Packard site has a number of indexes to individual sections, including HP ProLiant Server Solutions A-Z Index (h18000.www1.hp.com/solutions/proliant_azindex.html) and success stories a-z index (h18000.www1. hp.com/casestudies/atoz.html).

Penrith City has a Quick Index linking to general Council information (www.penrithcity.nsw.gov.au/Lib/ReaderServices/quickindex.htm) as well as an index specifically for library services (www.penrithcity.nsw.gov.au/Lib/AZServices.htm).

Format in which the index is provided

Discuss what format you should provide the web index in. HTML format will usually be appropriate. If using HTML Indexer (see ‘HTML Indexer’ in the ‘Software’ section) you will have to provide the webpages containing the metadata as well as the index so that it can be updated in the future.

See also ‘Index display’ in the ‘Structure and style of website indexes’ section.

Plan for updating

Before you start an index make sure there are arrangements in place for updating it, and announce an updating schedule.

Updating can be made quicker and easier by the use of databases, word processing software, and specialised indexing software. See the 'Software' section for details.

When updating you will need to co-ordinate with the web manager to ensure that you have the latest files to work on, and that they will load your updated index.

The biggest difference between web and print indexes is the fact that websites can change at any time, making the web index out-of-date; the advantage is that you can correct and improve the index at any time.

Index longevity

Maintaining an index requires commitment, organisation skills, stable staffing and an ongoing budget. Some long-lived indexes are those for the Australian Society of Indexers' website (www.aussi.org/indexer.htm, indexed by volunteer society members) and the Los Alamos National Laboratory Research newsletter (lib-www.lanl.gov/libinfo/news/newsindx.htm, indexed by Kathy Varjabedian, who is the newsletter editor and is on the library web team.

Maintaining any index is time-consuming. Reasons for the demise of indexes, apart from the amount of work involved, include changes in web authoring practices including the use of dynamic generation of webpages, and distributed content authoring. These both make centralised indexing practices less likely to continue.

Two innovative web indexes which no longer exist are the University of Texas Policies and Procedures Manual [10] (now replaced by search enhanced with metadata) and the index to Acxiom's *Case-in-Point* magazine (for which the target content is no longer available). [11] The online help-style index to the Quicken site (www.quicken.com) is also no longer in use.

4. STRUCTURE AND STYLE OF WEBSITE INDEXES

Back-of-book-style indexes for websites are based on the structure and style of traditional book indexes. There are, however, many differences because these indexes refer to webpages rather than book pages, so the locators are links rather than page numbers. A website index should be accessible from every webpage in a site, while a book index traditionally sits at the back of the book. Before reading this chapter, use the index to the Australian Society of Indexers website (www.aussi.org/indexer.htm) if you are not familiar with website indexes, as this will make it easier for you to identify the features that are discussed below.

Back-of-book-style website indexes should have the following features:

- a prominent link from the home page, a link back to the home page, and links to other main areas of the website
- navigation buttons for movement to the correct section of the index, and back to the top
- an introduction to explain the scope of the index and any important indexing policies
- a feedback link allowing users to let you know what they thought about the index (an electronic Suggestion Box)
- one heading per link, or symbols such as asterisks to show many links on the same topic, or frames, or links to mini-indexes (these are the equivalent of many page numbers for the same topic in a book index)
- alphabetical or other appropriate order (for example, chronological and geographical)
- subheadings and cross-references.

These features should be chosen with universal website indexing policies in mind. These include the potential users and the purpose of the website.

Style

> **Style sheet:** a block of text in which one or more formats for webpage display are defined. This may include redefinitions of standard formats such as <H1> or new formats specific to that page or site. Style sheets may be embedded in a particular webpage or stored as a separate text file to which some or all of the webpages on a site are linked. Where several style sheets are linked to one page, the order in which they are named determines which ones take precedence in the case of conflicting definitions. These are called cascading style sheets (CSS).

An indexer has to ensure that the format of their index is consistent with the rest of the site. Usually there is a universal 'style' defined for the website. This may be represented by a choice of theme (for example, in MS-FrontPage) or by a cascading style sheet (CSS) file which can be applied to the index.

The index will look different on different computers depending on the browser used, the monitor size and user settings. Indexes can be tested using different browsers or using software that simulates a range of browsers. Keeping a site simple is a good way of making sure it works well for most people.

It is often hard to reproduce the layout of a printed index on a webpage. Many website indexes use bullets as these are the easiest way to create indented subheadings. On some sites wraparound lines are not indented and line spaces are used unevenly.

Internal index links

When the index has finally been edited and checked and loaded to the web, make a link to it from the website's home page. You could also have a Search Page that lists all of the search options available on your site (for example, site map, search engine, index) and lets users choose the appropriate one from the list.

Links in the index will help people navigate within the index and the website. These include alphabet links (alpha bar) at the top of the page to headings starting with different letters, links to return to the top of the index page, and links to go to other pages. Assuming there are no entries for 'Q', 'X', or 'Z', three options are:

> A B C D E F G H I J K L M N O P R S T U V W Y
>
> A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
> (here Q, X and Z are greyed out)
>
> A-C D-F G-J K-O P-T U-Z

You can type the letters and add anchors for each letter in the index, or you can create fancy buttons. You can use only those letters that you have entries for in the index (as in the first example above); you can use all letters and mark unused ones somehow (the best option is to make them lighter) or you can group letters (as in the third example above).

Indexes using separate pages for each letter group

Some indexes provide just the alpha bar (and general information) on the entry page, loading only the index for specific letters as they are selected.

The Lancashire County Library site index (www.lancashire.gov.uk/libraries/atoz/atozindex.asp), for example, splits entries according to the letter they start with. Since there are often only two entries for each letter this is more time-consuming to search than a combined index would be. For example, under the letter D there are the following two entries:

- disabled – services to disabled people
- divisional maps for library locations

The index to the Society of Indexers (SI) website (www.socind.demon.co.uk/site/sitdex.htm) and to Linda Sutherland's site (www.users.zetnet.co.uk/tarristi/sitdex.htm) are both created using HTML/Prep which can automatically organise entries into separate pages for each letter group for faster access (as can HTML Indexer).

Introductions

The *introduction* to a web index should cover all points that are important in any index (for example, types of information indexed) plus anything specific to web indexing, remembering that most people have little experience with web indexes. Unlike book indexes, every user of the web index starts at the top, so you have the opportunity to make contact with them here.

For example, the introduction to an index for a website about pharmacology might read:

> This index contains entries for drugs, drug groups, diseases, disease groups and other subjects. Tables and figures (indicated by a **t** or **f** after the index entry) have been indexed selectively, often with only a heading for the main topic.
>
> All information about a drug is indexed under its generic name. Trade names are included with a reference to the generic name. Word-by-word filing has been used.

Feedback

Web indexes should give the indexer's email address so that users can give feedback on problems they had when searching, or changes they would like to suggest. The great thing about web indexes is that you can fix them!

Filing order

Most indexes are organised alphabetically, although occasionally another order is appropriate. Complicated filing orders should be described in the introduction to the index.

Chronological order might be appropriate for some sites, or for parts of indexes such as subheadings. Luckily alphabetical filing often also gives chronological filing, for example, *Web Indexing Prize 1998* files before *Web Indexing Prize 1999*. At times filing by day, month, season and year or by the logical order of events (*birth, marriage, divorce, death*) may be appropriate.

Sometimes important information is filed at the top of a list of subheadings, even though this puts it out of alphabetical order. One commonly used technique that usually avoids breaking alphabetical order is to index general information about a topic at a subheading called *about*. The advantage of this subheading compared to others such as *overview* and *defined* is that it files at or near the top of the subheadings, thus being more likely to be noticed. In the example below, the subheading *about* leads to a general discussion of the topic carpet laying.

carpet laying
- about
- floor preparation
- stretching carpet
- repairs

Alphabetical access and hierarchical access complement each other. An alphabetical index gives direct access to specific subjects by name, while hierarchical access classifies information allowing you to search from broad concepts to narrower ones, and to see related materials together.

Index display

Most indexes open the target webpage in a full screen, replacing the screen that was there before. This method is simple to implement and matches user expectations and de facto standards.

Frames

Frames are occasionally used in indexes so the alphabet bar and index can be seen at the same time. The index to Nancy Guenther's selected links uses frames (www.chesco.com/~nanguent), as does the Milan Jacovich detective story index shown in Figure 2.

KWIC indexes (Keyword-in-context)

The University of Bristol A-Z index (www.bris.ac.uk/index) is a dynamically generated KWIC (Keyword-in-context) index. It provides access through each 'key word' in a title, and provides the remainder of the title to show the 'context' of the keyword. KWIC indexes do not have any vocabulary control to group

similar information or clarify the meaning of terms, but they can be a 'quick and easy' approach for simple searching. A sample follows:

	Economics
Graduate School of	**Education**
CECAS (Continuing	**Education** Course Administration Service)

Indexes generated for content management systems

Many websites and intranets are now managed using content management systems (CMSs) that have their own templates for indexes. The indexer has to provide content suitable for display in the CMS.

Fred Leise writes about the creation of the PeopleSoft index (2002, www.boxesandarrows.com/archives/improving_usability_with_a_website_index.php) using regular indexing software from which he output the index with embedded HTML coding for inclusion in the CMS's index page template for later publishing to the website. He used codes within the indexing program to 'hide' the HTML coding so the program alphabetised only the index entries. He gave the following entry as an example:

```
<a href="/corp/en/about/pspartner/apply/apply_partner.
asp">Alliance partners, applying to become</a><br>
```

Online help (WinHelp) indexes

People have experimented with online help-style indexes with type-ahead features created using Java applets. These indexes are compressed into a small window so you only see a few entries at a time. They take up less space on the screen, but are less easily browsed. They are slow to load and to use and can function unreliably. See the review by David Brown of the quicken.com index, which no longer seems to be available on the Quicken site (www.aussi.org/webindexing/reviews/Quicken.html).

Geographical display

Maps can be used to allow searchers to browse by area, even if they do not know the name of a place. This should be complemented by an alphabetical list of places for people who know the name of a place, but not its location.

The Virtual Cruise Index has clickable maps and place names organised in hierarchies (1995-1999, continuouswave.com/north-channel/cruise.html), and the Australian Libraries Gateway from the National Library of Australia offers geographic access to help users find libraries in a given area (www.nla.gov.au/apps/libraries, and click on map search or www.nla.gov.au/apps/libraries?action=MapSearch).

The Western Australian Aboriginal languages index by Nick Thieberger shown in Figure 6 gives access by geographical area through maps as well as directly by the name of the language group (coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/WA/section1.htm#1.4).

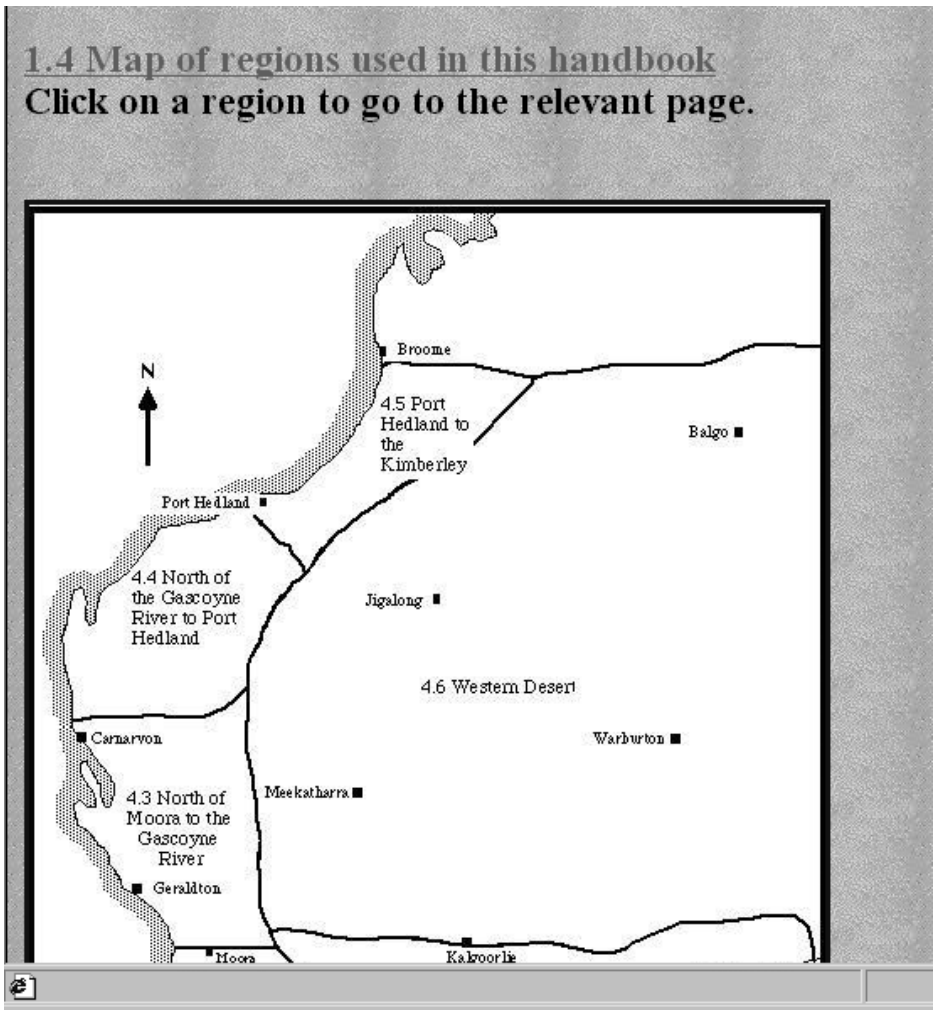

Figure 6. Geographic index to information on Aboriginal languages (Reproduced with the permission of Nicholas Thieberger)

Links to anchors within pages

Anchor (Bookmark): an HTML anchor makes the location in the file at which it is inserted available as a target for a link. It is written in the format *<A NAME=AnchorName>…</A>.*

Some webpages may be many lines long, meaning that users have to scroll as all of the information cannot be seen on the screen at one time. This causes problems as users will not always easily be able to find the information they need within the page, and they can waste time scrolling through unwanted information.

If there are anchors within the site you can use these in the link from the index (with a # symbol) to lead the user directly to the relevant part of the webpage. If there are no anchors, the indexer can add them or ask for them to be added.

If you have long pages with no links or anchors, you could consider splitting the page into shorter pages, or including a note at the top of the index reminding users

to scan the whole document, or giving advice on where specific sorts of information can be found within that long webpage.

While it is best to take users as close as possible to the information they need, the disadvantage is that linking to specific anchors within a webpage will take more time than simply making all links to the webpage as a whole. This is not because it is difficult to link to specific anchors, but because more links will be needed altogether. If, for example, a webpage has significant references to gardening in ten separate paragraphs, this requires ten different index entries if indexing is to paragraph level, but only one entry if indexing is simply to the top of the relevant webpage.

Relative and absolute links

When preparing an index to the site on which your page will be stored, you have the option of using either relative or absolute links. An absolute (remote) link includes the full URL of the page you are linking to, eg:

http://www.aussi.org/membership/memlist.htm

A relative (local) link merely describes the path necessary to get to that page from the index page, eg:

membership/memlist.htm

Advantages of absolute links are:

- anyone can download a copy of your index to their own computer (or another website) and the links will retain their full functionality
- most browsers allow a table of links to be printed out along with the page – for an index with remote links this will give a complete list of sites linked to.

Disadvantages of absolute links are:

- they make it impossible to test the links when working on your index offline
- if the provider changes, or the whole website is moved into another directory, every internal reference in the index will have to be changed. (On the other hand, if the internal structure of the website is changed, relative links can be changed automatically by authoring packages such as MS-FrontPage.)
- they make it difficult to reuse structures from existing websites (for example, alphabet buttons) without extensive modification.

HTML Indexer uses relative links, although absolute links can be added to link to external websites.

5. TERMS, REFERENCES AND LOCATORS

There is a lot of skill involved in selecting index terms and organising them in the most efficient manner. This section of the book looks at the choice of index terms and the use of subheadings and references. For more information about selection of index entries see 'Basic Indexing Principles' in Appendix 2.

Book index usability research

Unfortunately little research has been published about the way people use indexes, and much of that research is preliminary work, with small subject numbers, so indexers often rely on 'gut feeling' instead of empirical findings when choosing between alternative approaches. Research that has been done is discussed below, and key points from different projects have been summarised under a heading for the specific finding.

It is interesting to note that even when index test participants are editorial staff or technical writers, they still encounter difficulties using indexes. This suggests that the average user will find it even more difficult. On the other hand, children in Australian schools now learn formally about indexes and other book structures early in their schooling, so we may soon encounter a more index-literate cohort.

Research by indexing and editorial staff at **Macmillan** [12] identified some problems with index use and made recommendations for improvements. For the study they selected books that have multiple editions published, as usability testing is more worthwhile in these cases. Four books were studied using 22 Macmillan staff as participants (future tests will use nonstaff participants).

Paula Matthews and KGB Bakewell researched indexes to children's information books [13]. They found that children were aware of indexes and their role, but that they did not tend to use subheadings, and found cross-references difficult to understand. They found that children had difficulty scanning pages to find the information the index had directed them to – the use of bold to highlight key points on pages might be useful here. They also found that children were confused about page ranges – on encountering the locator '7-10' they asked 'what does 7 minus 10 make'. The initial suggestion was to use each page alone, for example, '7, 8, 9, 10', but a later (and much better) suggestion was to use words, for example, '7 to 10'. They also suggested printing the entire alphabet on the same page/s as the index, and including letters as section headers for parts of the index starting with each letter.

Corinne Jörgensen and Elizabeth Liddy investigated index usability with students from the School of Information Studies at Syracuse University as participants. Three features were investigated: [14]

- Divided (name/title and subject) versus combined indexes
- Absence of *see* and *see also* references
- Minimal use of concept words

They found a high number of successful searches, but also that subjects were sometimes satisfied with incomplete or peripherally related answers. 'Major categories of errors include stopping short of where the information could be found, problems understanding headings and format, general problems with comprehension, and problems with finding correct entry terms. Problems with entry terms included the subjects' use of adjectives or verbs as headings and finding the right level of granularity.' They found that 'Search strategies of users are unpredictable and not necessarily logical', but also that 'index users are creative and can draw upon internal knowledge to develop a search strategy or to make up for a deficiency in an index.' Their results on cross-references are discussed below.

Susan Olason [15] has examined the usability of indexes from a systems engineering and human factors perspective. Since starting a career in indexing she found users referred to indexes as 'confusing', and feels 'we may have fallen into the trap of indexing for indexers at the expense of our users.' She states that in systems engineering, the most important factor for quality is involvement of users throughout the lifecycle of the product. Her study examined the importance of the following features for index efficiency:

- Run-on versus indented style
- Sub-entries beginning with prepositions or conjunctions
- Access paths.

There were 126 participants 'representing a good cross-section of index users', made up of 'friends and friends-of-friends-of-friends'.

Determine users' needs

Indexers always need to think about their users' needs. This is particularly important on the web where traditional guides such as page numbers and tables of contents may not exist, or may exist in a different form to their print counterparts.

Decisions have to be made about the language used in the index, and this will depend on the specialised skills of the users of the site. For example, you might choose technical terminology for a professional site, and simple language for a site for general users.

Children's sites

The Oakland Zoo site has content aimed at children. The index (www.oaklandzoo.org/atoz/atoz.html) is grouped into categories (for example, 'Birds' and 'Mammals') and is all presented on one page. It uses inversion of terms to bring the main part of the animal name to the front – for example, 'Skink, Blue Tongued' and 'Cockatoo, Sulfur-Crested',

Multilingual audiences

Statistics Canada provides an A to Z index in French (www.statcan.com/francais/public/atozindex_f.htm) and English (www.statcan.com/english/public/atozindex.htm).

Locators (links as page number alternatives)

> **Locator:** the part of an index entry that tells the user where to look for information. In a book index locators are usually page numbers (but can also be references to items, paragraphs and so on). In a web index they are direct links to the information. The links can be the heading or subheadings of the index entry.

Seth Maislin wrote: 'The loss of page numbers and of global context presents a fundamental handicap to writing a good index…It is the indexer's responsibility to make accommodations for the environment.'[16]

Page numbers give clues about the type, importance, and amount of information on a topic. For example, in a book index the entry:

> cats 1, 15-27, *26*, 33, 95

suggests that there is introductory material on page 1, the major discussion on pages 15-27, an illustration (or other graphic material) on page 26, and extra, perhaps unrelated, information on pages 33 and 95.

In a web index, where each locator is simply a link, none of this information is available, and it is necessary for the indexer to add explicit information for the user. Words can be added to describe the sort of information (for example, *overview; defined*), the importance or size of the section of information (for example, *Chapter 7; 15 paragraphs; major discussion*) and the format of the information (for example, *illus.; downloadable file (300 kb*)). A web address (URL) can be added to index entries that link to external websites.

As well as helping users to select the links of most interest to them, this explicit information also helps users target the relevant information on the webpages they are taken to (for example, if they know they are looking for information in a table they won't spend time browsing the text), and, because there are no page ranges in websites, telling users the extent of the information helps them know when to stop reading.

Options to deal with multiple locators include the following (Ream, 2001, 'Web index preparation with HTML/Prep', www.levtechinc.com/Resources/DKRArts/Art_wi.htm_w):

Cats, ♦ , ♦ , ♦ , ♦

Dogs, 1, 2, 3, 4

Fish, 4:3, 6:15, 11:8

Horses, jumping, showing, training

Llamas

description

pictures

Subheadings

Index entry: record in an index, consisting of a main heading and any associated locators, subheadings, and cross-references. This means the whole 'metadata' example below is *one* entry. When indexers charge by the entry they usually define each cross-reference or locator as an entry, meaning the 'metadata' sample below would contain six entries, made up of one cross-reference and five locators.

Main heading: heading at the beginning of an index entry, either used alone or modified by subheadings. The main heading is an entry point into the index. (Cross-references are the other entry points).

Subheading/subentry/subdivision: headings that follow a main entry to modify it.

In the index sample below, metadata is the main heading, and 'Dublin Core' and 'misspellings useful in' are subheadings.

metadata, *see also* thesauri
Dublin Core 15, 33-37
misspellings useful in 14
website structure derived from 99-101, 105

More subheadings are used in web indexes than in book indexes because in a web index each locator (the equivalent of a page number) is a link. It is possible to group a number of links on one line (as in the examples above), but a simpler and more accessible approach is to use one line per link, that is, to make each locator into a subheading.

If a subheading itself has more than one locator (link) it is necessary to reword the subheadings to make them distinct, or to use asterisks to indicate each link.

These alternatives, and others, are discussed below.

Asterisked links on one line

In a book index if there are three page numbers for a subject they may be grouped on one line, eg:

Swimming prizes 15, 34, 72

In a web index the same approach could be taken, and all of the entries put on one line using symbols such as bracketed asterisks to indicate the links. For example:

Swimming prizes [*], [*], [*]

The user would have to click on each of these asterisks in turn. This method is appropriate where all of the links are of equal value. Numbers could be used instead of asterisks; however they have no inherent meaning, and could be confused with page numbers.

One line per link

The simplest method to implement, and the one that takes the user most directly to the information needed, is the use of one line for each link in the index. The link can be at a heading or subheading. The disadvantage of this method is that it makes the index longer, and can be more time-consuming to create. For example:

Swimming prizes (linked to /prizes) Junior (linked to /prizes/juniors) Senior (linked to /prizes/seniors)

If you have two links at a subheading you can reword the subheadings or use asterisked links, eg:

Swimming prizes Junior 1999 Junior 2000

Run-on subheadings

Another approach is to use run-on indexing, as follows:

Swimming prizes: announcements, junior, senior

This gives specific information without taking up as much space as the one-line-per-link method does.

'Continued' notes

When a book index has a long list of subheadings running on to a new column, the editor adds a *continued* note at the top of the new column. There is no comparable feature in online indexes, so it is important to ensure that long lists of subheadings do not get isolated from their main entry. The easiest way to do this is to make sure that there are never more subheadings than can fit on the screen at one time. This is also probably a reasonable limit for online selection by users.

If long lists of subheadings are needed, HTML/Prep can apply 'tips' that display the main heading in a popup box when a user hovers over a subheading. This is useful when the display window is small, or headings have long displays of subheadings under them, and means that context is not lost.

Research into wording of subheadings

Cecelia Wittmann [17] analysed subheadings in four award winning indexes. They were found to:

- Be on average five words long
- Start with a significant word (either a noun or a verb, especially avoiding beginnings such as 'and' and 'in')
- Not be related syntactically to their main headings (for example, 'Statistical material: units of measure in' is syntactically related, whereas 'Statistical material: units of measure' is not)
- Not exactly match words from the text.

These characteristics were not shared by non-award winning indexes to similar books. Conclusions were that the best subheadings are coined, rather than copied directly from the text, and sum up the topic aspect concisely and that prepositions at the beginning of subheadings should be avoided where they do not add meaning or clarity.

Susan Olason[18] found that subheadings that did not begin with prefix words (prepositions or conjunctions) had higher efficiencies and usefulness rankings than those that did. That is, a subheading such as 'emergencies' was preferred to 'in emergencies', and 'feeding and' was preferred to 'and feeding'. 'Comments about prefix words included frustration about being forced to read rather than scan, confusion about sorting (users did not realize that prefix words were ignored in sort), and confusion about their purpose (did not clarify the main entry/subheading relationship).'

This research confirms Wittmann's findings, and suggests that prefix words should be avoided when possible. To complete the research, however, it would be necessary to compare the lack of prefix words with the results when prefix words are *not* ignored in sorting.

Research into indented versus run-on subheadings

Indented style index: indented indexes start each subheading on a new line, indented under the main heading. For example:

names
    indexing rules for 41-42
    keyword searching and 5

Run-on (run-in) style index: run-on indexes list all subheadings in sequence, separated by punctuation such as semicolons. For example:

names: indexing rules for 41-42; keyword
searching and 5

Susan Olason found that indented-style indexes had higher efficiencies and higher usefulness rankings than run-on-style indexes. Indented indexes were ranked as user-friendly 90% of the time, while run-on indexes were never ranked as such. 'Comments about run-on indexes included frustration about being forced to read rather than scan, confusion about sorting and confusion about which page references went with which sub-entry.'

Research into distinguishing main entries from subheadings

Participants in the **Macmillan** [19] study commented that 'first-level entries [that is, main entries] in bold type greatly enhanced the usability of the index'.

Jörgensen and Liddy [20] suggest that 'more effort and thought needs to be put into making scanning an index an easy task, particularly in the area of distinguishing between headings and subheadings.'

Research into entry points

The **Macmillan** study recommends that indexers include more double and triple postings to provide as many entry points as possible. They wrote: 'Observers were surprised at what participants looked up…participants searched for terms they [that is, the observers] would never have thought of including in the index'.

As discussed above, **Susan Olason's** research suggested that users needed a broad entry point for the main topic of the book, with *see also* references leading to narrower topics. This entry acted as a miniature table of contents within the book.

Cross-references

Cross-reference: a *See* reference or *See also* reference leading the user from one part of the index to another.

***See also* reference:** directs index users to related topics that could be consulted in addition to the topic they are currently at: for example, 'beds, 26, *see also* cots'

***See* reference:** a way of indicating to a user that they should look elsewhere. A cross-reference may point to two or more locations: for example, 'rodents, *see* mice; rats'. The choice of which terms to use and which to refer from depends on the wording of the material being indexed and the target audience.

See references or double entry

A good index gives access to information no matter which synonym a user searches by. A choice has to be made about whether to use *See* references (for example, 'felines, *see* cats') or double entry (that is, having the same links at *felines* and at *cats*).

Use of double entry will save the user one step, but both of the terms will have to be updated whenever material on that subject is added or removed from the website. (If the index entries are stored in a database the double entries can be automatically generated.) When one of the index links is selected, all of the links to the same page will change colour (showing a link that has been followed) thus making sure that users do not select a link to the same webpage twice.

If a *See* reference is used (with a link from the unused term to the heading you have chosen to use) the user has one more step, but the index is easier to update and less likely to have inconsistencies. *See* references can also be useful when you want to lead users away from terminology because it is ambiguous, out-of-date or inaccurate (for example, 'quotes, *see* estimates'; or 'New Hebrides, *see* Vanuatu').

In the example below the link at *WHO* takes the user to the index entries for *World Health Organization*. They then have to select the appropriate subheading to be taken to the information on the website.

WHO, *see* World Health Organization [links to full entry]
World Health Organization (WHO)
funding
history

In the double entry case below, the links at WHO take the users directly to the webpage on which WHO is discussed. The users immediately find the information they need whether they look up the full name or the abbreviation, but the indexer has to remember to update both terms.

WHO (World Health Organization)
funding
history
World Health Organization (WHO)
funding
history

If items about the same subject are grouped in the index, but the terminology used on each webpage is different, the extra information should be added as a qualifier, as in the example above for *WHO*. This helps the user who has looked up *World Health Organization*, but only finds the abbreviation *WHO* on the target page. This is also important when you group near-synonyms such as *independence* and

autonomy. Indexers have to make it as easy as possible for the person who has looked up *independence* to find the word *autonomy* on the page.

The PeopleSoft Products (www.peoplesoft.com/corp/en/indices/prod_index.jsp) index provides *See* references, but links from the name of the *See* reference directly to the relevant text. This ensures that the users know the name of the page they are going to, but also get there as quickly as possible. For example:

> Application Integration Broker. *See* Process Integrator
> Asset Liability Management
> Asset Management. See Fixed Asset Accounting

See also references

See also references are used to guide readers to alternative or additional index entries. They should be at the top of the index entry so that they cannot be missed and should be reciprocal where appropriate (for example, 'metadata, *see also* web indexing' and 'web indexing, *see also* metadata').

Indexes use specific terms, while users often search more broadly. Someone wanting information on koalas might look up *marsupials* in the index. References to narrower terms can therefore be useful. For example,

marsupials, *see also* kangaroos; koalas

If there are no references to marsupials in general then this reference should be

marsupials, *see* kangaroos; koalas.

Where there are many narrower terms, or where the index is continually changing, a general reference may be more appropriate, eg: 'marsupials, *see also* names of specific marsupials, for example, koalas'.

Cross reference target highlighting

Using the JavaScript programming language it is possible to create an HTML index which will highlight the relevant term when the user clicks on a cross-reference: that is, if the index has a hyperlink reading 'goldfish *see* carp', then clicking on that link will take the user to the section of the index containing the term 'carp' and highlight that term by changing the background colour or otherwise modifying its format. An example can be seen at www.optusnet.com.au/~webindexing/testscript6.htm. [21]

Research into cross-references

The **Macmillan**[22] study found: 'that *see* references were not a problem…*See also* references, on the other hand, were confusing to some. For example, if they saw "Web pages. *See also* Web sites," some expected to see "Web pages" as a subheading under "Web sites." That is, they were reading *see also* as *see under*.

On the other hand, in an online help project I worked on we used *Search using* instead of *See* because of anecdotal evidence that users were confused by the *See* reference. One of the advantages of *Search using* is that it is clearly distinguished from *See also*.

Jörgensen and Liddy[23] found that when using an index without cross-references (*see* and *see also*) users were slower and made more errors than those using an index with cross-references. However, they also found that the overall success rate was higher for the index without cross-references. Common errors in using cross-references include 'reading the see/see also as part of a main heading, part of a subheading, running separate references together, or reading a heading and subheading as part of a *see also*. Across all uses of the Basic Index, many users did not understand the structure or the function of *see also* references, and many exhibited an openly hostile reaction to them, saying, for example: 'This thing is so trivial. [Why?] Because it keeps going back and forth and it doesn't ever give you a page for what you're looking for.'

6. SOFTWARE

This section discusses the software used in website indexing, ranging from general web tools (for example, HTML) through specialised indexing software and on to tools created specifically for the creation of website indexes (HTML Indexer and HTML/Prep). [24]

File formats

HTML

Anchor (Bookmark): an HTML anchor makes the location in the file at which it is inserted available as a target for a link. It is written in the format *<A NAME=AnchorName>…</A>.*

<HEAD> section: The <HEAD> section of an HTML document is placed at the top of the page between an opening tag <HEAD> and a closing tag </HEAD>, and contains metadata about the document, not the content that will be displayed on the page. It is followed by the <BODY> section.

Relative addressing (Local link): linking to another page on the same website through a local address rather than a URL: for instance, a link back to the homepage might take the form *<A HREF=/home.htm>Home page</A>*.

Tag: a piece of text that describes the semantics or structure of a unit of data (element) in HTML, XML or other markup language. Tags are surrounded by angle brackets (< and >) to distinguish them from text. 'Tags' is also used to describe the code indicating index entries in embedded indexing.

HTML files, which make up most of the material on the web, are plain text (ASCII) files. These files, when given an .htm or .html extension, can be opened by a browser program and viewed on the user's computer. HTML files found on the web are copied on to the user's hard disk and stored in a temporary cache directory while they appear on the user's screen: thus each user is actually viewing their own copy of the file, not the original.

HTML documents are made up of a 'HEAD' section and a 'BODY' section. The HEAD section contains any metadata that has been provided, while the BODY

section contains the text that will be displayed on the page (and its associated coding).

The production of a website index consists of setting up appropriate links from an index page to other pages on the site. To provide a link to an HTML file located outside the site the user must know its address (Uniform Resource Locator or URL). This usually takes the form *http://www.domain.xxx/filename.htm* (for example, *http://www.aussi.org/constitution.htm*). [25]

Tags of particular interest to indexers include anchors, which mark a specific place in a webpage so that a link can be made to it; and links that take a user from the webpage they are on to another place within that page, to another webpage, or to a specific place within another webpage. To give an idea of how these are structured, examples of three kinds of HTML are shown below.

1. An 'anchor' tag (also known as a 'bookmark') 'anchors' or identifies a particular section of the file by name, enabling a link to be made directly to that section. An anchor tag takes the form:

```
<A NAME="TopOfPage"> XXXXXXXXXXXXXXXXX </A>
```

where *XXXXXXXXXXXXXXXXX* is the section of the file which is anchored under the name *TopOfPage*.

2. A second type of tag provides a link from one section of the page to an anchored section elsewhere on the same page: this allows the user to click on a link to move automatically up or down the page to the location specified (for example, clicking on the phrase "Return to Top" anywhere within the index would take the user to the top of the index). This takes the form:

```
<A HREF="#TopOfPage"> Return to Top </A>
```

where *Return to Top* is the clickable text that appears on the page, and TopOfPage is the name of the anchored section that the link takes the user to.

3. A third type of tag provides a 'local' link from one webpage to another page in the same site. This takes the form:

```
<A HREF="Chapter2.htm"> Go to Chapter 2 </A>
```

where *Go to Chapter 2* is the clickable text that appears on the page, and Chapter2.htm is the name of the file that the link takes the user to.

Tags of types 2 and 3 can be combined to take the user to an anchored location on another webpage: for example, for the user to go directly to the Conclusion on the webpage containing Chapter 2, a link could be provided in this form:

```
<A HREF="Chapter2.htm#Conclusion"> Chapter 2 Conclusion </A>
```

Many of the tags found in web indexes will be of this type.

Although HTML tags can be created by hand, most large-scale web authoring and indexing programs use procedures that automate the coding process and allow users to concentrate on analysing the content of the site and providing appropriate connections. It is still useful to be able to read HTML to diagnose problems and make small changes.

XML

XML (eXtensible Markup Language) is related to HTML and allows material to be marked up semantically (by meaning) rather than for its desired appearance or role in a document. XML provides an open standard for compiling and accessing diverse data collections. Explicit specifications can be provided to control and validate XML data structures. XML also includes tools for converting between different sets of specifications. XML, like HTML, is a spin-off from the large and complex Standard Generalised Markup Language (SGML), ratified by ISO 8879 in 1986. The XML Working Group removed some of the less-used features of SGML and worked towards making a relatively simple language that could be used by a wide variety of applications over the Internet.

There is a family of related XML-based systems including (Jones, 2002, www.devx.com/devx/editorial/10244):

- XHTML. This is the preferred way of writing HTML as it is well-formed XML and can be manipulated more easily than the earlier HTML
- XSLT and XSL are languages that transform XML documents into something else, including text documents, PDF files, HTML or comma-delimited files
- DTDs (see below) and XML Schema describe the type of content that an XML file can contain, and let you validated the contents of XML documents
- Xpath and Xquery are query languages that let you extract single items or lists of items from XML documents
- SOAP is a standard communication protocol between web services.

XML has three main functions:

- To support a standard open system for representing tabular information of the kind found in proprietary databases and spreadsheets – ‘Database markup’
- To support a markup system for electronic documents so that they can be produced and distributed in a standard form – ‘Document markup’. This goes beyond the relatively simple markup provided by the use of HTML in web documents and allows for elaborate documents that can appear in many different ways depending on the user’s requirements at the time
- To allow for the representation of non-text information in a standard textual way across universal distribution systems (like the web).

XML files can be viewed (but not edited) in a web browser. Here they appear as a colour-coded and indented sequence of elements that can be expanded to show their contents or ‘rolled up’ with a click of the mouse. A plug-in is available for

MS-Internet Explorer that will also verify the correctness of XML. More complex formatting can be achieved through dedicated XML style sheets.

With a book marked up in this standard way an author, editor or publisher can routinely produce large-print copies for the visually impaired, extract tables of contents and first chapters for reviewers, collect together books with the same authors, arrange books in chronological order, check that a new book doesn't have the same title as an existing one, compile the book into a form suitable for reading on a hand-held computer, etc; and the same software could be used world-wide to do this.

XML is already in use at Cambridge University Press as a method of pageless indexing: sections of a book are tagged with opening and closing XML code which includes the index entries to be used for that section. The index can then be extracted regardless of the actual pagination of the work. You can download the 47-page author instruction manual at https://authornet.cambridge.org/information/productionguide.

World-wide standards (*schemas*) for XML applications are being developed. See the section on the 'Semantic web' below.

Document Type Definitions (DTDs)

DTD (Document Type Definition): schema specification method for XML documents. A DTD is a collection of XML markup declarations that define the structure, elements and attributes that can be used in a document that complies with the DTD. By consulting the DTD a parser can work with the tags from the markup language the document uses. DocBook is an example of a DTD often used with technical documentation to enable sharing and reuse.

Markup language: a way of depicting the logical structure or semantics of a document and providing instructions to computers on how to handle or display the contents of the file. HTML, XML and RDF are markup languages. Markup indicators are often called tags.

DTDs (Document Type Definitions) are XML markup declarations that define the structure, elements and attributes that can be used in a document.

DocBook is a markup language used for writing structured documents using XML. The DocBook DTD was developed for use in creating computer documentation, and can be used to create an embedded index (Brown, www.allegrotechindexing.com/news014.htm and 'DocBook: the definitive guide', www.oasis-open.org/docbook/documentation/reference/html/indexterm.html). David Ream and Seth Maislin have developed a draft DTD specifically for book indexes (follow links at www.levtechinc.com/ProdServ/Presentations.htm).

NKOS is working on the development of a DTD to enable reuse of knowledge organisation systems (KOS) such as taxonomies and authority files. Information about NKOS, the VocML DTD and a draft taxonomy of knowledge organisation systems can be found at www.alexandria.ucsb.edu/~lhill/nkos (scroll down).

There are a number of HTML DTDs (for example, 'HTML 4.01 Strict DTD', www.w3.org/TR/REC-html40/sgml/dtd.html). If you check the source code of some pages (View/Source) you can see the DTD referred to (for example, www.searchenginewatch.com and dcanzorg.ozstaging.com/mb.aspx).

Book indexing software

Like every profession, indexing uses specialised software. Book and website indexers work with indexing programs which allow the indexer to enter a large number of index terms, subheadings and page references, and organise them into alphabetical or other orders. They also allow for checking of cross-references and for output to a formatted word processing file for final printing and sending to the client. Database indexers often use database-specific software, and may also use thesaurus programs to maintain a controlled vocabulary.

The most popular programs for indexers are CINDEX and SKY Index (produced in the USA), and Macrex (produced in Britain). CINDEX and Macrex began as DOS programs but have produced Windows versions; SKY Index began as a Windows program. Michael Wyatt has compared CINDEX and SKY Index (1998 and 2000, www.aussi.org/resources/software/review.htm).[26]

Word processing software such as MS-Word and publishing software such as FrameMaker and PageMaker allow users to embed indexing terms in a document and to generate indexes. This allows indexing to be completed before page numbers have been finalised, and allows index tags to move with portions of the document if they are reused elsewhere.

'Indexing software' also includes programs that claim to analyse and index text automatically, without human intervention. These use structural features of the text (capital letters, words repeated several times) to try and make semantic judgements, identifying 'important' words and phrases and then listing these in alphabetical order with page numbers attached. The results are unimpressive for books and journals, but in a restricted environment such as database indexing these programs may have a role to play.[27, 28]

Finally, indexes created for PDF documents can be automatically linked to the text they refer to using a program called Sonar Activate. These indexes may be made available on the web. These programs will now be discussed in more detail.

CINDEX

CINDEX (www.indexres.com/cindex.html) is a book indexing program that is often used in conjunction with HTML/Prep for the creation of website indexes. It provides the ability to link cross-references within the file when used as a standalone product for website index creation.

For information on the CINDEX Users Group go to groups.yahoo.com/group/cindexusers/join.

Macrex

Macrex (www.macrex.cix.co.uk or www.macrex.com) is a book indexing program that creates indexes that link to web addresses. It can write a markup language coded index (not just HTML), either for posting to the web or for inclusion on disk or CD-ROM. The last subheading can be the link (good for indexing websites and single issue documents) or the final index entries can point to multiple targets.'[29]

To join the Macrex discussion group, send the message 'Join macrex list [your name] [your serial number]' to mailto:mailjoin@macrex.com.

SKY Index

SKY Index (www.sky-software.com) is a book indexing program that generates indexes that can be viewed with web browsers, and automatically links cross-references to the headings they refer to, but it does not create links from index locators to other sites on the web.

For information on the SKY Index Users Group go to groups.yahoo.com/group/skyindexusers/join.

MS-Word

MS-Word and other major word processing programs include an indexing feature designed for embedded indexing. Tagged index entries are inserted into document files and are later compiled to make an index. Because the indexing is embedded into the text, the page numbers in the index automatically change if the page numbers in the document change and the index entries are removed if the pages to which they refer are removed. While this is useful for certain types of indexes, for example, those for texts that are continually changing, or where the indexing must be started before the book has been completed, it is in general too cumbersome for most indexers.[30] Documents prepared in MS-Word can also be modified for the web using ReWorx, below.

ReWorx

ReWorx (www.republicorp.com/reworx.htm) is a program that converts Word documents to online documents. It can be used to generate XML documents, online help systems, intranets and websites. A demonstration version can be downloaded from the web.

The program recognises heading levels in the document and uses this information to generate a table of contents. This is placed at the left hand side of the screen and opens up to give more information as required (in a Windows hierarchical folder view). If the MS-Word document being converted has an embedded index, this will be used as the basis for an online index.

Sonar Activate for PDF documents

If you have a PDF document you can create links from the index to the text using Sonar Bookends Activate from Virginia Systems (www.virginiasystems.com).

The program is an Adobe Acrobat plug-in, so you need the full version of Acrobat as well.

With the text of the book in one PDF file, the indexer exports the index to PDF and inserts it into the file for the book. The program automatically generates hyperlinks for each page number in the index and table of contents. It works by looking for a comma-separated list of numbers after text and some white space, for example, 'gorillas 5, 22, 96'. It only works for one sequence of numbers, so if you have preliminary matter (for example, pp i-xii) you have to tell the program to ignore that section (you can add those links manually if needed).

Automatic indexing of documents using Syntactica

Syntactica (now apparently defunct) [31] is the latest 'automatic indexing' system intended for general use. Syntactica is a web-based service. Users pay a subscription fee which entitles them to upload their own documents to the Syntactica site. An opening balance of $US10 is credited to new users, allowing them to try Syntactica out on up to 100 pages of text before a payment is required. A registered user is given their own password-protected workspace in which the 'indexes' are kept and made available.

Syntactica can process text, RTF and Word files. Uploaded documents are analysed and the resulting 'indexes' become available for the user to download. Three indexes are available for each document; a short (Min) index with about one term per 100 words, a longer (Mid) one with about 4 terms per 100 words, and a longer one again (Max) with about 1 term for every 10 words. These ratios appear to decline for longer documents.

Once the indexes are produced the user can view any of these in a pop-up window alongside a plain text version of the input file, with hyperlinks from the index terms to the text they refer to. The indexes can be edited in another pop-up window by adding or removing entries and rewriting index terms. The indexes can be downloaded as text files and marked-up versions of the original documents are made available in Word format for making embedded Word indexes. The actual operation is slick and user-friendly.

The results, however, are disappointing. Syntactica does almost nothing that an indexer would recognise as analysis. It also makes blunders which any indexer would avoid:

- Every index entry is capitalised
- Sequencing is in ASCII order
- It doesn't use index terms that are not in the text
- It doesn't understand synonymy and it has trouble with names and plurals
- It can't distinguish most multi-word descriptions from nouns + verbs.

Website indexing software

You do not need specialised software to create a website index, but packages are now available that will save you time, especially for an index that needs regular updating. The software to use will also depend on your client's requirements and programming capabilities. Options are discussed below.

Basic unlinked HTML indexes

If an index is to be presented on the web without any links at all, simple HTML coding can be used. An easy approach is to use the 'Save as/ HTML' option in programs such as MS-Word. The result is simply the equivalent of a paper-based index on the web, and not a web index with links.

Linked indexes in HTML format

If you want links in your HTML index, you will have to add them manually using programs such as MS-Word, MS-Excel and MS-FrontPage, or use one of the tools mentioned below. It is often relatively easy to take convert an index in MS-Word format to text containing appropriate HTML tags using a **macro**. For example, a simple macro can convert

```
Greensleeves
greengages, unripe
Green, Harry
```

to

```
<A HREF="Greensleeves.htm">Greensleeves</A>
<A HREF="greengages#unripe">greengages, unripe</A>
<A HREF="Green#Harry">Green, Harry</A>
```

which can be saved as a text document with an .htm extension and used as a hyperlinked index. This assumes, however, that the linked pages and the sections on those pages are all named in accordance with the entries appearing in the index. The more deviation there is from this the more manual editing of the index will be required. Problems may also arise with capitalisation and the use of multiple subheadings.

Many early web indexes used HTML Definition Lists (Glossary Lists) to create the heading/subheading structure. The DL element provides a beginning and ending line break. In the DL container, the <DT> tag marks the term and the <DD> tag defines the paragraph. The standard format of a definition list is:

```
<DL>
<DT>Term
<DD>Definition of term
</DL>
```

The DD tag displays a single paragraph, continuously indented one or two spaces beneath the term element's text. For the creation of web indexes, the subheadings

are coded <DD> and are therefore indented. An article by Linda Fetters on indexing the University of Texas Policies and Procedures Website provides an example of this format.[32]

Databases

A database is a collection of records about individuals. Each record is made up of a number of fields relating to different properties of the individual. Many large indexes are stored in databases. In the case of a web index, each individual could be an image or a webpage. The back-of-book indexing programs discussed above are essentially database programs which have been modified to provide specialised facilities.

Databases save time on routine tagging, are relatively easy to update, allow automatic generation of double entries, and give indexers control over features such as sort order. Indexes stored in databases are usually managed by IT specialists in conjunction with indexers, and are only mentioned briefly here.

HTML indexes can be created from an MS-Access database, for instance, by producing a report which includes the HTML tags as well as the fields in each record: for example, if a field in Access is called 'Entry', a field in the report design saying

```
"<A HREF=" & Chr(34) & [Entry] & ".htm & Chr(34) >"
& [Entry] & "</A>"
```

will produce the corresponding output for each line:

```
<A HREF="Greensleeves.htm">Greensleeves</A>
```

Storage of indexing data in a database enables single sourcing – indexing once, and then reusing the index terms in different formats. See 'NSW Public Health Bulletin index' in the 'Single sourcing' section for an example.

Indexes generated automatically from metadata

Indexes can be automatically created from metadata instead of being manually crafted. This can have advantages including distributed responsibility for indexing and improved currency of the index, although it can also mean poor or no indexing. Indexes created from metadata are basically limited to one level, that is, they don't have individually coined subheadings to make aspects of topics explicit, although page titles or narrower terms can act as subheadings. They can include dynamic *see also* references linked within the index, and *see* references that take the user directly from the term in the reference to the information they require. That is, in the entry 'CSS, *see* Cascading style sheets', the words 'Cascading style sheets' would lead the user straight to the information they require.

The Montague Institute Review site provides a table of web indexes (www.montague.com/review/AtoZ.shtml) in which it is the only example that is dynamically updated. This is made possible by the use of metadata stored in a

database but does mean the index only has one level. The Montague site also has chronological, organisation and people indexes.

James Cook University automatically generates its alphabetical index from the contents of an A-Z metatag, which is added by content creators (www.jcu.edu.au/atoz and www.jcu.edu.au/office/itr/webmanagers/atozletter.html).

The index on the Meta Matters site (dcanzorg.ozstaging.com/mb.aspx) is automatically generated from metadata that has been created using the Metabrowser editor. The index includes document titles, categories and synonyms (metabrowser.spirit.net.au/prodServer.htm). Narrower terms from the taxonomy that has been used for term selection are shown as subheadings of terms in the index. Meta Matters also provides access through a faceted classification.

Wiki indexes

Wikis are communally created websites with information provided and edited by a range of contributors [33]. All webpages are given a topic (containing no spaces), and these topics can be searched in an alphabetical list, which acts as an index to the content of the wiki. They are not consistently constructed; for example, they may include both 'InformationArchitect' and 'InformationArchitects' as entries (the words are combined without a space to make a topic). They also only offer one level in each index entry, but they are a useful tool for browsing to find specific topics, or to explore the range of content on the site. The index to the IAwiki (iawiki.net/cgi-bin/wiki.pl?action=index) includes the following entries:

MailingList
MailingLists
MakovisionBlog
ManualCategoryGeneration
MarkBernstein
MemeKitchen
MetaData
MetaWiki
MindMapping
MuscleMemory
MyWhine

HTML/Prep

HTML/Prep from Leverage Technologies (www.levtechinc.com/ProdServ/LTUtils/HTMLPrep.htm) converts indexes that have been created using word processing programs or dedicated indexing packages into HTML files. It is useful for single sourcing where a document has to be maintained in print and online formats.

A PDF manual is delivered with HTML/Prep, and can be viewed on screen or printed by the user. The manual includes brief information on designing indexes for the web.

HTML/Prep has been used to create the index to the Society of Indexers (UK) website (www.socind.demon.co.uk/site/sitdex.htm) and the index to the Milan Jacovich detective series (see Figure 2; www.levtechinc.com/Milan/MilanHN.htm). [34]

HTML/Prep automatically creates links from page numbers (or other locators) to webpages with those locators as names (for example, page 53 is linked to 53.htm). Within the index it links cross-references with the main headings to which they refer, and can generate an alpha bar and 'Return to Top' links.

System features

HTML/Prep is a command-line invoked program running under the command prompt window, although it is 32-bit Windows software.

An alternative to working in the command prompt window is to use the program WinCommand, which lets you create an icon to run HTML/Prep from the Windows desktop or a folder. You can then type parameters into the WinCommand dialog box instead of typing them in the command prompt window (www.levtechinc.com/ProdServ/LTUtils/WinCmd.htm).

Input files

Files need to be tagged before they can be converted. Tags indicate which lines are headings and subheadings (so the subheadings can be indented), which ones are cross-references, and where the locator information starts.

There are two options when creating linked indexes. You can use the last subheading as the link text and the page field as the link value or you can put both the link text and the link value in the page field with the embedded tag <p> separating them.

HTML/Prep options

HTML/Prep options can be used to specify the appearance and content of the index. You can type the options in any order, but must separate them with spaces. For example, the option '-C*string*' specifies that the locators should be displayed and not treated as links. The *string* is the characters, if any, that the locator lead-in tag is replaced with. The option 'DT*text*' specifies the text to display as the link text. The default is 'Click here', as seen in Figure 7.

HTML/Prep can apply 'tips' that allow the user to hover over a subheading and see a popup box showing the heading structure above the current position in the index. This is useful when the display window is small, or headings have long displays of subheadings under them. You can see tips in the Yale Undergraduate Regulations Index (www.yale.edu/ycpo/undregs/pages/indexpage.html) and the Milan Jacovich detective series index (www.levtechinc.com/Milan/MilanHN.htm).

Other HTML/Prep options include the use of frames and the inclusion of tags to specify font, colour and so on. There are also specific instructions for the use of HTML/Prep with the specialised indexing program CINDEX.

The output files contain the index, main headings, and letter list formatted for web usage. The .htm files may need to be edited if extra tags or data are required. A separate document containing just the index's main headings and cross-references is also produced, enabling browsing of just the main headings in a large index. Links from the main headings take the user back to the full index when more detail is needed. An example of a 'main heading index' can be seen in the BNA Labor Relations Reporter Index (www.bna.com/lrr/lrrindx.htm).

An example of an index and its tags is given below:

Sample index:

live files 73
  defined 72, 85
  editing of 70
local links, *see* relative addressing
locators 57-58, 85, *see also* pageless indexing
pageless indexing 10, 66-69
relative addressing 24, 59

Tagged index:

<l0>live files <c>73
<l1>defined <c>72, 85
<l1>editing of <c>70
<l0>local links, <x><I>see</I> relative addressing
<l0>locators <c>57-58, 85, <x><I>, see also</I> pageless indexing
<l0>pageless indexing <c>10, 66-69
<l0>relative addressing <c>24, 59

Default index generated by HTML/Prep.

live files ...Click here
  defined ...Click here, ...Click here
  editing of ...Click here
local links, *see* relative addressing
locators ...Click here, ...Click here, ...Click here, *see also* pageless indexing
pageless indexing ...Click here, ...Click here
relative addressing ...Click here, ...Click here

Figure 7. HTML/Prep tagged index and default index

HTML Indexer

HTML Indexer is a program for IBM-compatible computers that automates the clerical aspects of creating web indexes. It produces a back-of-book-style index with an alpha bar (hyperlinked line of initial letters). Users can select default entries or add their own, and can set various output options. HTML Indexer uses embedded indexing of metadata tags within the webpages themselves to store indexing information, thus making the indexes it produces easily updatable. HTML Indexer also produces output for HTML Help and JavaHelp indexes.

A demo version is available on the Internet (www.html-indexer.com). It is fully functional except that indexing projects cannot be saved for updating, although the indexes created can be saved and used.

Default entries

When using HTML Indexer the webpages to be indexed in the project must be selected. The program then creates default entries for all titles and anchors within the webpages (using the first heading on each page and the text of the anchors). If a page has no heading, the title is used. If there is no title either, a blank entry is created. Blank entries are not included in the finished index.

Default entries can be deleted or edited and unneeded files can be blocked from the project so that default entries for them are not created. Some web authoring programs (e.g. MS-FrontPage) generate noncontent files; these can be blocked from the project. Most other default entries will need to be edited to remove formatting or to make the wording more appropriate for the index.

Adding entries

To add new entries users click on the Add button and type index entries into a dialog box. The dialog box can be kept open throughout the project as indexers move from one webpage to the next.

A specific page (or anchor) must be selected when entries are added. If the main folder is selected when entries are added they will be added to every page in the folder.

Subheadings are created by the use of a comma between the main entry and subheading. *See* and *See also* references can also be created.

The pages being worked on can be seen one at a time by selecting *View/Selected Source Page* from the menu. If indexers are working through a long document paragraph by paragraph it is helpful to be able to keep the whole document open in a second window.

A sample HTML Indexer index output is shown in Figure 9, and a sample project is shown in Figure 10.

Style

Style options that can be set include:

- run-in or indented style
- the number of columns for the index
- the style and position of hyperlinked initial letters (alpha bar).
- material to be added to the top or bottom of the index
- inclusion of Return to Top links.

Sort order

The sort order (that is, the order in which the words appear) can be different to the order of the index entries. This is necessary when there are formatting tags in the entry (as in the example below), or when the sort order does not reflect the exact alphabetical order of the entry (for example, if 'the' is to be ignored in filing). To specify a sort order different to the alphabetical order of the index entry, add the characters to be sorted on to the bottom section of the dialog box. For example, if you have added tags to make an entry italic, you will have to remove the initial tag in the 'Sort as' section, as shown in the screen shot below.

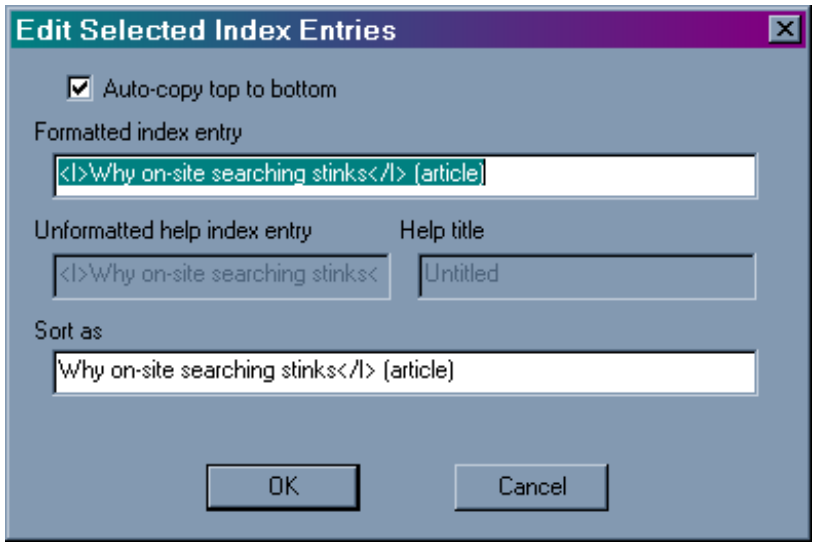

Figure 8. Initial italic coding in index entry removed in 'Sort as' box to ensure correct filing order

In cases where a heading with subheadings and a phrase start with the same word the sort order might have to be manually forced within HTML Indexer or in the final index. This is because sorting is done when headings and subheadings are together taking punctuation into account. Unfortunately, this sorting is not always appropriate to the final index, and the order needs to be changed manually. (This problem is common in other Windows applications, for example, online help, where various workarounds such as the addition of spaces are used to force the correct filing order).

For example, HTML Indexer gives the following filing order:

> Australian Society of Indexers Newsletter
> advertising guidelines
> Australian Society of Indexers
> officials

This is because the sorting is performed when the main heading and subheading are together. 'Newsletter' files before 'officials' giving:

> Australian Society of Indexers Newsletter, advertising guidelines
>
> Australian Society of Indexers, officials

Alpha bar and 'Return to top' links

HTML Indexer automatically adds an alpha bar at the top of the index with links to the letters of the alphabet (see Figure 9), and the indexer can choose to have Return to Top links between letter groups. The formatting of the index can be changed in various ways. Index titles can be added, and introductions can be put in a file called 'Top.htm' which goes at the top of the body text.

Metadata saved in pages being indexed

Embedded indexing: indexing method in which tagged index entries are inserted into document files. Tags are used to bracket blocks of text and to show headings and subheadings for index entries. Tagged index entries are not seen in the printed version, but can be compiled by software to make an index. If parts of the document are removed or rearranged the tagged index terms go with them. The index can then be recompiled to give an updated version. Embedded indexing is more time-consuming than normal indexing, but is efficient for documents that change often, or are not complete when indexing starts.

Live file: the copy of an electronic document that is currently being worked on, for example, by a writer or indexer. All changes must be made to the live file. If an indexer worked on one copy of a document, and an editor on another, the changes made by one of them would have to be incorporated into the document worked on by the other. ('Live' in a different context means that the file has been loaded onto the web and made available to users).

Indexes created using HTML Indexer are updateable because HTML Indexer embeds index entries in the documents to which they refer. Figure 11 shows the indexing metadata which is added to webpages.

Because the indexing process changes the files, the indexer must work with the live files; that is the only set of files that are currently being altered, and the ones that will eventually be uploaded to the website. If the indexer were to index one set of files while an editor worked on another set, neither would be fully

up-to-date, and work would be lost. It must also be understood that the indexing metadata is crucial for updating, and should not be removed from the documents.

Sample HTML Indexer indexes

The AusSI index (Figure 9) is an example of a back-of-book-style index to a whole website. It leads to information within the AusSI website, but does not index specific subsites, such as the *Australian Society of Indexers Newsletter*, in detail (www.aussi.org/indexer.htm). There is also a separate index to early issues of the newsletter (www.aussi.org/anl/AusSINews.htm).

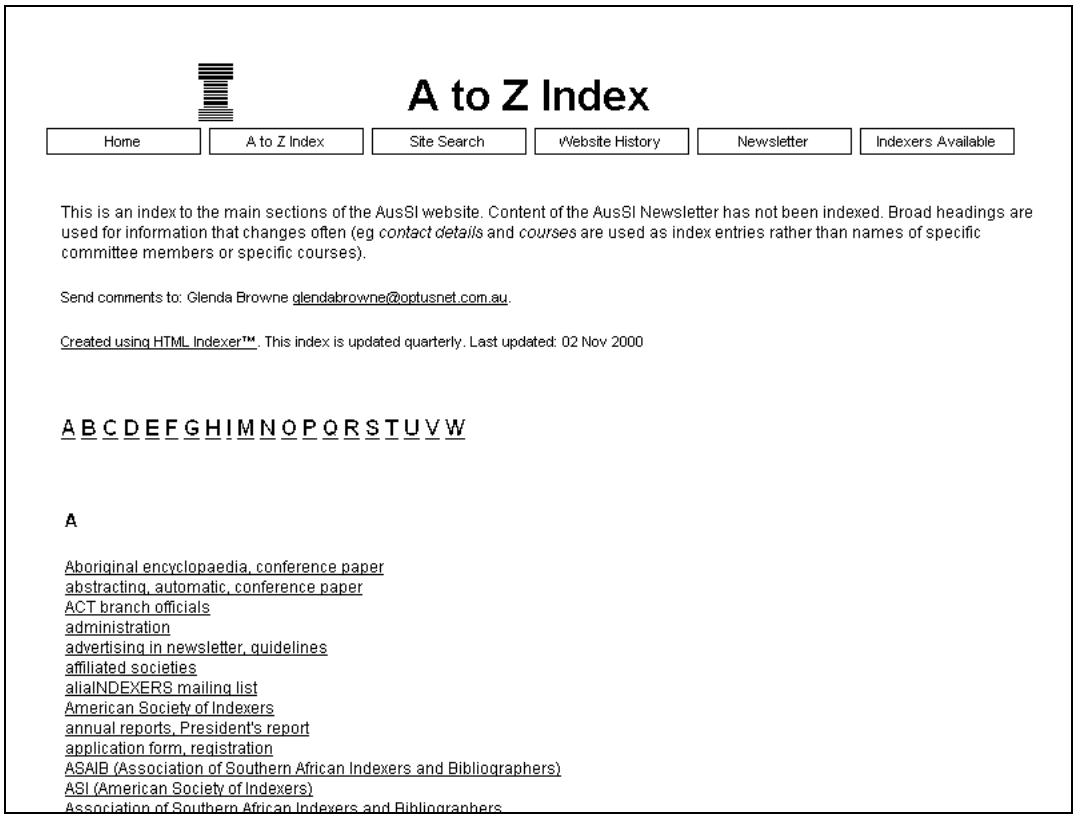

Figure 9. Index to AusSI website [35]

The first two entries in the index correspond to the first two entries in the HTML Indexer project shown in Figure 10, and the first two entries in the HTML coding for the index shown in Figure 12.

Figure 10 below shows the HTML Indexer project for the AusSI website in progress. The project files are shown on the left and index entries and their target URLs on the right. The 'conferences' folder has been selected (highlighted) in the list of folders on the left, and all of the index entries on the right apply to files within this folder. The index entries have been sorted alphabetically by clicking on the heading 'Index Entry'. Entries can also be sorted by Target URL (to compare all entries leading to the one URL while editing).

Figure 11 below shows the metadata that HTML Indexer inserted at the top of the webpage containing the Aboriginal encyclopaedia article in the conferences section of the AusSI website. For this page there are three index entries, and they

are each sorted exactly as they appear (so there is no difference between the *IndexAs* and *SortAs* wording).

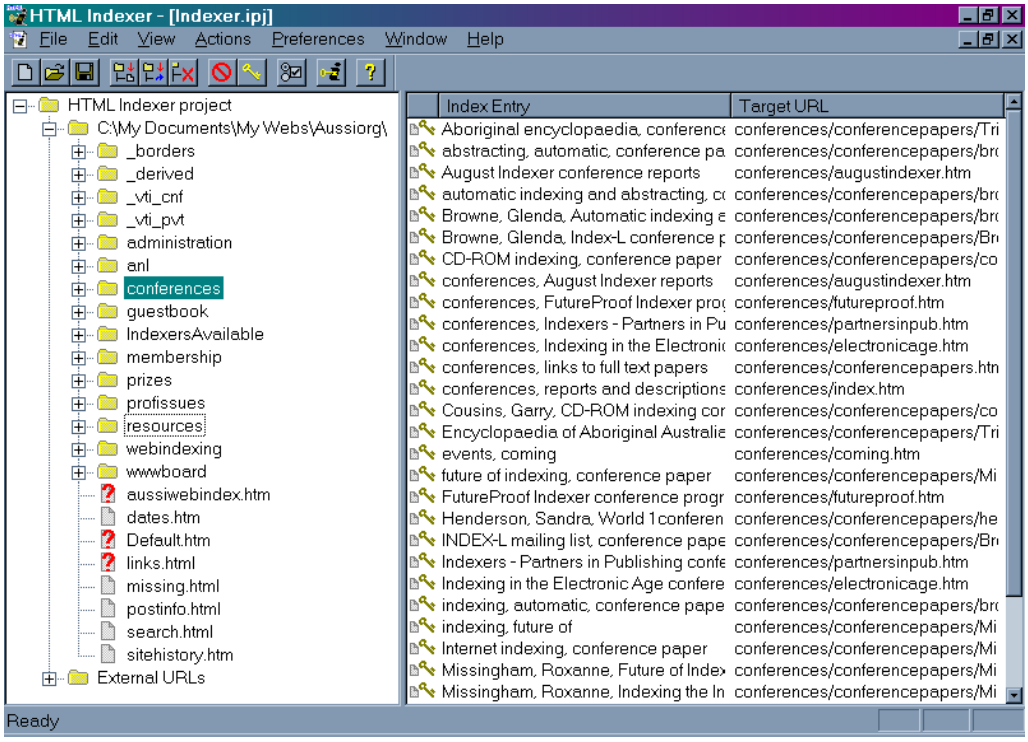

Figure 10. HTML Indexer project for the AusSI website, with project files listed at left and entries for the conferences section on right
(Reproduced with the permission of Brown Inc.)

```
<!-- HTML Indexer V3.0 - Do not edit this comment block.
 Target="#"
 IndexAs="Aboriginal encyclopaedia, conference paper", SortAs="Aboriginal encyclopaedia, conference paper", HelpAs="",HelpTitle="Untitled"
 IndexAs="Encyclopaedia of Aboriginal Australia, conference paper", SortAs="Encyclopaedia of Aboriginal Australia, conference paper", HelpAs="",HelpTitle="Untitled"
 IndexAs="Triffitt, Geraldine, Indexing the Encyclopaedia of Aboriginal Australia", SortAs="Triffitt, Geraldine, Indexing the Encyclopaedia of Aboriginal Australia", HelpAs="",HelpTitle="Untitled"
 HTML Indexer -->
```

Figure 11. Source code of the Aboriginal Encyclopaedia conference paper on the AusSI website. This metadata was embedded in the website by HTML Indexer, and is what makes the index updateable.

Figure 12 shows some of the HTML coding for the website index itself. The first two items correspond to the first two entries in Figure 10 (the 'Conferences' folder is selected in Figure 10, so it doesn't show all of the entries from the index). In the first line of code 'conferences/conferencepapers/TriffittG.html' is the address (URL) the index link leads to. It is a relative reference to the specific position on the AusSI site (www.aussi.org). *Aboriginal encyclopaedia, conference paper* is the wording of the index heading.

```
<A HREF="conferences/conferencepapers/TriffittG.html">Aboriginal
encyclopaedia, conference paper</A><BR>
<A HREF="conferences/conferencepapers/browneg.htm">abstracting,
automatic, conference paper</A><BR>
<A HREF="administration/officials.htm#ACTBranch">ACT branch
officials</A><BR>
<A HREF="administration/index.htm">administration</A><BR>
<A HREF="administration/editorguidelines.htm#9.
ADVERTISING">advertising in newsletter, guidelines</A><BR>
<A HREF="resources/organisations.htm">affiliated societies</A><BR>
<A HREF="resources/maillist.htm">aliaINDEXERS mailing list</A><BR>
<A HREF="resources/organisations.htm">American Society of
Indexers</A><BR>
```

Figure 12. Source code of the first entries in AusSI website index

Other indexes created using HTML Indexer include The Technical Editor's Eyrie shown in Figure 1 (www.jeanweber.com/Indexer.htm), Avalook at Australia (www.avalook.com.au/Indexer.htm), the Los Alamos Library Newsletter index (lib-www.lanl.gov/libinfo/news/newsindx.htm) and, of course, the index to the HTML Indexer site (finitesite.com/info/site-ndx.htm, or select from www.html-indexer.com).

Conclusion of the software section

Dwight Walker noted in 1998: 'One thing is for sure: since last year there are many more sophisticated methods for searching and indexing large amounts of online text.' [36] This development has slowed, and there have been no major new developments recently, although many more indexes now seem to be generated automatically using metadata.

Compared to the early days of hand-crafted web indexes using HTML, the current day website indexer is served by a range of tools. Which to use?

For single sourcing where the text and index already exist in print format, the best options are Sonar Activate (to create live links for PDF documents), Reworx (to prepare MS-Word documents for the web, including Tables of Contents and indexes) and HTML/Prep. Of these, HTML/Prep is the most work, but provides the most features as well. All three of these can update web-based indexes by

completely reformatting them after the indexes have been re-created in the source documents.

For single sourcing where an index is to be created from scratch and presented in a variety of formats, MS-Access or other database software provides a powerful tool that allows output of selected fields in a range of formats. It offers many of the advantages of specialised indexing software (such as the ability to sort by entry order, alphabetical order and so on) while also offering user-friendly features such as typeahead, and flexible report formats. However, it requires database skills to set up, and usually needs maintenance and tweaking as it is used. The indexer may also have to learn new skills.

To create a website index or web document index from scratch, the tool of choice is HTML Indexer. It provides in one user-friendly package the tools for creating, editing and displaying the index. Because it embeds the index entries in the pages they refer to it makes updating easy.

An alternative is HTML/Prep, which is a useful tool if the indexer wants the powerful index editing features available in a standard indexing program (CINDEX, Macrex or SKY Index) before using HTML/Prep as a last step to make the index useful for the web.

The indexer's choice of tool will depend on the specific requirements of the project, the existing data formats, and software tools and skills already available.

7. NAVIGATIONAL STRUCTURE AND TAXONOMIES

The focus of this book now moves away from back-of-book-style indexing towards alternative and complementary information access methods. The navigation structure of websites is crucial for the finding of information within them. This section discusses URLs and browsing hierarchies, as well as the depiction of structure in taxonomies and site maps.

This is followed by sections on onsite search engines (including the role of metadata and thesauri); the semantic web, search intermediation, and retrieval through external search engines or directories.

Physical structure of a website

> **URL:** Uniform Resource Locator – the address of a webpage or website. For example, *http://www.aussi.org* (also written as *www.aussi.org*).

The physical structure of a website is the arrangement of its files into directories and subdirectories. These may be stored in a single computer or scattered across one or more networks. Smaller websites usually have a relatively simple and static physical structure, while larger ones may have an elaborate and constantly changing arrangement of pages, including some that may be generated ad hoc from database content as required.

For simple systems, the physical location of a webpage can be deduced from its URL, for example, *http://www.tips-tweaks-training.com/courses/JavaScript/intro.html* is a file called *intro.html* in the *JavaScript* subdirectory under the *courses* directory on the *www.tips-tweaks-training.com* website (also known as the *domain*). Where a URL is given without a file name, the browser will add the file name 'index.htm' (or 'index.html') on the end and attempt to locate that file.

Sites are easier to maintain if the physical structure of the website mirrors the navigational structure (see below). Creating predictable URLs is also of value to users (Garrett, 'User-centered URL design', www.adaptivepath.com/publications/essays/archives/000058.php).

Many sites now use dynamically generated content with complex URLs that reflect formulas for retrieving content from the content management system. These can be recognised by characters such as "?" and "&" that indicate CGI variables (for example, webindex.com/paper?sku=123&uid=456). They can be

hard for users to read and share, and are not crawled by most search engines (hotwired.lycos.com/webmonkey/01/23/index1a_page3.html).

Navigational structure/Categorisation

Automated categorisation: the use of computer software to categorise webpages. It can be done using rule-based methods, in which the system is gradually trained, or by fully automated methods. Taxonomies for categorisation can also be created automatically.

Breadcrumb: link to all levels of the hierarchy above the current location, showing the route a searcher has taken, and the context of the current page. Breadcrumbs allow users to backtrack and to move up the hierarchy. For example, *Rhinitis>Allergic rhinitis>Perennial allergic rhinitis (Hayfever).*

Categorisation: the use of hierarchies based on words rather than notations. Each topic is allocated to a group, and that group is allocated to a more general group, and so on. Searching typically involves moving from more general to more specific topics; for example, to search for information on *children's birthday parties* you might first select the option *Celebrations*, then *Birthdays*, then *Children's parties*. Category structures are fairly arbitrary and may vary widely from one site to another; on a different site you might select *Catering*, then *Parties*, then *Children's parties*, then *Children's birthday parties*, for example. By using techniques such as double posting and cross-referencing, categorised sites can provide for access from several different directions.

Global navigation: generally applicable navigational links (for example, Search; Site Map) available from all pages of a website.

Hierarchy: a series of ordered groupings, moving from broader, general categories to narrower, more specific, ones. In a web directory you may only see one level of the hierarchy at a time. When you select a topic you are then shown the options at the next level.

Information architecture: design of the structure of information systems, particularly websites and intranets, including labelling and navigation schemes.

Local navigation: links that are specific to a section of a website, compared with global navigation which is available from all parts of a site.

Taxonomy: controlled vocabulary used primarily for the creation of navigation structures for websites. Often based on a thesaurus, but may have shallower hierarchies and less structure, for example, no related terms.

The navigational (browse) structure of a website is crucial for information retrieval by users browsing through the site. Navigation options are:

- Global navigation, which is available on every page
- Local navigation, which is applied to relevant pages

- Ad hoc (contextual) navigation, which is added as the context requires. Ad hoc navigation may be within the text or grouped at the bottom of a page for greater clarity.

The site structure is normally based on hierarchical (parent–child) categories which group related material. The categories can be represented in a taxonomy that shows broader and narrower terms (see below). Breadcrumb trails which indicate the context of the current page and allow the user to move directly to higher levels of the site are useful.

A navigational structure may be complemented by supplemental navigation, including site maps (see below) and indexes (see above). For detailed advice consult a book on information architecture such as the one by Rosenfeld and Morville [37].

Folk classifications

Putting things in categories is a basic human instinct. Marcia Bates ('Indexing and access for digital libraries and the Internet: human, database and domain factors.' www.gseis.ucla.edu/faculty/bates/articles/indexdlib.html) [38] has reported on linguistic and anthropological research into 'folk classifications', which indicates consistent characteristics across different cultures. People tend to create basic level categories of plants, animals, colours and so on in shallow hierarchies with from 250 to 800 terms, focussing on the generic level (for example, 'monkey', rather than 'howler monkey' or 'primates').

Folk classifications are also discussed by Lakoff in 'Women, fire and dangerous things: what categories reveal about the mind'. [39] The book is named after the 'balan' linguistic category of the Dyirbal tribe of Australia, which includes things such as women, bandicoots, dogs, platypus, echidna, some snakes, fireflies, scorpions, the hairy mary grub, anything connected with water or fire, and shields. Lakoff dicusses categories using the concepts functional embodiment (categories arrived at automatically) and basic level categories, as above.

Bates also discusses 'folk access' – the way people access information systems. Focus groups have found that users are positive about starting a search with some sort of classification tree, and that they often start with broader questions than the ones they really need answers to.

Categorisation

Because people tend to think in terms of categories, they are often used to provide the basic navigational structure of a website. The categories can be organised into taxonomies, which are discussed below.

Categories are hard to design and maintain because of the ambiguity of language and human subjectivity. For any set of content there are many possible groupings – the challenge is to find what works best for the majority of users. A keyword search engine can provide a good entry point into web-wide taxonomies such as the one used by Yahoo.

Categories should use one characteristic only for subdivision. For example, they can be task-oriented, audience-specific or metaphor-driven, but cannot be all of these things at the same time.

A categorisation can be established **bottom-up**, by looking at the information the website holds and working out the relationships between the different pieces of content and the categories into which it could be best organised, or **top-down**, by considering all of the theoretically possible types of information, establishing categories for them, and distributing individual topics between the categories. Top-down methods take into account business and user needs and are useful if the types of information collected could expand dramatically, that is, if existing content does not give a good indication of future content. In practice a combination of both methods will be needed (Fox, 2002, www.boxesandarrows.com/archives/rearchitecting_peoplesoftcom_from_the_bottomup.phpw).

Research for site design – free listing and card sorting

Free listing is a technique in which users are asked to name all the examples of a certain category that they know. This helps a developer determine the scope of a domain, and to become familiar with the user vocabulary in that domain. It can be a rough substitute for card sorting, but is most useful as a source of terms and domain limits for card sorting exercises (Sinha, www.boxesandarrows.com/archives/beyond_cardsorting_freelisting_methods_to_explore_user_categorizations.phpw).

Card sorting exercises in which study members sort cards with content labels into appropriate groups can be useful for determining the appropriate categories to use (www.infodesign.com.au/usabilityresources/design/cardsorting.asp; www.boxesandarrows.com/archives/analyzing_card_sort_results_with_a_spreadsheet_template.php). Open card sorts allow users to cluster labelled cards into their own categories, and add their own names for those categories. Closed card sorts required subjects to sort cards into existing categories, and are useful for examining prototype structures. Card sorts are informative, but don't present labels in the context of a site or a task, so the meaning of the labels is diminished. Affinity models can be generated from card sort results, showing the closeness of relationships between different topics.

Some people find card sorting ineffective as users seem to rush through the job grouping cards by characteristics other than those intrinsic to the topic (Donna Maurer, www.maadmob.net/donna/blog/archives/000248.html). One suggestion posted to the blog was to allow users to select the topics of interest to them personally, and just to group them.

Depictions of site design - blueprints and wireframes

Blueprints show the relationship between pages and other content components in a website, and can be used to show organisation, navigation and labelling systems. They are similar to site maps in that they act as a map to give an overview of the site. The first blueprint is normally the one showing connections from the home page. Blueprints can be drawn by hand or using diagramming software such as

Visio. Blueprints describe different levels of information, from content areas including a number of webpages, to individual webpages, to distinct content 'chunks' that stand alone even though they might be used within a page (for example: 'Contact Us' information is usually treated as a 'chunk').

Wireframes (page description diagrams), on the other hand, show how an individual page should be organised including the placement of navigation (for example: top, bottom, left or right-hand) and content elements. Individual wireframes are usually prepared for a site's main pages, while generally-applicable templates are applied to the other pages. An **HTML prototype** also links navigation from one wireframe to other wireframes.

Evaluation of site design

In 'Toward usable browse hierarchies for the web', Risden (www.microsoft.com/usability/UEPostings/HCI-kirstenrisden.doc) has reported a method for evaluating the usability of browse hierarchies. Hierarchies such as those used by Yahoo should be coherent and learnable. This is made easier if they have high within-category similarity and high between-category discriminability. That is, all the topics categorised in one group should be closely related to other topics in the same group, and clearly different from topics in other groups. Categories also need specific labels that make the content clear ('interests' and 'information' were found to be overly general). Where this doesn't happen, users tend to have difficulty choosing appropriate paths through the hierarchy, and in learning where topics should be categorised in the hierarchy.

Factors in site design – Depth versus breadth

Breadth: the number of navigation options available at each stage. A home page that provides links to 20 subsections has more breadth than one that says 'Click here to select a department'.

Depth: the number of levels in the navigation hierarchy to the most specific topics. A site where you select 'amphibians' then 'frogs' is shallower than one where you select 'animals', 'vertebrates', 'amphibians', and 'frogs' then 'green tree frogs'.

Information scent: visual and linguistic cues that indicate to a searcher whether a website has the information they seek, and help the searcher navigate to the required information.

Three-click rule: the three-click rule suggests that if a user has to click more than three times to find the information they are looking for they will give up the search.

Some designers aim to let users get to any page they want with a maximum of three clicks. On very large sites this may be impossible to achieve, as a balance has to be sought between breadth (the number of options at each stage) and depth (the number of levels to be clicked through). User Interface Engineering (www.uie.com/Articles/three_click_rule.htm) suggests that the number of clicks isn't what is important to users, but whether or not they're successful at finding

what they're seeking. More clicks are not usually a problem so long as it is clear to the user that they are heading in the right direction.

In 'Depth vs breadth in the arrangement of web links' (otal.umd.edu/SHORE/bs04/index.html) Zaphiris and Mtei describe a study which evaluated five different webpage linking strategies with varying depth and breadth. They found that with deeper web structures (that is, more links to click through), response time increased. They suggest that links should be collected and arranged so they can be presented simultaneously. This is the same recommendation as made by the NCI usability site (usability.gov). It also confirms the general approach recommended by Edward Tufte which is to include as much content as possible on a page, that is, 'Maximise the data-ink ratio' (now the data/pixel ratio).

Larson and Czerwinski also found that increased depth harmed search performance, but that a compromise with medium depth and breadth performed better than the broadest, shallowest web structure ('Web page design: implications of memory, structure and scent for information retrieval' research.microsoft.com/users/marycz/chi981.htm). They also discussed the importance of '**information scent'**, which means giving as many clues as possible to let the user know which links are likely to be useful for them to follow. This is often best done by clear delineation and useful labelling of categories, for example, the heading 'A to Z indexes for websites' provides a lot more scent than the heading 'Take your users where they want to go…fast!'.

Bernard and Hamblin ('Cascading versus indexed menu design', psychology.wichita.edu/surl/usabilitynews/51/menu.htm) evaluated what they call an 'Index menu layout', with categories and some subcategories shown in the centre of the screen (similar to the Yahoo layout), and two cascading layouts, with options listed either at the top or left side of the screen. In the cascading layouts subcategories were shown when the mouse hovered over the category. They found that users were faster with the 'Index menu layout' and that they chose this option more, although they did not explicitly express a preference for it.

Thus it seems that presenting as much information as possible on the home page and using as few levels as possible helps users find what they are looking for more quickly.

Research-based web design and usability guidelines

The NCI (National Cancer Institute, usability.gov) has established a website providing evidence-based guidelines on web design and usability issues. These are organised into groups (for example, 'search', 'titles/headings', 'navigation' and 'accessibility') and are ranked according to the quality of evidence available. For example, 'Put important information at top of hierarchy' rates four out of five for 'strength of the evidence'.

Some of the NCI guidelines are:

- Use well-designed headings to orient users and classify information on the page.

- Establish the level of importance of each category – important categories should appear higher on the page so users can locate them quickly.
- Put important information at top of hierarchy – flattening the hierarchy helps you provide more information sooner. In the case of the NCI guidelines, they have named 14 main categories and their subcategories on the home page.
- Provide printing options.

Accessibility testing

The Bobby Online Free Portal (bobby.watchfire.com/bobby/html/en/index.jsp) is a free service that allows you to test webpages to discover barriers to accessibility. Pages are tested against the US Section 508 guidelines and the W3C's Web Content Accessibility Guidelines (WCAG). Their site map (bobby.watchfire.com/bobby/html/en/sitemap.jsp) lists access features they have provided including a 'fallback style sheet' for sites that do not support CSS (cascading style sheets) and shortcut keys for quick access to important navigation links.

Taxonomies

Controlled vocabulary: a list of terms to be used in indexing (or cataloguing); often a thesaurus or synonym ring. Use of the same list by all indexers enhances consistency. Most libraries use the *Library of Congress Subject Headings* as a controlled vocabulary for cataloguing books and other library items.

Taxonomy: controlled vocabulary used primarily for the creation of navigation structures for websites. Often based on a thesaurus, but may have shallower hierarchies and less structure, for example, no related terms.

Taxonomies are controlled vocabularies used primarily for the creation of navigation structures for websites. A large company, for instance, may set up a taxonomy to direct the structure of their website or intranet. A thesaurus can be used as a taxonomy, but a taxonomy will often be shallower than a thesaurus, and have a simpler structure (for example, with no related terms or scope notes) and less strict rules for the creation of narrower terms. Whereas the focus for a thesaurus is on linguistic relationships, taxonomies focus on conceptual relationships within a particular environment: for instance, 'pay queries' might be linked to 'Financial Services' in one company taxonomy and to 'Human Resources' in another.

A sample extract from a taxonomy (designed by an enthusiast) is shown below. One dot indicates a child term while two dots indicate a grandchild. A sample thesaurus display for a single term from the same thesaurus is shown in Figure 19.

corporeal undead
. bodaks
. ghouls and ghasts
.. ghasts
.. ghouls
. liches
. skeletons
. vampires
. zombies

Figure 13. Hierarchy showing narrowing of the topic 'corporeal undead'

Lane Becker from Adaptive Path (www.adaptivepath.com/publications/essays/archives/000032.php) provides an introduction to the use of taxonomies in web-facing businesses.

Taxonomies can be bought through the Taxonomy Warehouse where users can search by category, browse a list of taxonomies and request a quote via an online form (www.taxonomywarehouse.com). Verity now sells six specialist taxonomies based on open industry standards (17 November 2003, www.kmworld.com/news/index.cfm?action=readnews&news_id=2930).

See also 'Thesauri for metadata creation' in the 'Onsite search engines' section.

Bitpipe, a site providing IT research documents, makes their thesaurus/taxonomy visible to users so they can manipulate their searches. For example, for a search on 'eCommerce' Bitpipe provides a list of document matches and a link: 'SEE: RELATED TOPICS'. Selecting 'Related topics' opens a page with detailed information on the term 'eCommerce', including its hierarchy, subcategories and *see also* references. A sample is shown below:

eCommerce
Business of Information Technology > Business Processes > eCommerce
eCommerce (247 Documents)
Subcategories:
Collaborative Commerce (16 Documents) – Related topics
eTailing (21 Documents) – Related topics
See Also:
B2B (153 Documents) – Related Topics

Figure 14. Bitpipe thesaurus presented for browsing

Metadata-driven websites

An alternative to hard-coding of navigational links to create the site structure is the generation of relevant content from a database as required. This depends on

the use of keyword metadata organised in a taxonomy or ontology (see below) to describe the content. Site creators have to decide how to describe a document rather than thinking where to slot it into a hierarchy. For a practical discussion see 'Building a metadata-based website' by Lider and Mosoiu (21 April 2003, www.boxesandarrows.com/archives/building_a_metadatabased_website.php).

See also 'Faceted metadata classification' in the 'Onsite search engines' section below for another approach that makes metadata central to resource discovery.

Automatic taxonomy generation

Automatic categorisation software is available with a number of content management and portal software systems. This software creates a taxonomy from the content on a site, and can populate that taxonomy with links to specific content items. Most software packages now allow for human checking and editing of the automatic results.

Automatic categorisation can be used to create the navigation structure of a site, and to organise search hits into logical groups.

Automatic categorisation relies on the ability of computers to process enormous amounts of information. Some approaches follow explicit rules, others learn from 'exemplar' documents provided to them, and still others use statistical clustering to group related information. The value of each approach depends to some extent on the nature of the documents being categorised, and most software packages now use a combination of approaches.

The Verity white paper 'The ABCs of content organization' (2002, www.verity.com/pdf/white_papers/MK0391a_ContentOrg_WP.pdf) shows a 'category hierarchy' (taxonomy) that was automatically created by Verity's Thematic Mapping for San Jose Mercury News articles. The sample includes the following hierarchy:

```
soviet_president
        bush
         housing
         cuomo_trade
         buchanan
         yes
```

Figure 15. Verity automatically generated taxonomy

There is widespread agreement now that you don't get optimal results using computers alone, and that the most cost-effective solution is 'cyborg categorisation' – the use of both human and computer input. A strategy might involve manual taxonomy generation or automated taxonomy generation with human review, followed by training or rule creation, then automated categorisation of documents. Human review is crucial while the system is being established, and to check that standards are maintained. Human input may also be

valuable for areas identified in user studies (or from analysis of search logs and other records) as very important and in those areas in which computers do not perform well, such as the allocation of documents into 'genre' categories ('overview', 'technical information', 'for children'). It may also be important to identify key tasks that are not well represented in the historical documents on which the system has been trained. For instance, each new legislative or regulatory requirement involves new systems and training for staff. In this way top quality can be maintained for the most important documents, and high throughput for the remainder. For a detailed discussion of automated categorisation see a series of four articles by Glenda Browne [40].

The Klarity suite of software products (www.klarity.com.au) is an Australian product that has been developed by Intology (www.intology.com.au, www.topic.com.au**)**, a subsidiary of the tSA Group. Different tools are used to categorise text (Klarity), to generate keywords and build taxonomies and thesauri (Keyword and Taxonomy Builder), and to provide an alternative virtual file structure based on a subject hierarchy for information on shared drives (Network Neighborhood).

Site maps

The basic navigational structure of a website can be visualised through a site map. This overview can quickly orient users and show them the scope of the site. Site maps are relatively easy to set up and maintain, as the information is inherent to the site. Site maps can be textual – using indents and other features to show the levels, or visual – using graphical links to show the structure.

Tables of contents

Some site maps are displayed like detailed tables of contents, showing all of the pages in the site, with the related pages next to them. Research into websites has found that showing more than one level of information at a time makes website use more effective. This is because users can orient themselves within the site and zero in on the information they want.

Jakob Nielsen has found that many users don't know about site maps, and can't find them easily when they are looking for them. He recommends keeping a map simple and labelling it clearly (www.useit.com/alertbox/20020106.html).

Some site maps present all of the information on one screen, while others allow the user to select areas for expansion. The ATO (Australian Tax Office) site map (www.ato.gov.au) first offers nine general options listing user groups (Individuals, Businesses, Non-Profit Organisations, Tax Professionals and so on). When you select one of those options the display opens up and shows narrower levels. The Volkswagen site map (www.vw.com/sitemap/indexmacnn.htm) has four categories (Cars, Culture, Commerce and About Us) with sections below that can be expanded through use of a dropdown menu.

Site maps can also be seen at the Dublin Core (dublincore.org/sitemap.shtml), Karger (content.karger.com/sitemap/index.asp) and Excite for Web Servers

(companyinfo.excite.com/site_map/index.html) sites. The Search Engine Watch site offers a site map (www.searchenginewatch.com/article.php/2148801), and also a welcome page that lists the six main sections of the site in a table with the people they are designed for (for example, site owners) and a brief description of the content (www.searchenginewatch.com/firsttime.html). The Quicken site provides a site map (help.quicken.com/support/sitemap) and also site help, which offers a smaller number of topics for access for immediate practical advice (www.quicken.com/support/help).

Chiara Fox (www.boxesandarrows.com/archives/sitemaps_and_site_indexes_what_they_are_and_why_you_should_have_them.php) praises the Bean sitemap for clearly displaying three levels accurately according to the navigation structure of the site. She castigates the Harvard University sitemap for not following the structure or wording of the site, and for mixing global navigation and footer elements.

Visual site maps

Visual site maps depict the structure of the site using written and/or pictorial tags. Some use graphics as an attractive way of providing information. For example, the QLS Group of library suppliers (www.qls.net.au) uses a signpost with differently coloured signs for different states.

Inxight uses a star tree to display their site map (www.inxight.com/map). Clicking on one part of the display moves that term to the centre, and brings in more detail around it. Jakob Nielsen found that the greatest failures come from site maps that attempt to 'lure the user into a dynamically twisting and expanding view, rather than presenting a simple, static representation of the information architecture' (www.useit.com/alertbox/20020106.html).

Classification

> **Classification**: refers to formal established classification schemes, for example, the Dewey Decimal Classification (DC) and Library of Congress Classification (LC), which use a notation to describe classes.
>
> **Notation:** code used in formal classification schemes. In the Dewey Decimal Classification the notation *993* refers to the history of New Zealand, and the notation *994* refers to the history of Australia.

Classification schemes assign numbers or letters (notations) to topics. This allows users to browse through related links, or to go directly to links on topics of interest. They can be take time to learn, but once you know the classification number for a subject of interest you can very quickly find information on a specific subject (for example, select *500* in the Dewey Decimal Classification for general items about science).

Classifications are useful for international access to information because they use numbers not letters. This means the classified information can be accessed equally by speakers of any language. (Of course, explanatory material must be in the

user's own language, but the classified content is re-usable). They are also familiar to library users. However, many people are reluctant to use them, especially as shelf location (as in a traditional library) has no significance on the web.

Another disadvantage is the cost of paying someone to implement the scheme, and the time this takes. Once implemented, it is also difficult to adjust a classification scheme – a lot of work is required to reclassify old material. For these reasons it appears that many of the earlier web experiments using classification schemes are no longer being maintained. Few of the projects listed by OCLC as being classification schemes for web resources (orc.rsch.oclc.org:6109/classification) are still current.

Dewey Classification

Some libraries and Internet information services have adopted the basic structure of the Dewey Decimal Classification as a way of organising and navigating resources on the web.

BUBL LINK (Bulletin Board for Libraries, www.bubl.ac.uk/link/ddc.html) groups selected sites according to the Dewey Decimal Classification. Users start with the ten main classes, then browse step-by-step through the classification hierarchy. The classification scheme on BUBL is complemented by alphabetical subject access (just as a library classification can be accessed through the alphabetic catalogue) and broad subject groupings as well as search, thus giving a choice of access points.

WWlib Browse Interface (Burden, www.scit.wlv.ac.uk/wwlib/browse.html) is another website organised according to the Dewey Classification, as is 'Canadian Information by Subject' from the National Library of Canada, although it is not labelled as such (www.nlc-bnc.ca/caninfo/esub.htm). Search can be narrowed from the ten main classes down to specific subjects, where users can click on links to individual websites. A sample is shown below:

000 Computers, information & general reference
001.9 **Curiosities and wonders**
Alberta UFO Study Group
BCSCC – British Columbia Scientific Cryptozoology Club
002 **Book**
Canadiana Book Collecting (Buriedantiques.com)
004 **Computer science**
ActDEN
Adventures in Science and Technology
Atout Micro – magazine des utilisateurs d'ordinateurs

Figure 16. Dewey Classification for information structure

Library of Congress Classification

Cyberstacks (www.public.iastate.edu/~CYBERSTACKS/homepage.html) is a web-wide directory organised according to the Library of Congress Classification.

Formal subject classifications

Formal subject-specific classifications are used by a number of websites including 'Materials organized by mathematical subject classification' (www.ams.org/mathweb/mi-mathbyclass.html) and EELS (eels.lub.lu.se). EELS is an engineering subject gateway, but is no longer maintained, although pre-existing content is still available. To find information on 'bridges', for example, you browse from 400 (Civil Engineering) to 401 (Bridges and Tunnels) to 401.1 (Bridges). (EELS is to be replaced with a new service consisting of harvested records automatically selected for relevance in accordance with the Engineering Index Thesaurus (EI)).

Visualisation of library classifications

> **Visualisation:** graphical presentation of information, often dependent on categorisation or clustering techniques to bring out patterns in the information.

Antarctica is a Canadian company founded by Tim Bray, co-creator of XML. It produces Visual Net software which maps various types of data to create large-scale browsable maps (www.antarctica.net). A white paper for libraries (February 2003) is available after registering at antarctica.net/request.html.

Visual Net has been used to display a library catalogue, medical bibliographic data (Pub Med, pubmed.antarcti.ca, authorisation required), financial data (Macdonald and Associates Limited, public.vn.canadavc.com/start), and a map of the Internet using Open Directory input (maps.map.net). Metaphors for the visual display include geographic maps and blocks to imply physical dimensions similar to library shelves.

The library catalogue at Belmont Abbey College, North Carolina (Belmont.antarcti.ca/help/help_2D.html) is presented using Visual Net software in an attempt to replicate the experience of shelf browsing. The library is organised according to the Library of Congress Classification (LC), and resources are grouped by classification number. The size of the blocks indicates the number of holdings for each category, and protruding blocks indicate subcategories. Selected resources (for example, new or popular ones) within each category are indicated by dots or other icons along with the title. Icons are used to indicate computer accessible resources, videos, maps and so on, and the colour within a dot indicates the type of book (print or eBook). Coloured circles of different thicknesses around the dots provide further information, such as the newness or language of the item.

Users can browse the collection by clicking on sections of interest, and can limit the holdings that are visible on the maps by typing a keyword into the filter. They can also view text-based search results by clicking the List button.

8. ONSITE SEARCH ENGINES, METADATA AND THESAURI

Search engines

Site-specific search engines give access to information within one site (for example, www.oclc.org, www.nla.gov.au searchenginewatch.com, and www.willpowerinfo.co.uk). Onsite search engines are useful if:

- there is enough content to warrant one (the site is too big to browse)
- users come to the site to search for information, rather than to perform a few simple tasks
- the site is dynamic, with daily changes in content
- the site is fragmented, with related information spread throughout the site.

Search buttons should be available on every page, and more advanced search options should be provided on a separate page. Search effectiveness often depends on metadata, which is discussed below.

> **Indexing:** often used to refer to the automatic selection and compilation of 'meaningful' words from a website into a list that can be used by a search system to retrieve pages. This list is more properly called a **concordance**. As this procedure involves no intellectual effort indexers distinguish their own work by calling it intellectual indexing, manual indexing, human indexing, or back-of-book-style indexing.
>
> **Search engine:** server that 'indexes' webpages, stores the results, and uses them to return lists of pages which match users' search queries.
>
> **Search log:** record of searches performed

User search behaviour

Jakob Nielsen found that about half of all users are **search-dominant** (that is, prefer to find information using search engines), about a fifth are **link-dominant** (they prefer to follow links and browse through the hierarchy of the site), and the rest show mixed behaviour (Nielsen, 1997, 'Search and you *may* find' www.useit.com/alertbox/9707b.html). It is hard to find up-to-date research on this subject (except for UIE below), although many people quote the original research.

Jared Spool has disputed the statement that there are search-dominant people. Research by User Interface Engineering (2001, www.uie.com/Articles/always_search.htm) found that out of 30 users performing shopping tasks, none were search-dominant (that is, always used the search engine first) although about 20% were link-dominant. What they did find was that on 21% of sites every user who visited used search – that is, some *sites* were search dominant, while on 32% of sites users only used links – that is, those *sites* were link dominant. They found that some products, such as books, were well-suited to search, while others, such as clothing, were found better using browsing. They suggest choosing one access method as a priority and putting the most effort into that one. This suggestion, however, goes against usability heuristics which suggest offering user control and flexibility, thus implying that user choice of approach is a good thing (see 'Jakob Nielsen's usability heuristics' in the section 'Users and usability').

Another important finding is that onsite searching is not always a successful strategy. Jared Spool ('Why on-site searching stinks', www.uie.com/searchar.htm) compared users searching using a site-specific search engine with those following links, and found that provision of a site-specific search engine actually reduced the quality of the results. This is because when search engines fail users conclude that the topic is not discussed on the site, rather than that they should try another search.

Search logs

Examination of existing search logs (records of searches performed) can be very useful in identifying the sort of information looked for, and the ways that users search for it. A way of learning about general search approaches is to check a live search display such as Metacrawler's Metaspy (www.metaspy.com/info.metac.spy/metaspy).

Search tools

To offer search facilities on a website you can install search engine software on your own server, or you can use a remote service which takes care of the technical configuration and server strain. For up-to-date information see the 'Search Tools Product Listing in Alphabetical Order' (23 July 2003, www.searchtools.com/tools/tools.html) or 'Search Engine Software For Your Website' (Sullivan, 28 May 2003, searchenginewatch.internet.com/resources/software.html).

Issues in the selection of a search engine are:

- the type of search needed (simple word, or based on concept)
- the platform being used
- whether the search engine accesses files from the server's file system or via http; accessing files via http increases server load

- the file types that the search engine indexes; if a site contains a lot of Adobe Acrobat PDF files, for instance, it is important to select a search engine that will index PDF files
- whether the search engine highlights the search word in retrieved pages (hit highlighting), making it easy to find the reference in the file.

Google for site search

Google can be used by users as a substitute site search engine by prefacing a search with 'site:URL'. For example, to search for the words 'automatic indexing' on the AusSI website a searcher types 'site:http://www.aussi.org automatic indexing' into the Google search box.

Scoped searches

Scoped search (search zones) is best offered as an alternative when users start to search, rather than being the only option available on a page. Search may be limited to specific categories, products, audiences or dates. If users enter a scoped search without selecting the zone they wish to limit a search to, they are often unaware that their search has not included the whole site. Results lists should indicate the zone that was searched, eg:

You searched for: course enrolment
You searched in: International students

Optimising search

Once you have installed a search engine, you still have to optimise it for your users. This includes setting the relative importance of words from titles, metadata keywords, descriptions and text in **relevance ranking**. (See also 'Metadata to enhance search' below). Depending on the software used for indexing the website the query interface (searching mechanism) can be set to do field searches (similar to traditional database searching).

You can set the standard for the information that is displayed to the user. A summary in addition to the title is useful for selecting links to follow – options are to use the first 25 or so words on the page, or to use the metadata description which can be specifically worded to be useful as a summary.

Best bets are webpages which have been identified as being particularly pertinent to certain commonly performed searches and have been manually tagged. When the terms they have been tagged with are searched for, the best bets sites are displayed at the top of the list, often in a separate section labelled 'Best bets'. The disadvantages with this approach are the time taken in tagging, and the risk that the selected best bets do not suit the user any better than the standard search would.

Boolean searching

Boolean 'and': Use of the Boolean operator 'and' in a query means that all of the terms in the query must be present in a document for it to be retrieved. For example, 'automated and categorisation' means that a document must contain the term 'automated' and the term 'categorisation'.

Boolean 'not': Use of the Boolean operator 'not' in a query means that if the search term is present in a document, that document will not be retrieved. For example, 'bear not market' will *not* retrieve a document with the sentence 'Share prices have gone down in the bear market'.

Boolean 'or': Use of the Boolean operator 'or' in a query means that any one of the terms in the query must be present in a document for it to be retrieved. For example, 'categorisation and categorization' means that a document must contain either the term 'categorisation' or the term 'categorization'.

Keyword: a) In the search engine section keywords are words that are used to search for a topic. Also called 'search terms'. b) In the metadata section, keywords are subject metadata terms.

Keyword searching: typing significant words and phrases that relate to a topic into a search engine. For example, to find information about your pet, Gerby, you might type the keywords *gerbils* and *sand rats*. If you wanted scientific information you could try the scientific terms *Gerbillus*, *Tatera*, *Taterillus gracilis* and so on. If you need to find more general information you could broaden your search with the terms *domestic animals* and *pets*. Note that a 'keyword' may actually be a phrase consisting of several words.

Keyword searching involves typing significant words and phrases that relate to a topic into a search engine. Many people do simple one or two word searches, but the Boolean operators 'and', 'or' and 'not' can be used to do more complex searches. In addition, 'advanced' search pages (for example, sunsite.berkeley.edu/KidsClick!/search.html) often allow users to specify that the whole phrase must be found, or that a search term is in a specific field such as title or URL.

Search engines use different **default operators** (that is, Boolean operators that are applied without the user typing them in). For example, Google defaults to 'and' (but now also allows searchers to use the tilde symbol (~) to search for synonyms of a term, which automatically uses 'or') while Hewlett-Packard defaults to 'or' (search.hp.com/gwuseng/boolean.html?uf=1). Inktomi uses a proprietary algorithm which basically provides search results in the following order:

- Exact phrase matches (so a search for 'automated categorisation' would require these two words to be present together in that order
- Boolean 'and' (so a search for 'automated categorisation' would require both of these words to be present somewhere in the document)

- Boolean 'or' (so a search for 'automated categorisation' would require one or other of these words to be present in the document).

This method seems to offer the best possible solution – the most relevant documents are likely to be displayed first, but the user will be provided alternatives in case nothing useful has been retrieved with phrase matching or Boolean 'and'.

Problems with recall – synonyms

> **Recall:** the proportion of relevant information that is retrieved by a search. If a search only retrieves one hundred relevant documents out of three thousand that are available, that search has low recall. If it retrieves all the available documents on the topic, it has high recall.

Search engines often have poor recall and precision because natural language is not specific – one word can have many meanings, and one meaning can be described by many words.

Recall is often poor because the language used by the person who created a web document is different to the language used by the person performing a web search. For example, the writer may have used the word 'holidays', while the searcher typed 'vacations', or the writer used 'accommodation' while the searcher typed 'accomodation'. Recall is also low when the text discusses concepts without explicitly naming them. For example, the text 'John Howard's wife' will not be retrieved by a search on 'Janette Howard'.

Recall is improved when the person searching uses a range of terms in their search combined with Boolean 'or'.

Recall can also be improved if the site creator uses a variety of terms in the document or in metadata so they are more likely to match users' search terms. Keyword metadata is an ideal way to consistently present **synonyms**, and works well for sites such as intranets. It doesn't currently work on the web as no major search engines except Inktomi pay attention to keyword metadata as they suspect that it is spam.

Intranets often use **synonym lists** (also known as synonym rings and unlimited aliasing**,** Fast, Leise and Steckel, 26 August 2003, www.boxesandarrows.com/archives/synonym_rings_and_authority_files.php) or thesauri to automatically expand searches, or to offer suggestions for expansion to users. Synonym lists should include:

- Synonyms, for example, *vitamin C* and *ascorbic acid*
- Abbreviations and full versions, for example, CD and compact disc
- Alternative spellings and word forms, *fibreglass* and *fiberglass* as well as *glass fibre*
- Common misspellings (*milennium* and *millenium*).

It is also helpful to users if you can reveal **query transformations** in the results. That is, if the user typed 'ascorbic acid' and the thesaurus identified 'vitamin C' as a synonym, 'vitamin C' should be highlighted in the retrieved text. See also 'Search transformations' in the 'Search engines' section below.

Search logs from a search engine can be used to identify searches that fail, and enable web managers to try and remedy this by changing the site or providing more information for searchers. Search logs at the web portal Lycos in 1999 showed that users typed the incorrect spelling, *millenium*, significantly more often than they typed the correct spelling, *millennium*. This suggests that it is worthwhile adding *millenium* as an extra metadata term. [41] For websites some people use alternative synonyms and spellings on different pages within the site as it looks sloppy to use two different spellings within the same page.

Alternatively, if a search retrieves no hits, search engines such as Inktomi use spell checkers to offer alternative searches (for example, saying, 'There were no hits for 'exema'; would you like to search 'eczema'?). Others provide general advice such as 'Your search retrieved no hits. Check your spelling, or try using alternative search words, or less words in your search'.

Problems with precision – false drops

False drop: document retrieved by a search but not relevant to the searcher's needs. False drops occur because of words that are written the same but have different meanings (for example, 'squash' can refer to a game, a vegetable or an action).

Precision: the relevance to the searcher of the items that are retrieved. If a search retrieves one hundred documents of which ninety-five are very relevant, that search has high precision.

Precision is often poor because the words people search for have more than one meaning (**false drops**), or are used in passing without being the topic of the document (**passing mentions**). False drops are caused by homographs that are written the same but have different meanings, for example, *briefs* (legal or underwear), *sleepers* (people asleep, earrings, spies, railway line parts), *aussi* (the French word for 'also', and the abbreviation of the Australian Society of Indexers, AusSI), and *lead* (the metal, the dog-walking device, or the action). An American correspondent to an indexing mailing list wrote that in her workplace a *committee* is a person who has been committed to an institution!

Precision can be improved by **clustering** search results into related groups and then browsing only the relevant groups. The metasearch engines Vivisimo (www.vivisimo.com) and Kartoo (www.kartoo.com) automatically group search results, thus giving a smaller set for browsing. Clustered Hits (www.clusteredhits.com) uses the taxonomy of the Open Directory Project (DMOZ) to cluster the results of keyword searches, and the Microsoft site search engine (search.microsoft.com) uses a taxonomy to group results in categories such as 'downloads' and 'product information'.

It is impossible to avoid retrieving passing mentions, but relevance ranking based on the number of times a term is retrieved should give these a low ranking, thus allowing more relevant information to show first. More complex searches using Boolean 'and' with a number of terms are also likely to limit the number of hits that are just passing mentions.

Search tips

Field searching: ability to limit a search by requiring that the search term is present in a specific 'field' (category of data) in the record. Field searching is often done with categories such as author and date that are common to most records.

Search engines can be more effective if they provide assistance to users. This can include general advice as well as behind the scenes assistance.

A website with a search engine should tell people how to plan and refine search queries, including how to broaden and narrow them. Advanced search pages should include advice on Boolean searching and **field searching** (that is, limiting a search to specific fields such as the title). This sort of information is often best provided when a search has failed and the user is receptive to information on new ways of searching.

A good example of search tips is found at the StepTwo search page (www.steptwo.com.au/search/index.html). There are seven tips on entering keywords including:

- 'Searches are not case-sensitive: "sgml" matches "SGML", "Sgml" and "sgmL"
- Searches only match whole words: "omni" will not match "omnimark"
- Add a "-" in front of words that you do not want to appear in the search results.'

Search transformations

Search engines process and modify a user's search to suggest approaches for the user to take, or to automatically provide them with what has been judged the best results.

One transformation is the use of default Boolean operators as discussed in the section 'Boolean searching' above. Another is the use of synonym lists or thesauri to search for synonyms of the search term as discussed in the section 'Problems with recall – synonyms' above.

Another modification is the use of **stemming** to expand searches to include plural forms and other word variations. This means that a search for 'print' also retrieves documents containing the words 'prints', 'printing' and 'printer'. In some cases stemming is limited to words from the same part of speech (for example, a search for a verb will only stem to create other verbs); in other cases it depends on the characters at the start of the word. Stemming is an advantage in most cases as it

means that searchers and metadata creators don't have to worry about inclusion of singular and plural forms. It has also been found that most users expect some form of stemming (Muramatsu and Pratt, 'Transparent queries: investigating users' mental models of search engines, 2001, ai.rightnow.com/colloquium/papers/search_engine_mental_models.pdf) . Experienced searchers, however, often don't like stemming as it means they can't make a distinction between pairs of terms such as 'trust' and 'trusts' and 'electronic journals' and 'electronics journals'.

Muramatsu and Pratt also found that users 'have multiple misconceptions about how search engines process their queries' and that without feedback on query transformations they create erroneous mental models of search engine operation. They recommend increased transparency, but acknowledge that it can be difficult to provide complex information to users without adding to search complexity. One simple approach is hit highlighting, in which the terms which resulted in a document being retrieved are highlighted in the results list.

Metadata to enhance search

Keyword: subject metadata terms. In the search engine section 'keyword' is used to mean words that are used to search for a topic.

Metadata: structured data about data, which may include information about the author, title and subject of web resources. Metadata is added in the <HEAD> section of a webpage or is stored in a database. It is available for searching but is not displayed in a browser.

Metadata is structured data about data, and is similar in concept to author, title and subject information in traditional library catalogues. Keyword, title and description metadata are important for the retrieval of documents by search engines as they provide additional, appropriate and consistent keywords for searching.

Title and author metadata can often be generated automatically from webpages, but subject keywords (words describing the subjects of webpages) and descriptions (abstracts or summaries of the content for display by search engines) usually have to be generated by people, making them more expensive to implement.

Information about metadata can be found at the Meta Matters site (dcanzorg.ozstaging.com/mb.aspx) and at Search Engine Watch (searchenginewatch.internet.com/webmasters/meta.html).

While keyword metadata is now of little importance to web-wide search engines, it is still important in site search (Goodman, 'Google uses meta tags sparingly, but should you?, www.traffick.com/article.asp?aID=105), where search engines can be set to take it into account, and spamming is of little or no significance.

See also 'Metadata-driven websites' in the section 'Navigational structure/Categorisation' and 'Problems with recall – synonyms' above.

Metadata tags

<HEAD> section: The <HEAD> section of an HTML document is placed at the top of the page between an opening tag <HEAD> and a closing tag </HEAD> and contains metadata about the document, not the content that will be displayed on the page. It is followed by the <BODY> section.

Tag: a piece of text that describes the semantics or structure of a unit of data (element) in HTML, XML or other markup language. Tags are surrounded by angle brackets (< and >) to distinguish them from text.

Metadata tags are listed in the <HEAD> portion of an HTML webpage, or in a database. Metadata keywords are listed in any order, with commas between them. Metadata titles are those used in the title tag and displayed in the banner at the top of the webpage. Metadata descriptions are summaries (abstracts) of the webpage that may be displayed by search engines in lists of retrieved hits. These tags are written in the following format:

```
<HEAD>
<title>E-book indexing</TITLE>
<META NAME="keywords" CONTENT="e-books, electronic books, eBooks, e-book indexing, electronic book indexing, eBook indexing, embedded indexing, indexers, information access">
<META NAME="description" CONTENT="Indexing of electronic books for better searching. E-books can be used on a number of handheld devices. "<DESCRIPTION/>
</HEAD>
```

Figure 17. Standard HTML encoded metadata

Metadata standards – Dublin Core

Metadata standards such as the Dublin Core (purl.oclc.org/dc or dublincore.org) have been developed by interested groups to try and enforce consistency on the web. The use of standards means that metadata can be shared with other systems using the same standard. Dublin Core is an international metadata standard that defines 15 elements to be used to describe the content and authors of web resources. The Dublin Core standard can be downloaded from the National Information Standards Organization (NISO) site (www.niso.org/standards/index.html).

Dublin Core elements fall into three groups which roughly indicate the class or scope of information stored in them:

- Elements related to the content of the resource (title, subject, description, source, language, relation and coverage)
- Elements related to the resource as intellectual property (creator, publisher, contributor, rights)

- Elements related to the instantiation (the electronic or physical manifestation) of the resource (date, type, format, identifier).

Only the subject-related elements are discussed further here.

```
<HEAD>
<title>E-book indexing</TITLE>
<meta name="DC.Subject" scheme="GBJJ" content="e-books">.
<meta name="DC.Subject" scheme="GBJJ" content="information access">.
<meta name="DC.Description CONTENT="Indexing of electronic books for
better searching. E-books can be used on a number of handheld devices.">
</HEAD>
```

Figure 18. Dublin Core encoded metadata

Metadata standards – AGLS

The AGLS (Australian Government Locator Service) metadata standard is maintained by the National Archives of Australia. It contains 19 elements, and is based on the Dublin Core element set, with the addition of:

- Three elements related to the resource as intellectual property (function, audience and mandate)
- One element related to the instantiation of the resource (availability).

Six of the AGLS elements must be present in any AGLS record. These are: Creator, Publisher, Title, Date, Function or Subject, Identifier or Availability.

More information is available in the 'AGLS Metadata Element Set' (www.naa.gov.au/recordkeeping/gov_online/agls/metadata_element_set.html). This version uses qualifiers to give extra information, for example, to name the thesaurus from which subject headings (keywords) have been selected.

Metadata standards – EAD

The Encoded Archival Description (EAD, www.loc.gov/ead) metadata standard has been developed for encoding archival and manuscript collections. For an introduction to archiving practice and the importance of EAD see 'Encoded Archival Description: an introduction and overview' by Daniel Pitti (*D-Lib Magazine* v.5 n.11 November 1999, www.dlib.org/dlib/november99/11pitti.html).

Metadata guidelines online

There are some extensive organisational metadata guidelines on the web. These give a good idea of general principles, and specific decisions made by individual organisations. For example, Edna (Education.au, www.edna.edu.au/metadata) provides guidelines as well as online training (www.educationau.edu.au/edna_cataloguing).

The Justice Sector Metadata Standard (info.lawaccess.nsw.gov.au/lawaccess/lawaccess.nsf/print/jsms) is designed for organisations providing legal information on the Internet. Schemes include: LIAC (DC.Subject element, for example, 'Emergency services'); JSMS (categories, for example, 'Act', 'Factsheet'); JSMSType (for example, 'document', 'image'); and JSMSAudience (for example, 'all', 'low income earners', 'transgender').

Tools for creating metadata

There are two types of tools on the web that can be used for creating Dublin Core metadata:

Metadata editors provide a template for entering new metadata content, and then place the content into HTML <META> tags. Examples of editors are:

- Nordic Metadata Project (www.lub.lu.se/metadata/DC_creator.html)
- Reggie (metadata.net/dstc). Reggie outputs in HTML and RDF formats
- Metabrowser (metabrowser.spirit.net.au).

Metabrowser can also be used to generate an alphabetical index. See 'Indexes generated automatically from metadata' in the 'Website indexing software' section.

Metadata generators extract metadata from existing HTML-encoded documents and place it into the HTML <META> tag. An example of a generator is:

- DC-dot (www.ukoln.ac.uk/metadata/dcdot).

Blocking access to sites with metadata

Metadata can be used to block access to sites as well as to enhance it. Sites that are considered unsuitable for certain classes of users (children, employees during working hours) can be blocked by a browser on the basis of the information they contain.

'Rating systems' such as those developed by ICRS allow users to block websites based on rating set by their authors. Criteria used to block or allow sites can be based on any or all of the criteria: Language; Nudity; Sex; and Violence. This system can be switched on and adjusted in a standard browser (MS-Internet Explorer or Netscape Navigator) and then password protected to prevent unauthorised changes. The person setting up the system may decide to block all unrated sites (opt-in) or allow all unrated sites (opt-out). In practice very few site designers can be bothered researching and allocating ratings for their sites, and the ratings system has fallen into disuse. The ICRS home page, with details of its rating criteria, can be viewed at www.icra.org.

Proprietary software such as CyberSitter and NetNanny uses a ratings system in conjunction with: a) actively examining webpages (and other Internet communications) before downloading to detect 'unauthorised' words and phrases; b) comparing site addresses with ongoing analyses of 'good' and 'bad' sites maintained online by the software suppliers; c) allowing users in authority to block or allow sites of their own choice.

Most major search engines, including Google and Yahoo, now offer a 'safe' search setting which will supposedly screen out sites with objectionable content. Although this is a handy feature it can be easily turned off, and it suffers from the same problems in blocking legitimate sites and allowing unwanted ones to come through as any other text-based filtering system.

Thesauri for metadata creation

Controlled vocabulary: a list of terms to be used in indexing (or cataloguing); often a thesaurus of synonym ring. Use of the same list by all indexers enhances consistency. Most libraries use the *Library of Congress Subject Headings* as a controlled vocabulary for cataloguing books and other library items.

Thesaurus: a structured list of approved subject headings (preferred terms) showing the relationships between them. The relationships include broader (parent) terms, narrower (child) terms, and other related terms. The thesaurus may also show terms that are *not* to be used in indexing (nonpreferred terms) with references to the terms that should be used instead (for example, automobiles, use cars).

A thesaurus can be used to maintain consistency between metadata contributors. It provides:

- A list of terms to be used in indexing (preferred terms, keywords, descriptors, or subject headings)
- Synonyms of those terms (nonpreferred terms or variant terms), with references to the preferred terms
- A hierarchical structure (parent–child relationships) showing broader terms (BT) and narrower terms (NT)
- Related terms (associative relationship).

In traditional database indexing only preferred terms are used as indexing terms. Indexers and searchers consult an approved thesaurus to find the appropriate term to use. For websites or intranets, access is usually provided on all synonyms (that is, preferred and nonpreferred terms). This can be done by including all of the preferred and nonpreferred terms in the metadata, or by putting the terms in a synonym ring so that entry of any one term searches for all the other terms as well (See 'Problems with recall – synonyms' above).

As well as being used for consistency in indexing, a thesaurus can be used by searchers for term selection and query expansion and refinement. For example, after consulting a thesaurus, a searcher may choose to search on a narrower term than the one they had previously thought of. A thesaurus may also be used to further business goals using related terms to enhance cross-selling (for example, 'you bought some paint, would you like to see our catalogue of paintbrushes?').

The thesaurus display for a single term is shown below. A sample of the hierarchical output from the same thesaurus is shown in Figure 13.

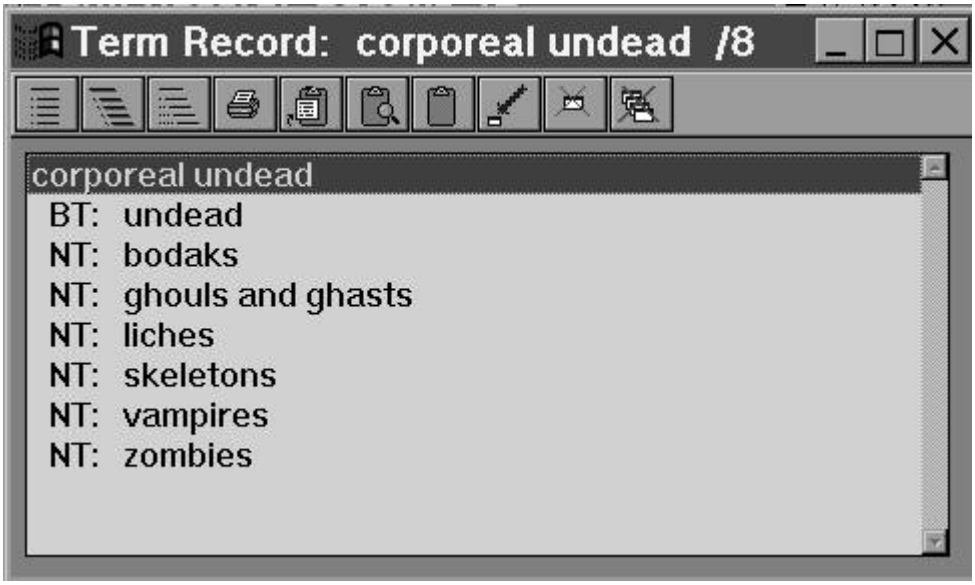

Figure 19. Term record from MultiTes for 'corporeal undead'

Linked thesauri on the web

A thesaurus can be used most fully when it is linked to the resources it describes, not just presented for viewing. Because people tend to search broadly access to a thesaurus can help them narrow their searches, but as yet there are relatively few linked thesauri on the web.

The **PICMAN Topic Thesaurus** (www.sl.nsw.gov.au/picman/subj.cfm)[42] is used for indexing and searching for photos held at the State Library of NSW. It contains 15,000 topic terms used to index the 314,000 records in the PICMAN database. Users who simply type a search term using Basic search are taken immediately to the results of their search. Users who select the thesaurus option (by clicking on 'Subject') and type a word that is in the thesaurus are taken to relevant information from the thesaurus, and can choose broader, narrower and related terms as appropriate. [43] Selection of an index term leads to a display of pictures with descriptive information. The thesaurus is detailed, allowing for specific searches. Figure 21 shows the results of a search for *barbecues*.

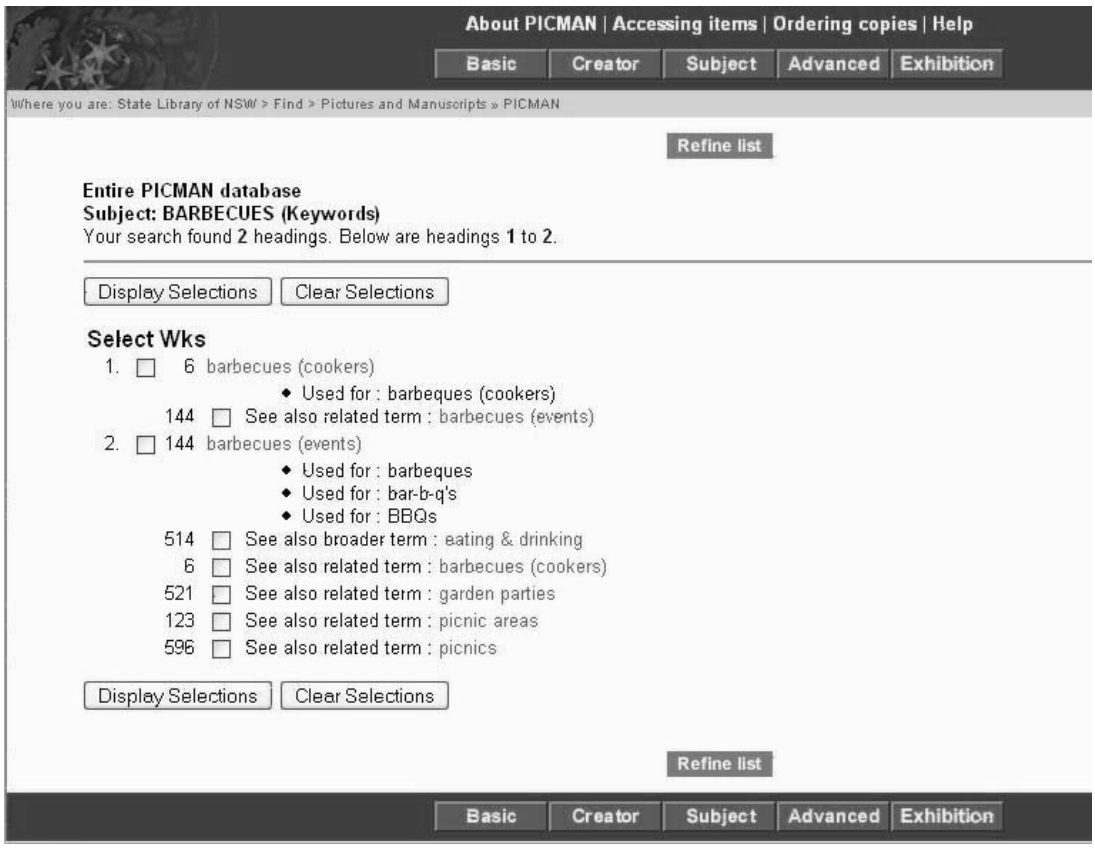

Figure 20. Subject search for 'barbecues' in PICMAN database (Reproduced with the permission of the State Library of NSW)

The PICMAN thesaurus, with enhanced functionality, has been made freely available on the web (www.picturethesaurus.gov.au) as the **Australian Pictorial Thesaurus.** Its aims are to 'promote the use of Australian headings for the indexing of documentary images and other suitable collections'[44] and to become a national standard for the indexing of Australian pictures. Users will be able to add candidate terms, which will remain distinct until approved as thesaurus terms and integrated into the structure.

HealthInsite is another site with a linked thesaurus (www.healthinsite.gov.au/search/thesaurus_levels.cfm), although again, thesaurus search is not the default option, but has to be selected by the user. Much thought has gone into the development of this system, which has been discussed at conferences and is well-documented. [45]

TAGS – Thesaurus of Australian Government Subjects

TAGS – Thesaurus of Australian Government Subjects (www.noie.gov.au/projects/egovernment/better_practice/tags/tags.htm) describes Commonwealth information and services at a high level, providing broader terms that can be supplemented by other thesauri as needed. TAGS is the preferred tool for Commonwealth government agencies to use in AGLS subject description (see above), and includes all APAIS (Australian Public Affairs Information Service Thesaurus, www.nla.gov.au/apais/thesaurus) terms that are within its scope.

AGIFT

AGIFT (Australian Governments' Interactive Functions Thesaurus, www.naa.gov.au/recordkeeping/gov_online/agift/extract.html) lists terms to be used in the function element of AGLS. There are 24 broad terms, with narrower terms organised alphabetically within each level.

Library of Congress Subject Headings (LCSH)

LCSH (Library of Congress Subject Headings, print purchase details at lcWeb.loc.gov/cds/lcsh.html) is the most widely used controlled vocabulary in the world. It has been used mainly for the cataloguing of library resources. FAST (O'Neill and Mai Chan, August 2003, 'FAST: faceted application of subject terminology', www.ifla.org/IV/ifla69/papers/ 010e-ONeill_Mai-Chan.pdf) is a project that aims to simplify LCSH to make it more generally applicable in a web environment. It will be organised into eight distinct facets: topical, geographic (place), personal name, corporate name, form (type, genre), chronological (time, period), title and meeting name. The topics themselves will not be divided into facets (such as materials, processes and so on, as would be done in a fully faceted classification – see below).

If this project is successful it could mean a relatively quick, cost-effective means of developing a generally-applicable, easy-to-use list of terms. On the other hand, it might mean building the future on a system that was not designed for the web, and has grown increasingly inconsistent over the last hundred years.

Thesaurus construction

Keyword creators may use or adapt a pre-existing thesaurus, or create one from scratch. For corporate intranets a new thesaurus is usually created to cater for the specific products, interests, and terminology of the company. Specialist thesauri can also be bought from the Taxonomy Warehouse and providers such as Factiva (www.factiva.com).

To create a thesaurus may take anything from one month to three years, depending on its size and complexity, and updating each edition can take up to six months. There are several software packages that can assist with the clerical aspects of thesaurus construction, including the creation of reciprocal entries and error checking. Two packages commonly used in Australia are MultiTes (www.multites.com) and TermTree (www.termtree.com.au). A newer product which is focussed on web delivery is WebChoir (www.webchoir.com). The Willpower site (www.willpowerinfo.co.uk) has a useful checklist to use when selecting thesaurus management software.

Further information

Mike Middleton's site at QUT (sky.fit.qut.edu.au/~middletm/cont_voc.html) and Mary Sue Stephenson's resources (www.slais.ubc.ca/resources/indexing/index.htm) [46] provide links to information about thesauri and thesaurus creation software. The Bayside Indexing site (www.bayside-indexing.com /milstead_articles.html) now hosts Jessica

Milstead's article on thesauri in the online environment. David Batty has an article on thesauri online ('WWW – Wealth, weariness or waste', www.dlib.org/dlib/november98/11batty.html). Tim Craven provides a 'Thesaurus construction tutorial' at instruct.uwo.ca/gplis/ 677/thesaur/main00.htm, and the ANSI/NISO Z39.19 – 2003 standard 'Guidelines for the construction, format and management of monolingual thesauri' is available free on the web at www.niso.org/standards/standard_detail.cfm?std_id=518. It is currently being revised to make it more appropriate for web-based thesauri. A number of websites list other online thesauri. [47]

Faceted metadata classification

Breadcrumb: link to all levels of the hierarchy above the current location, showing the route a searcher has taken, and the context of the current page. Breadcrumbs allow users to backtrack and to move up the hierarchy. For example, *Rhinitis>Allergic rhinitis>Perennial allergic rhinitis (Hayfever).*

Faceted metadata classification: breaking subjects into standard component parts (facets) and presenting these to users as search options. A topic such as wine might be divided into the facets such as *country of origin, variety* and *price*. In the best faceted search systems the user is provided with feedback about the number of terms retrieved at each stage.

Faceted classification depends on separating subjects into their component parts, and allowing access through one or more of those parts according to user needs. It is considered to be an ideal approach for combining the best of browsing and searching online. It works particularly well for online retrieval as facets can be combined post-coordinately (that is, while searching), rather than having to be combined in a set citation order for shelf arrangement. Faceting is easiest to implement with uniform collections (for example, wines and recipes) but may have the most impact in complex multidisciplinary environments.

Faceted classification often works well for ecommerce applications, where specific attributes are applicable to all products. For example, someone might want a toy for less than $10, or for a newborn baby, or with Harry Potter on it. They can find these by searching for the required values in the facets *cost*, *age*, and *character*. Other common facets in business applications are topic, product, document type, audience, geography, price and ratings.

While some websites use facets for search only, or for navigation only (details in the survey by Web Design Practices, October 2003, www.webdesignpractices.com/navigation/facets.html, the best of them allow users to combine searching and browsing. At all stages **breadcrumbs** (a chain of links from general to specific) are displayed showing the path they have taken, and allowing them to backtrack if needed. When browsing the user can refine searches by drilling down a hierarchy to more specific terms, or by adding values from different hierarchies. For example, users can narrow a search by drilling down from 'vertebrates' to 'mammals' to 'whales' in the *animals* facet, or by

selecting 'vertebrates' from the *animals* facet and then selecting 'Australia' from the *place* facet. Because available options are presented to the user at each step, they only have to recognise the term of interest, not decide what search term to search on. Faceted classification can also be used to refine searches according to generally applicable facets such as format, user appropriateness, type of material (for example, overview), genre, time and place. As the user refines their search, the display shows the number of hits at each point, letting the user know whether they need to further refine their search, or whether the number is small enough for them to look at them all.

This general approach to faceted searching on the web has been informed by the **Flamenco Search Interface Project** (FLexible information Access using MEtadata in Novel COmbinations; bailando.sims.berkeley.edu/flamenco.html) at the University of California, Berkeley. In 'Finding the flow in Web site search' (www.sims.berkeley.edu/~hearst/papers/cacm02.pdf), Marti Hearst and others discuss a project implementing retrieval for architectural images using combined browsing and searching. In usability tests they found that users were satisfied with the system, despite the presence of unfamiliar features, which often deter users. Ideas from this research are now found in the commercial products Endeca (www.endeca.com), Siderean Seamark Server (previously bpallen Teapot; www.siderean.com), and i411 (www.i411.com).

There is a standard for faceted classifications which is compatible with the topic map standard. Details of the XFML standard are at xfml.org.

Two areas in which faceted classification may bring great benefits are inter-disciplinary studies and the search for information in specific formats, or from a specific point of view. Inter-disciplinary studies such as 'the effect of computers on education' versus 'the effect of education on computers' can be difficult to distinguish in simple keyword searches. Faceted classification allows clarification of the roles of each of the disciplines.

Similarly, it can be difficult to separate search results that retrieve a specific type of document, with sites that are *about* that type of document, for example, information on indexes, best bets, or classifications. This has been confirmed in automatic classification studies by Dumais and Chen (research.microsoft.com/~sdumais/sigir00.pdf), who found that computers can identify subjects much better than they can identify genres or predicted users (for example, 'for children'). The Kosmoi page on classification highlights this problem. It is a webpage *about* classification, but a column on the left headed 'Amazon.com' lists supposedly related books such as 'Paterson First Guide to Caterpillars of North America', 'DSM-IV made easy: the clinician's guide to diagnosis' and 'Crime classification manual' (dixionary.com/Reference/Library/Classification). These items *are* classifications, but are not *about* classification.

Facet Map

You can make your own simple faceted classification using FacetMap (facetmap.com/index.jsp), a software package that allows you to create and test

your own faceted classification scheme on the web. The site offers '3-minute concept info' which provides information while leading you through a demonstration of FacetMap.

Figures 21 to 23 below show a faceted classification in action.

Faceted classifications on the web

A number of websites now use faceted organisation. Examples include:

- Annotated Wordnet (www.siderean.com/wordnet17.jsp)
- CMS Faceted Product Directory (www.cmsreview.com/timelines/ShopFeatureDirectory.html)
- Epicurious (www.epicurious.com). Start at eat.epicurious.com/recipes/enhanced_search/index.ssf?/recipes/enhanced_search/index.html or go to eat.epicurious.com/recipes and select 'Enhanced Search'
- Meta Matters (dcanzorg.ozstaging.com/mb.aspx)
- Online proceedings of the DC- 2002 Dublin Core conference (www.siderean.com/dc2002.jsp). The proceedings can be searched by: Subject by category (for example, 'activities', 'organizations'); Subject (for example, 'Dublin Core', 'RDF'); Creator (for example, 'Jane Greenberg'); Time of event (for example, '14 October'); and Type of event (for example, 'Plenary Session')
- Tower records (www.towerrecords.com). Users first search by typing in a keyword or filling in a form with various categories. Later pages allow for refinement of search according to facets which are displayed at the left of the page. These facets include genre, feature (such as 'boxed sets' or 'in stock"), price and artist. Results are grouped into useful categories, for example, a keyword search for 'Mozart' retrieves hits with Mozart in the composer name, the Ensemble name, and so on
- Langemarks Cafe (www.langemark.com/taxonomy_search/blog) offers 'taxonomy search' which is a simple faceted search system in which you can select 'Subject, 'Content Type' or 'Feeling' from a dropdown list. This site indicates some of the problems with search systems that use facets but don't indicate the number of hits for each option. It is particularly problematic as some of the dropdown options have no content attached at all (for example, a search for all subjects and content types with the feeling 'Silly' yielded no results).

Figure 21. Facet map starting page

Figure 22. Facet map after selecting 'French'

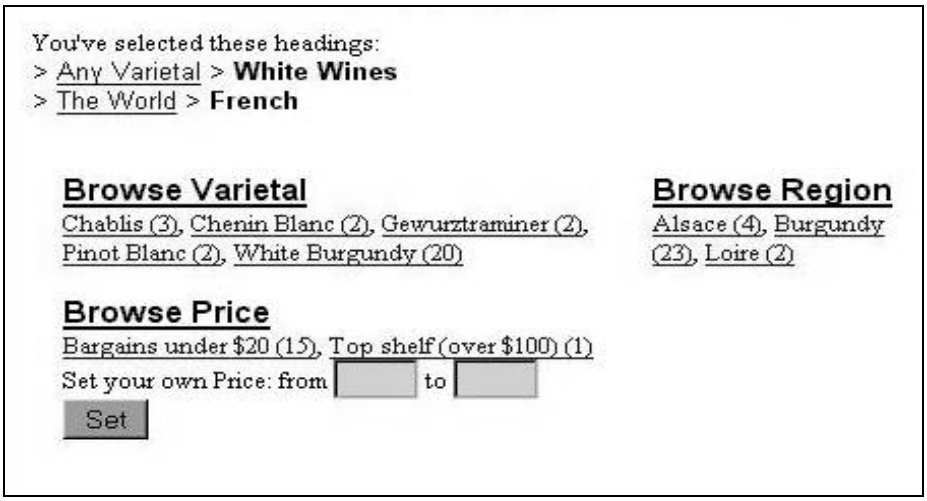

Figure 23. Facet map after selecting 'French' and 'white wines' (reproduced with the permission of facetmap.com)

9. SEMANTIC WEB – RDF, DAML+OIL AND ONTOLOGIES

Semantic web

RDF and RDF Schema

Ontologies

Topic maps

Semantic web

Markup language: a way of depicting the logical structure or semantics of a document and providing instructions to computers on how to handle or display the contents of the file. HTML, XML and RDF are markup languages. Markup indicators are often called tags.

Semantic web: project of the W3C in which automated methods based on machine processable metadata are envisaged to replace much human searching of the web. Relies on ontologies, XML and RDF.

Semantics: meaning. If a computer understands the semantics of a document, it understands the meaning, rather than just interpreting a series of characters.

URI (Uniform Resource Identifier): unique identifier of the location of a resource. In many cases the URI will be a URL (that is, webpage address, eg: http://www.aussi.org).

The semantic web depends on precise description of data to enable machine-processable transactions – since computers find it difficult to understand the world, humans are now trying to describe the world in ways that are easier for computers to understand. Tim Berners-Lee, creator of the web and now with the W3C (World Wide Web Consortium) considers the semantic web to be the next step in the optimal use of the web as a tool for task performance as well as resource discovery.

For the semantic web to work it requires the following ('Introduction to semantic web technologies', www.hpl.hp.com/semweb/sw-technology.htm):

- A global naming scheme (URIs – Universal Resource Identifiers. In most cases URLs serve this purpose)
- A standard syntax for describing data (RDF – Resource Description Framework). RDF lets you assert facts, for example, 'person X is named Shannon'
- A standard method of describing the properties (attributes) of that data (RDFS – RDF Schema and other markup languages including DAML+OIL). RDFS

lets you describe vocabularies and use them to describe things, eg 'person X is an AustralianResident'

- A standard method of describing the relationships between data items (ontologies). An ontology lets you establish relationships between vocabularies, for example, 'visitors' in schema A are the same as 'users' in schema B
- A way to support trust and security.

Examples of the tasks that a computer might perform for a user include selection of and payment for goods, and booking medical appointments based on the computer's knowledge of the person's health condition, regular medical practice, health fund requirements, specialist needs, schedule availability, transport timetables and so on. For this to work a lot of information has to be documented, including ephemeral things like a person's priorities – for example, is the choice of doctor or the time of the appointment more important?

In a criticism of the hype about the semantic web, Clay Shirky (7 November 2003, 'The Semantic Web, syllogism, and worldview', www.shirky.com/writings/semantic_syllogism.html) points out that the only tasks that computers will be able to perform for us via the web are those that depend on **syllogisms**. The canonical syllogism is:

Humans are mortal Greeks are human Therefore, Greeks are mortal

with the third statement derived from the previous two.

The semantic web specifies ways of stating these kinds of assertions on the web so that they can be combined to discover things that are true but not specified directly. Shirky states: 'This is the promise of the Semantic Web – it will improve all the areas of your life where you currently use syllogisms. Which is to say, almost nowhere.' One problem the semantic web faces is the fact that people speak in generalisations which they are (often) able to apply correctly given context and experience that cannot be replicated in machine-readable form.

RDF (Resource Description Framework) and RDF Schema

Namespace: a closed set of names or a place where a schema (set of names) is stored. Namespaces are identified via a URI (for example, a URL) and are a mechanism to resolve naming conflicts. Within a given namespace all names must be unique, although the same name may be used with a different meaning in a different namespace.

RDF: a formal data model from the W3C using XML for the description of web resources using machine readable metadata. It has potential for use in the semantic web.

RDF schema (RDFS): defines a set of metadata properties (for example, 'Creator') that can be associated with resources.

Schema: a description of the structure and rules a document must satisfy for an XML document type. Includes the formal declaration of the elements that make up a document.

RDF is an extension of XML, and was developed under the auspices of the W3C (World Wide Web Consortium). RDF is designed to provide for the expression of semantic information – information about things. At the core of RDF is the notion of a resource description. A resource is something: it could be a document, a book, a company or a person, or any other object or concept of interest. A description is a set of information which represents the resource. The information is obviously selected to be of value to users or searchers; thus your resource description would probably include information about your job and phone number but not whether you can waggle your ears.

Information is given in a resource description via defined properties. Only certain data count as properties, and these are specified through a list of valid property-types, called a schema. Property-types should be logically and practically appropriate to the type of resource. Thus 'weight' would be a valid property type for describing a vehicle, and 'CEO' would be a valid property type for describing a company, but not vice versa. Schemas are stored in namespaces.

The properties in a resource description are assigned values, of which the simplest are just strings of text – 'Harry Smith', '1200 kg' and so on. The *value – property-type – resource* triad is typically represented in plain English as '*value* is the *property-type* of *resource'* or 'the *property-type* of the *resource* is *value*'; for example, 'Harry Smith is the CEO of Snibbo Enterprises'; 'The weight of the Toyota Camry is 1200 kg'. These triples are assumed to be logically independent assertions which are mutually compatible: RDF as such does not attempt to check on the logical consistency of what is asserted, although this is a possibility within the system.

Values can be links to resource descriptions: thus 'Harry Smith' may have its own collection of values – 38 years old, male, born in Toronto – which may in turn be called on by the person who wants to find out about Snibbo Enterprises. In plain English we can think of these as adjectives or subordinate clauses: 'Male, Toronto-born Harry Smith, 38, is the CEO of Snibbo Enterprises'. In RDF this involves a cross-referential structure, in this case from a collection of information about companies to a collection of information about people. This information can also be shown diagrammatically as a network of connections between nodes.

RDF syntax is a simplified form of XML. In RDF no identifier (resource description) has more than one property, and all identifiers in a collection have the same property set. As in XML, each identifier must be declared before use, and in RDF this is done by providing a URI (Unique Resource Identifier) which identifies that resource. URIs are obviously related to the URLs used to access websites, but RDF does not put any constraints on them other than that they are not duplicated: thus a URI could in principle refer to a library shelf list of physical books, a company's employee database, or any other unique way of locating a

resource. If the resource of interest happens to be a webpage, the URI is identical with its URL.

For an alternative description of RDF, read 'An introduction to the hard Semantic Web …in simple Haiku' (infomesh.net/2002/swhaiku). [48] A sample follows:

> RDF is made
> of triples—only triples,
> also called "statements".
>
> Three **terms** make triples:
> just binary relations,
> never more nor less.
>
> The triple-nature:
> **subject**s first, then **predicate**,
> followed by **object**.

See also 'Topic maps' below and 'XML' in the 'File formats' section.

Ontologies

> **Ontology:** specification of a conceptualisation of a knowledge domain. An ontology is a controlled vocabulary that describes objects and the relations between them in a formal way, and has a grammar for using the vocabulary terms to express something meaningful within a specified domain of interest. The vocabulary is used to make queries and assertions. Ontological commitments are agreements to use the vocabulary in a consistent way for knowledge sharing. Ontologies can include glossaries, taxonomies and thesauri, but normally have greater expressivity and stricter rules than these tools. A formal ontology is a controlled vocabulary expressed in an ontology representation language (for example, OWL). [49]

An ontology is a specification of a conceptualisation of a knowledge domain and provides a consistent way of describing the relationships between things. In its broadest sense, the term is used to include a range of controlled vocabularies including glossaries, taxonomies and thesauri. However, in the stricter sense an ontology is a controlled vocabulary that can be expressed in an ontology representation language and can be used for automated reasoning support (can be understood by a machine and used to make logical inferences). These principles are important in the development of the semantic web.

The mandatory requirements of an ontology according to Deborah McGuinness ('Ontologies Come of Age', 2001, www.ksl.stanford.edu/people/dlm/papers/ontologies-come-of-age-mit-press-(with-citation).htm) are a controlled vocabulary, unambiguous interpretation of classes and term relationships and strict hierarchical subclass relationships. These all apply to a well-constructed thesaurus.

In addition, typical, but not mandatory requirements of an ontology are: property specification of a per-class basis; individual inclusion in the ontology; and value restriction specification on a per-class basis.

DAML+OIL

DAML (DARPA Markup Language) is a more expressive extension of RDFS. The latest release of the language is DAML+OIL (www.daml.org/about.html). OIL (ontology inference layer) includes precise semantics for describing term meanings (and thus also for describing implied information). OIL is compatible with RDFS and allows its extension to provide fully featured ontologies (Bechhofer and others, 2000, 'An informal description of Standard OIL and Instance OIL', www.ontoknowledge.org/oil/downl/oil-whitepaper.pdf).

An ontology written in OIL contains descriptions of classe, slots, and individuals. Classes are collections of objects (for example, person), and may be subclasses of another class (for example, person is a subclass of mammal). Anything that is an instance of a subclass must also be an instance of the class (for example, an instance of a person must be an instance of a mammal, that is, if 'GlendaBrowne' is a person, then 'GlendaBrowne' must also be a mammal). Person is defined in the following format:

class-def person

subclass-of mammal

Classes typically contain information about how their members relate to other (slots), eg:

slot-def has-daughter

These are binary relations, meaning that for every 'has-daughter' there must also be a 'has-parent' relation.

Slots can have constraints that limit the type of slot filler they can have (that is, limit the values that can be entered for that slot). So that only females can be entered as daughters, the following slot-constraint is defined:

slot-constraint has-daughter

value-type female

This is useful for automated data checking.

Instance OIL is a strict superset of Standard OIL, and allows the creation of 'instance-of' and 'related' statements. An 'instance-of' statement asserts that an individual is an instance of a class, eg:

instance-of Zoe zebra

states that Zoe is a zebra.

A 'related' statement asserts that an individual is related to another individual or data value via a slot relation, eg:

related age Zoe 35

states that Zoe is age 35.

OWL (Web Ontology Language)

OWL (Web Ontology Language, www.w3.org/TR/webont-req) from the W3C is derived from OIL, but uses element names that are easier to understand. OWL statements define classes, properties and individuals, together with relationships such as 'subclassOf', 'domain' and 'inverseOf' with a set of 'axioms' for defining restrictions such as 'oneOf', 'disjointWith' and 'intersectionOf'.

OWL also has a built-in general class named 'Thing' that is both the class of all individuals and the superclass of all classes, and a special class with the name 'Nothing' that is the empty class'. OWL Lite provides a subset of OWL based on commonly used features of DAML+OIL.

OWL ontologies are web documents that can be referenced by means of a URI. The official exchange syntax for OWL is RDF ('Diffuse – Guide to the semantic web', www.diffuse.org/semantic-web.html).

Topic maps

RDF: a formal data model from the W3C using XML for the description of web resources using machine readable metadata. It has potential for use in the semantic web.

Topic map: tool for representation of model-based data on the web for enhanced access. Topic maps are based on topics, associations and occurrences. In comparison with RDF, topic maps are developed separately from the documents they refer to.

Topic maps and RDF are both used for the representation of model-based data on the web, and have similar goals and methodologies. Topic maps differ from RDF in that they are topic-centric (the topic map exists independently of the resources it describes), while RDF is resource-centric, with resources annotated directly. For comparisons, see Pepper, 2002, '10 theses on topic maps and RDF' (www.ontopia.net/topicmaps/materials/rdf.html) and ceres.ca.gov/thesaurus/solutions.html.

Topic maps therefore resemble a subject catalogue in a library, while RDF resembles one catalogue entry for one document. It would be useful to be able to integrate topic maps and RDF, to avoid the development of collections of incompatible resources on the web. Research is under way on how to accomplish this goal (www.semanticweb.org/SWWS/ program/full/paper53.pdf).

Topic maps are based on principles used in traditional indexes and thesauri, with inspiration from semantic networks. They involve the application of human intellect to create structured views of information. Topic maps are based on topics, associations (the relationships between topics) and occurrences (resources that discuss the topics).

Topic maps 'float above' the resources they provide access to. They are therefore reusable over any number of resources, and provide valuable information about

relationships between topics. To provide access to information on topics, the topic map is linked (for example, by URL) to the resource it is describing. Linking is done manually in many cases, but can be automated for structured information.

The first step in the creation of a topic map is to define its theme. This can be broad or narrow, for example, 'gardening on the East coast of Australia' or 'digital libraries'. The second step is to gather topics relevant to the theme (thesauri and indexes might be useful sources) along with relevant information resources such as websites. The third step is to structure the associations (relationships) between the collected topics.

Topics are defined broadly to include anything that can be discussed, whether real or not. Topics are multiheaded links, pointing to all occurrences of a subject. Topics are therefore similar to terms in a thesaurus, which are applied to all documents on that particular subject.

Topics can be grouped into classes called topic types, for example, 'person', 'city', 'product', 'part', and so on. These are defined by the designers of each topic map, and make it possible to build specialised indexes. This is one feature of topic maps that makes them expressive, and provides for useful grouping of topics (for example, you could include all cities in France, all people who work in a certain department of a company, or those whose birthdays fall in October).

Topics have an internal identification so they can be referenced, as well as an external representation, or base name, which is used to present the topic to the end user. These are often the same, but don't need to be. In the XML example below, 'bond-uni' is the internal identification, and 'Bond University' is the name that will be seen by users.[51]

```
<topic id="bond-uni"
<baseName>
      <baseNameString>Bond University</baseNameString>
</baseName>
:
</topic>
```

Topics also have variants, which are alternative forms of the base name optimised for computational purposes. These include display names, which may be graphics, and sort names. Sort names are used when the base name is not appropriate as a sort key, for example, where it contains roman numerals or words in languages with a sort order different to that used by computers.[52]

Multiple names can be used when different access points are wanted, for example, 'art museum' and 'museum of art'. (In printed indexes and thesauri, *see* or *use* cross-references would be used between these terms). Multiple names are also used in multilingual topic maps where speakers of different languages will use the same topic.

Just as book indexes show the page number at which information on a subject can be found, so topic maps link to occurrences of information on a subject. A topic 'indexing' might link to the URL of the Australian Society of Indexers (www.aussi.org), as this is an occurrence of the topic 'indexing'.

Below is the XML for a reference to a resource about Bond University, the topic described in the box above:

```
<occurrence>
        <resourceRef xlink:href="http://www.
        bond.edu.au/">
</occurrence>
```

Occurrences can be split into groups depending on the role played by the resource being linked to. Occurrence roles include mentions, definitions, graphics, introductory material, expert opinion, and so on. A link to the Australian Society of Indexers website could have an occurrence role of 'web home page' or 'society webpage' and so on, depending on the roles created in that instance.

The separation of the topic map itself from the occurrences it refers to is one of the key features of topic maps. This allows the development of topic maps independently of the resources they describe, enabling them to be used with a range of information sources. This is particularly useful for topics such as places that are used by very many organisations; it will also apply to some extent to more general subject areas, although these may have to be adapted for individual use. For example, it is unlikely that a topic map created by one insurance company would be immediately usable by another, as the companies would have different products and processes, and would use different words to describe many of the things they do.

Another key feature of topic maps is the use of associations between topics, showing the specific relationship that is applicable to each individual case. Each topic in a topic map may have many different associations. For example, 'Tim Winton *works as a* writer', 'Tim Winton *lives in* Fremantle', 'Fremantle *is in* Western Australia', and 'WA *is an abbreviation of* Western Australia'. Many of these relationships would be important in a traditional thesaurus, but whereas in a thesaurus WA/Western Australia might just have a general synonymous relationship (USE/USE FOR), and Western Australia/Fremantle might have a hierarchical relationship (Broader term/Narrower term), in a topic map the exact nature of the relationship is spelt out.

Topics can be also classified into a hierarchical structure, in which children inherit properties of parents. For example, 'Bond University' can be classified as an 'instanceOf' the more general topic 'university'.

Topic names, topic occurrences, and topic associations comprise the topic characteristics. These are assigned to a topic, and the assignment can be qualified by its scope to increase its relevance. Scopes are made of a set of components called themes, and can be used to distinguish between various names for the same topic (for example, nicknames and formal names), or to characterise the domain

of knowledge in which an assertion is valid. This allows a search for 'springs' in the travel domain to retrieve information on hot springs and artesian bores, while a search for 'springs' in the mechanical engineering domain will retrieve information about metal coils.

Facets are used to filter portions of information, for example, to extract information in a given language, or relevant to a targeted audience. They are similar to the check tags used in Medline indexing, in which it is very simple to search for documents relevant to humans (no rat or rabbit studies), pregnancy (for example, effect of aspirin in), and so on.

Topic maps can be useful for structuring information repositories and navigating through them. Application areas include web portals, intranets, and content management systems. The topic map data model allows automated merging of information from diverse sources including databases, thesauri, automatic tagging tools, and metadata from RDF Dublin Core documents. This feature allows the integration of information from many sources within an organisation into a coherent whole.

To see a simple topic map in action go to the Infoloom site (Biezunski, www.infoloom.com/tmsample/bie0.htm) and select the 'Indexes' link. You will see a list of different indexes (topic types) that you can search within.

The **Diffuse** site also uses a topic map. If you search the 'Alphabetical index of standards and specifications' (www.diffuse.org/alpha.html#1) some entries are tagged 'Display Topic Map' – clicking on this button takes you to a topic map for that standard. The topic map for 'HTTP: HyperText Transfer Protocol' (www.diffuse.org/TopicMaps/HTTP.xml) [53], for example, includes the acronym and full name, standards that are based on HTTP, standards that are partly based on HTTP, standards referenced by HTTP and news items about HTTP. This standard layout is used in all the topic maps.

A topic map for the composer Puccini is shown in Figure 24, below:

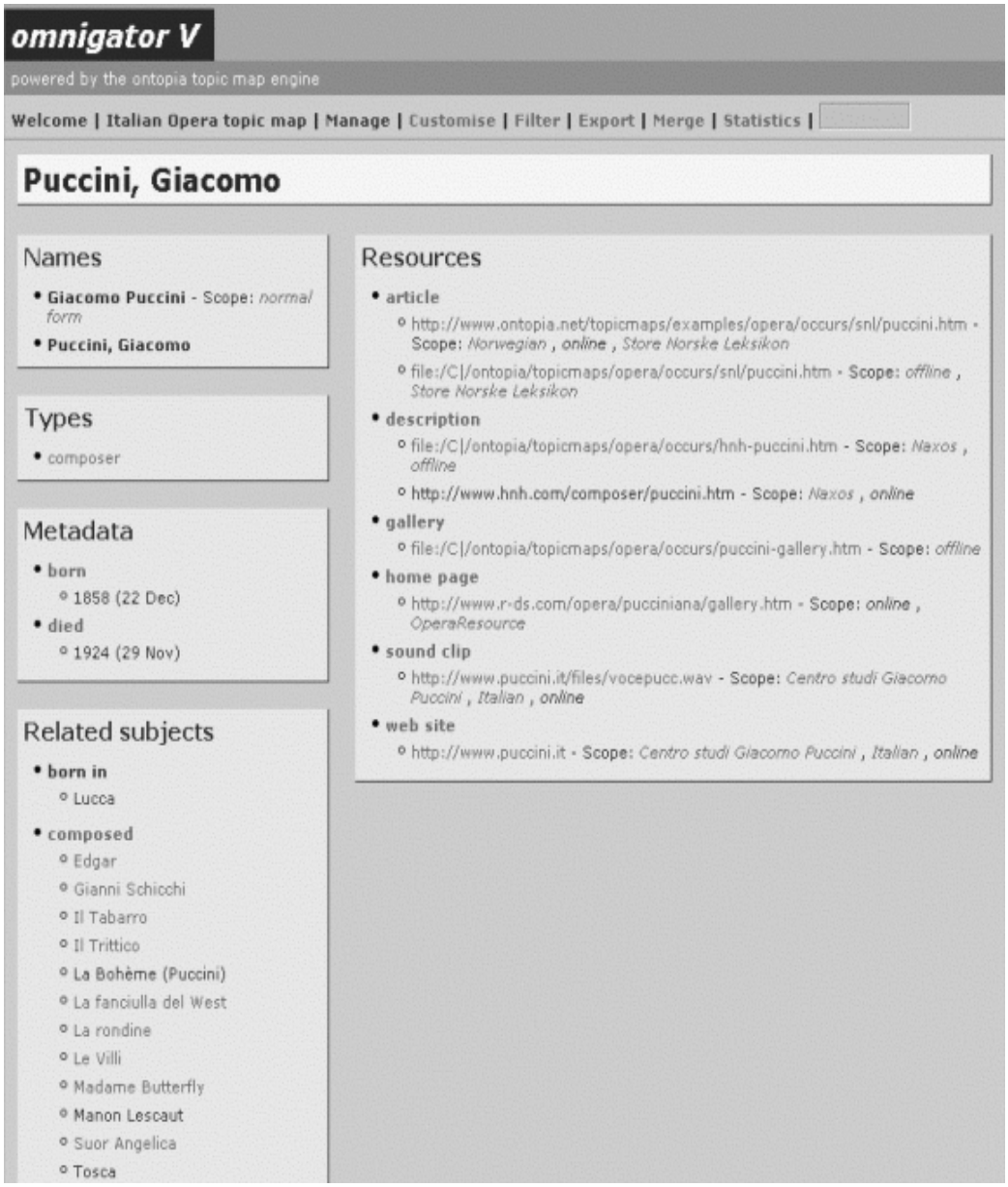

Figure 24. Topic map for the topic 'Puccini'

For more information download the starter pack from the Ontopia website at www.ontopia.net/topicmaps/what.html. This includes the 'TAO of topic maps' article by Steve Pepper (www.ontopia.net/topicmaps/materials/tao.pdf). The articles by Michel Biezunski ('Topic maps at a glance', www.infoloom.com/tmsample/bie0.htm) and Rafal Ksiezyk ('Trying not to get lost with a topic map', www.infoloom.com/tmsample/ksi2.htm) also provide excellent introductions.

10. SEARCH INTERMEDIATION

Social navigation

Mediated information access

To find information on the web we rely largely on our own search skills, while the semantic web promises automated support in the future. Nonetheless, humans help remains important for information discovery in at least two ways – the first is through the automatic derivation of suggestions about information use based on decisions made by 'similar' people. The second is the reliance on paid or unpaid help from expert searchers.

Social navigation

Social navigation is an access method based on the principle that we all benefit from experience – especially someone else's! It is a broad term that covers a range of mechanisms whereby the content that will satisfy a person's information need is judged by reference to actions or comments from a community of like-minded people. Examples of social navigation include Epinions' **recommendation engine**, where users provide ratings and reviews (www.epinions.com) and Amazon's recommendations based on **collaborative filtering** in which they provide book purchasers with lists of other books bought by the people who bought the book they are interested in (Linden and others, *IEEE Internet Computing*, January/February 2003, 'Amazon.com recommendations: item-to-item collaborative filtering', dsonline.computer.org/0301/d/w1lind.htm).

See also 'History-rich tools for social navigation' (Wexelblat, web.media.mit.edu/~wex/Footprints2/fp-v2.html).

Mediated information access

When all else fails, and a user just can't find what they want, librarians and commercial services provide a safety net. There are now a number of free services in which people can email a query to a librarian, and there is a paid service offered by Google. [54]

Librarian-assisted information access

AskNow! is a free Australian question-answering service available via the web at www.asknow.gov.au. It enlists State and Federal government librarians from around Australia to provide the online equivalent of a reference desk.

Google Answers

Google Answers (www.answers.google.com/answers) is a paid system: people submit queries along with a nominated fee (from $US2 upwards). Google researchers who elect to try and answer the question receive 75% of the fee if they

succeed; the rest goes to Google to cover setup and administration costs. Google researchers are freelancers who are vetted by the company and put through a battery of test queries. Google then keeps in touch with them through weekly newsletters. Researchers also have their own chat system.

In general, Google researchers are motivated and competent. Short of raising their fee, however, questioners have no way to influence when their questions will be taken up or how long will be spent on them. Some questions are answered within hours; others remain on the system unanswered for months. There is a presumption that the answer to a Google question will not be found on the Internet through normal search procedures, and the most effective searchers often have access to research databases or other collections of material not easily available to the general public.

11. SUBMISSION TO OR FINDING BY EXTERNAL SEARCH ENGINES OR DIRECTORIES

Once someone has created a site, they want users to be able to find it. This will happen through announcements and word of mouth, and by them finding it through web-wide search engines and directories. The directories might be general (for example, www.yahoo.com), or specific to a certain subject or user group (for example, legal information at www.austlii.edu.au). The next step, therefore, is to maximise the chances of the site being found by the people to whom it will be relevant.

This can be done by submitting a site to search engines and directories. Before doing so it should be optimised so that it is likely to be ranked highly by the search engine. Ranking criteria vary between engines, but in general the things that matter are keywords in titles, headings and page text and links to the site, especially from sites that are themselves ranked highly. These are discussed in the section on 'Search engine optimisation'.

As well as aiming for a high ranking by optimising a site, people can also pay for inclusion in a directory, for quicker listing and more frequent spidering, and for special listing at the top of the list of returned hits. These options are discussed in the section below on 'Paid search services'.

Subject-specific directories are a good way to target searchers with a specific interest in your topic. They are discussed at the end of this section.

Once a site can be found after appropriate searches, regular checks have to be done to ensure that the pages remain listed.

Search engine optimisation (SEO)

Bot (Agent, Robot): programs with some artificial intelligence that are sent to do a task in lieu of a real person. They run automatically and act autonomously. Spiders are one example.

Spider (crawler, web crawler): bot that visits publicly accessible websites following all links it comes across collecting data for search engine 'indexes'. A spider discovers new sites and updates information from sites previously visited. A spider can also be used to check links within a website.

Google and other search engines often retrieve thousands or millions of hits in response to a search. Most users don't look beyond the first 10 to 30, with a few

persistent users searching hundreds. For a site to be included in the first listing it must be optimised for retrieval. There are simple steps to do this; there are also many Search Engine Optimisation (SEO) companies which offer this service.

Firstly, the target market has to be determined, along with the keywords being used for searching. Sites of colleagues and competitors should be examined to see which ones rank highly. Secondly, the criteria used by major search engines for ranking sites should be investigated and fulfilled if possible.

Keywords – metadata for web-wide search

Most search engines no longer pay any or much attention to keyword metadata for web-wide search (Goodman, 2002, 'An end to metatags' www.traffick.com/article.asp?aID=102), although Inktomi still does to some extent (Lloyd, www.searchengineguide.com/articles/2003/0718_rc1.html).

On the other hand, they do rank sites more highly if they have search terms in their titles, headings and early paragraphs in their sites (as well as in the rest of the text). In image-heavy sites it is particularly important to add descriptions of the site and its major sections to the home page.

Without keyword metadata it is difficult to cater for all possible term variations. One suggestion is to use as many **synonyms** as possible in the title and heading. Another is to use word variations in different pages of your site, so one page might say 'optimisation' throughout, while another says 'optimization'. Unfortunately putting both on the same page looks sloppy. (If you have an accommodation site make sure you put the spelling 'accomodation' on it somewhere, as it is used almost as frequently as the correct spelling.)

Now that Google offers a synonym search option this issue may become less significant. (In Google putting a tilde sign (~) before a term expands the search to include synonyms of that term).

While metadata might be inappropriate for use in general search engines, due to problems with spamming, it may be more useful in the closed web – 'communities of trust where the structure and meaning of webpages can be anticipated.' (Brooks, informationr.net/ir/8-3/paper154.html). See also the section on the 'Semantic web' above.

URLs

It is important to choose a URL that contains your organisation's name, or a description of an important role you do and to provide clear navigation paths within the site, using target keywords as link names within the site.

Dynamically generated URLs from content management systems containing characters such as "?" and "&" (for example, webindex.com/paper?sku=123 &uid=456) are recognised by Google, but not by other search engines. Where possible these should be converted to a syntax that search engines can crawl.

Amazon has done this, for example, changing:

> amazon.com/store?shop=cd&sku=B00004FIZ&ref=p_ir_m&sessionID=107-6571839-6268523
>
> to
>
> amazon.com/exec/obidos/ASIN/ref=B00004FIZ/ref=pd_ir_m/107-6571839-6268523

This works because Amazon's application server knows the fields in the URL are CGI parameters in a certain order (hotwired.lycos.com/webmonkey/01/23/index1a_page3.html).

Frames and Flash

Sites should be kept as simple as possible. The use of frames is no longer the major problem it used to be, as both Google and Inktomi now crawl them. [55] Flash content is not likely to be crawled in the near future as the files rarely contain text.

Links to the page

The best way to get listed with most of the major crawlers is to build links to your webpage. Search engines such as Google rank pages highly if they have many links leading to them, especially if the pages providing the links are themselves highly linked (www.google.com/technology/index.html). This has led to the development of 'link farms', where people swap links to try and increase their ranking. In response to this, pages are now also ranked more highly if the links come from sites in the same focus area (Nobles, 21 October 2003, www.searchengineguide.com/aws/2003/1021_aws1.html). The best way to increase links is to improve the site quality so that people *want* to link to it.

Valuable well-written content

The best way to make people link to your site is by adding valuable content and reducing information pollution that might put users off. Jakob (www.useit.com/alertbox/20030811.html) reports that 'information pollution' – that is, excessive words and meaningless details – is making it harder for people to find useful information. 'The more you say, the more people tune out your message'.

Site submission

Crawler-based search engines automatically visit webpages to compile their listings, so a page can be included without being submitted, although submission will make inclusion faster and more certain. The fastest guarantee of inclusion in a search engine database is through paid submission (see below). Submitting to directories is also useful (see below).

The top seven search engines identified by Useit.com are Google, Yahoo, MSN, AskJeeves, Lycos, AltaVista and AOL, which together accounted for 97% of

traffic to their site (www.useit.com/about/searchreferrals.html). Results will be different for different types of searches.

There is usually information or a link on a search engine home page about how to submit a site. Once a site has been submitted it can take months for it to be included. A crawler will find all pages that are linked to from the page submitted, so the home page and one or two other key pages can be enough. Some people submit a crawler page containing links to every page that should be crawled. This can be labelled 'site map' on the home page.

Google, AllTheWeb and AltaVista provide 'Add URL' pages that let you submit a URL directly to the crawler, free. AllTheWeb and AltaVista also offer a paid inclusion program, as do Inktomi and Teoma (Sullivan, 29 April 2003, www.searchenginewatch.com/webmasters/article.php/2167871, see also below). Google offers sponsorship not paid inclusion (see below).

Ross Dunn offers useful suggestions in his '10 minute search engine optimization' (21 April 2003, www.searchengineguide.com/dunn/2003/0421_rd1.html), while SearchEngineWatch (2003, searchenginewatch.com/webmasters) offers detailed submission tips for specific situations (submitting to directories, crawlers, and paid listings).

Paid search services

Paid inclusion/Paid submission/Paid registration

Editorial results: search engine hits dependent on content and not influenced by payment.

Paid inclusion: payment for inclusion of a site in a search engine's editorial listings, without an artificial boost in ranking.

Paid submission: payment for guaranteed consideration of a site for inclusion in a directory.

Pay-per-click listing (PPC): search engine advertising in which payment is based on the number of times the website is selected (clicked) from the results list. Is used in **paid placement advertising** and **paid inclusion**.

Paid submission is offered by Yahoo, and guarantees that a site will be *considered* for inclusion in the Yahoo directory (although there are no guarantees that it will actually *be* included).

Paid inclusion guarantees inclusion in a search engine's editorial listings, without an artificial boost in ranking. It is useful for getting quickly recognised by search engines, and for guaranteeing that all significant pages from a website are included.

Search engines such as Inktomi, AllTheWeb, Teoma and AltaVista offer a paid inclusion program that guarantees to include a site within a time period of between two days and two weeks (usually a week) and to regularly revisit the page for a year. AllTheWeb offers its program indirectly through AllTheWeb

resellers, and Teoma's is done through the Ask Jeeves Site Submit program (Sullivan, 29 April 2003, www.searchenginewatch.com/webmasters/article.php/2167871, ask.ineedhits.com/sitesubmit.asp?id=30129&o=0, or select 'Submit a site' at www.askjeeves.com). Ask Jeeves is the only metasearch engine that takes submissions – the others all rely on submissions to the individual search engines included in their metasearches. Google doesn't offer paid inclusion, fearing it would dilute the quality of its ranking, but does offer paid placement (see below).

XML paid inclusion (Lee, June 6 2003, www.clickz.com/search/strat/article.php/2217951) is used by search engines such as Inktomi as an efficient way to spider paid inclusion files. Sites provide Inktomi with an information-rich XML file containing information about the site's pages including title, description, keywords and text. This information resides on the host site or the site of a value-added XML agency (VAXA) or the servers of an XML reseller. Inktomi refreshes every 48 hours, thus ensuring that the search engine database is kept up-to-date. Because Inktomi is refreshed so frequently it can be a good place to check search engine optimisation approaches (Lloyd, 'Optimising for Inktomi – and how it can help on other SEs', www.searchengineguide.com/articles/2003/0718_rc1.html).

Many people argue that paid inclusion has the potential to distort **search rankings**, while others state that paid inclusion is a way of increasing the relevancy of search results (Boswell, 16 October 2003, 'Paid inclusion is online search at its purest', www.marketleap.com/report/ml_report_47.htm). Others (including Sullivan, 5 November 2003, Search Engine Report n.84, searchenginewatch.com/sereport/article.php/3104511#inclusion) argue that paid inclusion is acceptable if it is labelled as such, but advertisers are reluctant to accept this as they feel that people would then ignore their sites (although if they had truly improved relevancy for users this should not be a concern).

Although paid inclusion is only meant to speed inclusion, some people claim that it can also offer improved ranking. An investigation by *Business Week* (www.businessweek.com/magazine/content/03_40/b3852098_mz063.htm) found that 10 out of 20 'advertisers and online marketing pros' had experienced firsthand a rise in search engine rankings when they signed up for paid inclusion[56].

Advertising and sponsorship (Paid placement/Paid listings/Pay-per-click)

Paid placement (Advertising): listings in search engine results where advertisers pay for a guaranteed high ranking, usually dependent on specified keywords being used in a search. These listings are usually segregated from editorial results and labelled to indicate that they are ads. Also known as 'pay for placement', 'pay for performance', 'CPC (cost-per-click) listings, or PPC (pay-per-click) listings.

Sponsorship: sponsored ads are normally located in a separate boxed and labelled section at the top of a search engine's results list.

'Buying your way in: search engine advertising chart' by Danny Sullivan (30 May 2003, www.searchenginewatch.com/webmasters/print.php/34751_2167941) gives a clear overview of the types of paid listings with a chart indicating how the search engines display the paid content. Those that do not conform to US Federal Trade Commission guidelines as ranked by SearchEngineWatch are labelled 'FAIL'. The guide also mentions content promotion and banner ads, which are not discussed here. Google also provides sponsorship options. For more information see the vast selection of articles on search engine advertising at www.searchenginewatch.com/resources/article.php/2156561.

Advertisers can pay for placement on Google through **Google AdWords**, which displays an ad at the right-hand side of the page, and the more expensive sponsorship option, which places the ad at the top of the editorial listings. **Google Premium Sponsorships** (www.google.com/ads/overview.html) require advertisers to spend at least US$10000 over a three-month period.

Google AdWords (https://adwords.google.com/select) is a self-service system in which users select the keywords which they want to trigger their ad, and set a cost per click. Payment is based on the number of times that the link is followed. Recent changes have meant that Google now uses 'expanded matching capability', generating ads by synonyms and other phrases related to the selected keywords. Advertisers can test keywords to see which more specific keywords will also trigger their ads (if they are using the expanded matching capability) and to get ideas of alternative keywords to use. Keywords can also be defined as negative matches, ensuring that certain combinations of terms do not trigger the ad (www.clickz.com/search/strat/article.php/3092841).

Overture (formerly **GoTo**, www.overture.com/d/about/advertisers and www.content.overture.com/d/home) provides placement for a pay-per-click fee. Sites 'bid' on the amount they will pay for each click on their URL – the higher the payment per click, the higher their ranking if more than one site has chosen the same keyword. These sites are labelled 'sponsored listing' in the results list, and are followed by sites provided by Inktomi which are labelled 'additional listing'. The additional listings are not ranked according to payment. Overture listings appear in a number of other search engines including AOL Search. Danny Sullivan ('Submitting via paid listings: Overture (GoTo), 2002, FindWhat & Google', www.searchenginewatch.com/webmasters/article.php/2167821) says of Overture: 'No other route can put you in the top results of many major search engines in such a short period of time.'

FindWhat (www.findwhat.com) is a pay-per-click engine similar to Overture, but with less traffic. Its listings appear on Excite's search results page.

Sponsorship is often general, but pretends to be targeted. For example, if you search **Alta Vista** (www.altavista.com) for 'Glenda Browne', the second sponsored link is for eBay and says 'Find Browne items at low prices', while in a search for 'thesaurus construction' the third sponsored match reads: 'Surprise? eBay has construction supplies'.

Yahoo Sponsor Listings (https://ecom.yahoo.com/fast/sponsor?spLMC) are available to commercial websites that are already listed in the Yahoo Directory.

Participating sites are featured in a 'Sponsor Listings' module within the appropriate category, and are also highlighted in the regular alphabetic list of sites.

Search engine submission services

Automated search engine submission services submit a site to as many search engines as possible. One example is **Submit It** (www.submit-it.com), which is part of MSN bCentral. It charges US$79 to keep one page submitted for a year (US$30 for extra pages). The service includes:

- Keyword research
- Search engine optimisation advice
- Guaranteed submission/registration (Priority Listing through Inktomi)
- Ability to monitor search engine rankings
- Checks on link popularity, and alerts to broken links.

Submitting sites to directories and subject gateways

> **Directory:** collection of evaluated links to websites, usually categorised by subject. Many search engines, for example, Yahoo and Google, have associated directories. When directories are limited to information on a specific subject or discipline they are often called subject gateways.
>
> **Subject gateway:** a directory limited to a specific subject area such as *education*, or *Tasmania*. Sometimes called a 'portal'.

The web has been described as the greatest source of unsubstantiated opinion ever seen. One approach to effective searching is the use of general directories and subject-specific gateways that lead to *selected* resources on the web. While not guaranteed, these sources have at least been vetted and considered useful by some overseer.

General directories may be nonprofit (for example, BUBL and LII) or commercial (for example, Yahoo and Google). Subject gateways normally focus on a place (for example, Tasmania), format (for example, ebooks) or discipline (for example, education). The best ones offer a range of access methods including categorisation and controlled vocabulary searching using terms from a thesaurus as well as search.

Most directories offer people the chance to submit sites for inclusion, either free or for a fee, through a button labelled 'Suggest a URL' or similar. Subject gateways only accept sites that fit their selection criteria.

The DESIRE project [57] has produced a long list of selection criteria from which individual gateways can select those most relevant to their own situations. The criteria are in five groups: scope, content, form, process, and collection

management. Content criteria, such as comprehensiveness of the information and authority of the author are usually the most important.

General directories

Some general sites which are worth considering are:

- BUBL LINK (www.bubl.ac.uk/submiturl.html) – submit a site suggestion via the URL suggestion form, or send an email to mailto:bubl@bubl.ac.uk including the URL and a short description of the purpose and contents of the site you are suggesting
- Google – the Google directory uses sites from the Netscape Open Directory Project (dmoz.org/add.html)
- Infomine (infomine.ucr.edu/feedback/suggest.php) – in October 2003 they had stopped taking suggestions as they had been plagued by bulk inappropriate commercial submissions. It appears they have worked out a way to filter junk as the form for site submission is now active again
- LII (lii.org/search?suggest=1) – has a form which you can fill in to make your submission
- LookSmart offers listings on a pay-per-click basis (listings.looksmart.com/home/details.jhtml). Non-commercial sites can be submitted via Zeal (www.zeal.com/users/non_profit.jhtml?rpc=49) at no charge.

Yahoo!

The Yahoo directory is an important place to be listed, both in its own right and because it is an important source of links and ranking information for Google.

At the Yahoo home page (www.yahoo.com) clicking 'Suggest your site' at the bottom of the first screen takes you to the 'Express submission with payment' page. This provides a quick assessment of your site (for a fee) with a yearly inclusion fee if accepted. From this page you can also select sponsorship (rank higher in the directory, see above) and inclusion in Inktomi search (for those who search rather than browse). If you scroll to the bottom of the home page and click on 'How to suggest a site' you get general information on suggesting any site for inclusion.

If you select Yahoo Express (https://ecom.yahoo.com/dir/express/intro) your site will be reviewed within 7 business days, leading to quicker inclusion in the directory (although there is still no guarantee that your site will be included).

To 'Suggest Your Site' (docs.yahoo.com/info/suggest) you go to the appropriate Category and click on 'Suggest a Site'. This submission is free except for business categories, but there is no guarantee that your site will be investigated or included.

Subject-specific or user-specific directories

Subject-specific or user-specific sites include:

- Arbor Nutrition Guide (www.arborcom.com; send an email by selecting 'Comments, new sites')
- AustLII (www.austlii.edu.au; select 'Feedback' at the bottom of the home page, then 'New report' then select 'Content request' on the feedback form)
- AVEL (Australasian engineering & IT resources, avel.edu.au/docs.html)
- Sosig (Social Science Information Gateway, www.sosig.ac.uk/new_resource.html)
- Tasmania Online (www.tas.gov.au/tasonline/submiturl.asp)
- Yahooligans! (go to www.yahooligans.com and select 'Suggest a site').

Webrings

Webrings, which link sites about the same topic, act as self-selected subject gateways. You can browse a directory of rings at dir.webring.com/rw. The Indexers' Webring (www.geocities.com/Athens/4537/indxr.html) is an example of a professional ring.

12. BRINGING IT ALL TOGETHER

This book has covered a range of methods for accessing information on the web, with examples of each type. Each method has its strengths and weaknesses, and a web manager has to consider these when choosing which methods to implement.

Categorisation and classification/Navigational structure/Taxonomies/Site maps

Within individual websites, categorisation is important in the basic navigational structure of the site and enables users to browse related material. Categories need to be determined with users in mind, and are depicted in a taxonomy.

Site maps (visual and tables of contents) give a useful overview of a site and allow users to move directly to a webpage of interest without having to click through many levels of the hierarchy.

Search engines

Search engines can be useful on **individual websites** as they:

- allow access to all information on the website (some sections might not be manually indexed)
- allow detailed access (manual indexes don't include every mention of a word)
- give immediate access to newly added material (when manual indexing might lag)
- help people find topics with well-defined names
- give access for people who prefer computer searching.

Search engines are essential for searching the **whole web** as they are automated and can cope with the huge amounts of information available. They offer some sort of relevance ranking, and allow users to redefine their searches if they retrieve too much or too little information.

Overall, search engines are a crucial tool for information retrieval. Their use is so entrenched that to a lot of people 'indexing the web' means indexing of sites by a search engine.

Metadata, thesauri, ontologies and topic maps

The addition of metadata to a webpage can enhance its retrieval by search engines, especially on individual websites. Use of a thesaurus enables more consistent metadata creation, and provides a useful linguistic tool for searchers. Other controlled vocabularies of value on the web include ontologies, which are predicted to be crucial in the development of computer intelligence in the semantic web, faceted classifications and topic maps.

Back-of-book-style indexing

Back-of-book-style indexes are the most granular information retrieval tool on the web (along with search engines); that is, they give access to specific chunks of information rather than whole webpages or websites. For this reason they are best suited to individual websites or to documents within websites.

The advantages of back-of-book-style indexes are that they provide:

- A familiar format (primary school children learn index use)
- Access to *specific* information (index terms are as specific as the information available)
- Access to *selected* information (someone has deemed these items worth indexing)
- Immediate access to information (one click takes you there)
- Browsable lists (see all the options and choose the best)
- Multiple entry points to information
- Subheadings (to help you choose between options)
- Cross-references (to lead you to alternative or extra information).

The main disadvantages of back-of-book-style indexing are the costs, the time delay between creation of the resource and its addition to the index, the need for regular updating if the content changes, and the need for skilled indexers.

The best sites offer a range of information retrieval tools to cater for people with different searching preferences and for people with different information needs. It is important to maintain them well, and to make the options you have chosen to provide clear to users.

Access to sites on the web

In addition to providing access to information within a website, it is important to make sure the site is accessible to people searching the web. To do this it is important to optimise the site for retrieval – this includes monitoring the ranking of the site, and keeping up-to-date with changes in search engine policy. There are then options to pay for inclusion or sponsorship if desired. In addition, submission to subject gateways and communication with groups that are likely to be interested gives a targeted means of promotion.

The future of information access on the web

In the future it is likely that human intervention to aid access to information will continue, often in more sophisticated and formal ways. A range of access methods will continue to be important, for different situations, and for different types of users, and human intermediation will always be needed as a last – if not first – resort.

Scenario

Rhiannon is interested in web indexing. She has chosen this topic for a university assignment because she thinks she might like to work as an indexer.

She has no idea what organisations are involved in indexing, so she starts with a keyword search over the whole web. Going to Google, she types in 'indexes', later refining her search with other terms. One of the first links from the search on 'web indexing' is to information about the AusSI Web Indexing Prize. She navigates through the Australian Society of Indexers' website to sections of interest, and finds lots of other information about indexing in Australia. She looks up web indexing in the A to Z site index and finds information about a specialised mailing list as well.

In the index there is a *See also* reference to metadata, and she finds this related information interesting too. She follows the link to the MetaMatters site and adds it to her list of favourites.

She retries Google using the expanded term 'website indexing' and finds a different set of links. The first one links to www.optusnet.com.au/~webindexing, where she finds details of this book. A 'sponsored link' at the right leads to the same site, under another link which says 'Add a search engine to your website today'. Both of the sponsored links are ranked equally for interest, although they reflect totally different approaches. This is a problem with the broad range of meanings associated with 'website indexing'.

A search for 'website indexing' at Britannica.com (to which she subscribes) shows articles on machine indexing, economic indices (for example, Laspeyres price index; index funds) and some specific indexes. She finds that the different meanings of *index* often cause problems in her searches.

She goes to the BUBL Link site, which she has used as a student, and drills down through the Dewey Decimal Classification hierarchy finding links to the Society of Indexers and the American Society of Indexers at notation 025.3.

Her search for background information leads her to the Diffuse site, where a topic map provides structured content on RDF.

In this short search session Rhiannon used a global search engine to get started, and through it she found a useful general site. While on this site she browsed through the site hierarchy, and then homed in on her special area of interest using the back-of-book-style site index. She found useful information on one site through the classification structure. Some of her searches might have been more successful because of metadata, but this was not apparent to her, as metadata is a hidden tool.

Overall, she was well-served by a number of different tools with different aims and creators, different areas to cover, and different levels of specificity.

13. CONCLUSION

In just over a decade, the web has become the medium of choice for storing and distributing vast quantities of information. Web-wide search engines use sophisticated search techniques to locate sites on the web. But more and more sites are competing for the searcher's attention. And locating a site is like locating a book in a library—once found, the user must still go through the book to find the topic of interest. As individual sites become larger and more complex, they will need increasingly sophisticated retrieval methods to direct users to the material that they are looking for.

The alphabetical back-of-book-style index is a familiar, tried-and-true, user-friendly way of providing direct access to topics covered in a book or on a website. Good indexes are effective tools for improving the speed and accuracy of user searches. As sites grow and develop, and user frustration increases proportionally, we can expect to see more demand for indexing as a retrieval method.

If you think a website needs an index, let the web manager know. If you manage a website use this book to create an index for your site. The more visible website indexing becomes, the more users will realise the benefits!

APPENDIX 1: FURTHER INFORMATION

If you decide to index a website you may want more information and support. Some sources are listed below.

Australian Society of Indexers' website

The Australian Society of Indexers website (www.aussi.org) has information on many topics, including:

- courses and events
- mailing lists
- overseas indexing societies
- the AusSI Web Indexing Award
- monthly newsletter
- professional indexers available for freelance work (click on the link to *Indexers Available* on the home page).

Mailing lists

Mailing lists are sources of information and support for web indexers. These include:

- **aliaINDEXERS** for Australian indexing: alia.org.au/alianet/e-lists/subscribe.html
- **CHI-Web** for usability**:** www.sigchi.org/web/index.html
- **DC-ANZ** for Dublin Core**:** groups.yahoo.com/group/DC_ANZ
- **eBook community** for electronic books**:** groups.yahoo.com/group/ebook-community
- **Faceted classification (FCD):** groups.yahoo.com/group/facetedclassification
- **Index-L** for all types of indexing**:** indexpup.com/index-list/faq.html
- **SIGIA-L** for information architecture**:** www.asis.org/AboutASIS/SIGEmailLists/ia.html.

Other indexing-related mailing lists are listed at www.asindexing.org/site/discgrps.shtml.

Courses

The SISTM Continuing Education Department at the University of NSW runs courses in indexing, thesaurus construction, and web and intranet indexing (mailto:M.Henninger@unsw.edu.au , cpd.sistm.unsw.edu.au/bestsellers.html). The Australian Society of Indexers (www.aussi.org) runs occasional courses in book and website indexing.

Jobs in website indexing

There is no clear-cut road to jobs in website indexing. After doing training and gaining experience (in a voluntary capacity if no other way) the best approach is to network with other indexers, with information architects and with knowledge managers. Library and information job agencies sometimes have metadata work on their books. In Australia try the following:

- Information Management Staff by Zenith (www.zenman.com.au)
- One Umbrella (www.oneumbrella.com.au).

APPENDIX 2: BASIC INDEXING PRINCIPLES

Creating an index in any format requires knowledge and skills relating to indexing, and the ability to understand the users of the document or website and their needs. The best indexes will be created by professional indexers or by people with a good knowledge of the site and its users, and a willingness to learn the skill of indexing. This section introduces general indexing principles.[58]

Analysis of text

Scope

The first step in indexing is to analyse the text, and decide what sorts of entries should be included in the index. To do this the indexer considers the potential audience (or audiences) for the index, the amount of material to be indexed, and the space that is available for the index.

The indexer decides what types of information (for example, tables and illustrations) and what sorts of content (for example, names of people and organisations; names of authors) should be indexed, and to what level of detail.

Selection of information to be indexed

The indexer then has to decide which specific items in the text (or other information source) should be indexed. This is done by judging the importance of the information, the needs of the users, and the context.

When deciding whether to add an index term, the indexer should ask: 'If someone looked up this term in the index, would they be pleased to find this paragraph?' If the answer is 'Yes' then the index term should be included.

Selection of terms

Choosing content to be indexed, and choosing the terms under which it is indexed, are two different steps. You might decide that a paragraph on the Goods and Services Tax should be indexed. The next step is to decide how to index it. Obvious options are to have entries for *Goods and Services Tax* and *GST*. If you think people would not be familiar with the abbreviation you should add the full name in parentheses, for example, *GST (Goods and Services Tax).* Another possible entry is *tax* (maybe with a subheading for *GST*); alternatively you should have a reference from *tax* ('tax, *see also* Goods and Services Tax').

Indexers select terms both for people who are new to the book (terms that may not be in the book at all) and for people who have read the book (terms which readers of the book will be familiar with).

Wording of index entries

There are a number of indexing principles that should be used unless there is a good reason to follow another method. Some of these principles are shared with other professions such as librarianship, and include:

- Use the language of the text: thus if the book consistently says 'sodium chloride', this should be the main term in the index, though the indexer may want to include a cross-reference from 'salt'
- Use natural (normal) language unless scientific or other terms are appropriate, as above
- Use direct order (for example, *hospital libraries*). If necessary, also use indirect order, or include a reference from the term in indirect order (for example, *libraries, hospital,* see *hospital libraries*)
- Enter terms under their specific name, for example, if the subject is Siamese cats, use the index entry *Siamese cats*, not *cats*, or *pets*. Entries that turn out to be too specific are a lot easier to fix than those that turn out to not be specific enough!

Cross-references

See references direct users from nonpreferred to preferred terms. For example, 'shiraz, *see* syrah' (these are alternative names for the same grape variety). If space permits indexers can create a double entry instead; that is, include the same page numbers under *shiraz* and *syrah*. *See* references are sometimes written differently in online indexes, for example, 'Search using' or 'Go to'.

See also references are used at preferred terms to suggest additional places to look, for example, 'meningitis, *see also Haemophilus influenzae'* (a reference from the disease to an organism that causes it). *See also* references are usually reciprocal, and you would also expect to find a reference '*Haemophilus influenzae*, *see also* meningitis'. (References that contain one scientific name unfortunately have to have an italic name next to an italic *see* or *see also*. When both index terms in a reference are in italic, the *see* or *see also* is written in normal style: '*And Then There Were None,* see *Ten Little Indians'*).

Names

There are complex rules for indexing names. Consult a library reference such as AACR2[59] or an indexing textbook for more information. Biographical dictionaries and library catalogues (many are online) will show you what form of name other people have used.

Issues to be considered include name variants (for example, *Mao Zedong* and *Mao Tse-tung*), earlier and later names (for example, single and married names), entry part of name (for example, *van Gogh, Vincent* or *Gogh, Vincent van*) and made up names (for example, *Wile E. Coyote*). If in doubt about which form of a name to

use in an index, and if space permits, use both (for example, *Malcolm X* should be entered in direct order under *M* for Malcolm, however, if you think your users might look under *X* you should also enter it as *X, Malcolm*).

If two people with identical names need to be indexed, a qualifier should be added in parentheses to distinguish between them. Dates of birth or death and occupations are commonly used, for example, *Joan (d.1891)* and *Joan (shopkeeper)*. Place names are often qualified with the area they are in, especially if there could be ambiguity, for example, *London (Ontario)* and *London (England)*.

Corporate names (names of organisations) are entered in direct order, for example, *Martha Agnesi Automotive Services*. You may need a reference from, or double entry at, an alternative form of the name. For example, also add an entry for *Agnesi Automotive Services* or *Blaxland Automotive Services*, if these are other names the organisation is known by.

Plural or singular word forms

Most indexers use the plural form for all items that are countable; this allows the use of the singular for other meanings. For example, all of the following pairs have different meanings: *trust* and *trusts*, *interest* (in banking) and *interests* (hobbies and activities), *storm cover* (insurance against storms) and *storm covers* (covers for boats for use during storms). The choice will sometimes differ for different users; for example, construction indexes might use the term *cements* to indicate the different types of cement available.

Capitalisation

The Australian and New Zealand indexing standard[60] (which is based on the international standard), advocates the use of initial lower case letters except for proper names. This means that proper names can then be easily distinguished from other entries by their initial capital.

Lynn Moncrief writes ‘Using initial capital letters in main headings can create havoc with the usability of computer-related texts. Case helps in readily distinguishing between commands, GUI components, and program components’[61] (yet the index to that book uses all initial capitals!)

Filing order (sequencing)

The two main filing rules (that is, rules used for sorting index entries) are word-by-word and letter-by-letter. With word-by-word filing, a space files before anything else; with letter-by-letter filing, spaces are ignored. In word-by-word filing *New York* files before *Newark*, because the space between *New* and *York* files before the ‘a’ in *Newark*. In letter-by-letter filing *Newark* files first because the ‘a’ in *Newark* files before the ‘Y’ in *New York*.

Word-by-word filing is more likely to ensure that similar terms file together. In word-by-word filing every phrase starting with the word cat will file before other words such as catalogue that start with 'cat' but have no meaningful relationship to cats. (Cat naps and cats, however, will be separated by the word catalogue). Letter-by-letter filing ensures that words are not separated because they use spaces or hyphens (for example, in letter-by-letter filing, *on line, on-line* and *online* will file together).

Word-by-word filing	Letter-by-letter filing
dog biscuits	dog biscuits
dog-catchers	dogcarts
dog collars	dog-catchers
dog paddle	dog collars
dog sleds	doges
dogcarts	dogfights
doges	doggerel
dogfights	dogma
doggerel	dog paddle
dogma	dogs
dogs	dogsbody
dogsbodies	dog sleds
dogtooth violets	dogtooth violets

There are also rules for filing symbols (in general they file before letters or are ignored) and ampersand, which is either ignored or filed as if it was written as 'and'. Symbols such as Greek letters may be ignored, but in other cases they are crucial parts of the name and should be filed as if spelt out (for example, *ß blockers*, filed as *beta blockers*).

Numbers are filed after symbols and before letters in ascending numerical order. For example, 6 comes before 55, which comes before 1000.

In all indexing, but in web indexing in particular, you have to beware of automatic computer filing. For instance, in ASCII order upper case letters file before lower case letters, giving two separate A to Z sequences.

Articles (a, an, the) at the beginning of index entries are often ignored in filing. (However, articles in place names such as *The Hague* are taken into account).[62] They may appear in the index entry but not affect the position of the entry, be inverted (moved) to the end of the entry, or be omitted entirely. In all three cases they will be sorted in the same position in the index. For example,

goals
Good Weekend

Good Weekend, The
The Good Weekend
gooey sweets

On the web many lists take articles into account when they sort entries, as this is the default when using computer sorting rules. If automatic computer sorting is used, a double entry should be made without the article. For example, if you have an entry '*The Australian'*, you should also make one for '*Australian* (newspaper)'.

In some indexes and lists (for example, phone books) *Mount* and *Mt* and *Saint* and *St* are interfiled as if the contractions are written in full. In the UK names starting with *Mac*, *Mc* and *M'* may be interfiled. The illogicality of this becomes apparent when the name is not Scottish, for example, *M'bwango, Mary* filed as if spelt *Macbwango*. Modern filing practice is based on the principle 'File on what it says, not what it represents'.

Editing

About one-third of the time taken to create an index should be used for editing. This is the process by which the raw entries are examined and are written and grouped to make them as useful as possible. Editing is necessary because there are many ways to write the same concept. It is only when the whole text has been indexed that final decisions about what to include and how to group entries can be made.

Entries in a raw index might read:

Cultural symbols in play 101
Mathematical symbols in play 18
Play with symbols 25-26
Pretend play 66-67
Symbolic play 67-92
Symbolic play, pretend 104
Symbols in play 11

This might end up as:

play, *see* pretend play; symbolic play
pretend play 66-67, 104, *see also* symbolic play
symbolic play 11, 25-26, 67-92, *see also* pretend play
cultural 101
mathematical 18

During editing you should also check spelling, punctuation and format (for example, italics for scientific names). Check that *See* and *See also* references all

lead to appropriate topics, and make *See* references into double entries if appropriate.

APPENDIX 3: AUSSI WEB INDEXING PRIZE

1996 AusSI Web Indexing Prize

The Australian Society of Indexers started a web indexing prize in 1996. [63]

In reporting the results (www.aussi.org/prizes/webindresults96.htm) [64] Dwight Walker noted that there were many different views on what a web index was. These included:

- Linear back-of-book-style indexes
- Hierarchical Yahoo-style indexes
- Annotated bibliographies
- Ebook tables of contents.

The prize-winners were Alan Wilson, for the Australian Parliamentary Library Index (www.aph.gov.au/parlindx.html, now www.aph.gov.au/find/find_index.htm); Graham Greenleaf, Geoff King and Andrew Mowbray for AustLII (www.austlii.edu.au); and Nancy Guenther (www.chesco.com/~nanguent) and Jeffery Telk Hock Lee. All but the last were still going strong at the end of 2003.

1997 AusSI Web Indexing Prize

The Web Indexing prizewinners for 1997 (www.aussi.org/prizes/webindresults97.htm) [65] were Christabel Wescombe for the University of Sydney Faculty of Education Internet Guide (www.library.usyd.edu.au/Guides/Education); Ann Treacy for the Index for Minnesota websites (www.mnonline.org/uffda; no longer there); and Graham Greenleaf, Geoffrey King, and Daniel Austin for AustLII's World Law Index (including Project DIAL) (www.austlii.edu.au/links/World; now www.worldlii.org).

1998 AusSI Web Indexing Prize

The Web Indexing prizewinners for 1998 (www.aussi.org/prizes/webindresults98.htm) [66] were Lloyd Sokvitne, Liz Holliday and Elizabeth Louden for Tasmania Online (www.tas.gov.au); Marilyn Rowland, for Case-in-POINT (Acxiom Corp); and Peter Langmead for Screen Network Australia (www.sna.net.au). Tasmania Online is still running, but Case-in-POINT is no longer accessible.

1999 AusSI Web Indexing Prize

From 1999 a decision was made that gateway sites, where one webpage links to many other pages scattered throughout the web, would not be eligible for the award as the skills involved in maintaining a gateway are not necessarily the same as those involved in indexing, and it is very hard to evaluate the inclusiveness of a gateway without specialised knowledge of the subject area.

Perhaps because of this limitation, there were only three entries for the prize that year so a decision was made to send each of them the prize and to review all of

the indexes (*AusSI Newsletter*, v.24, n.3 April 2000, www.aussi.org/prizes/webindresults99.htm). The entrants (none of which still exist) were:

- Case In Point Index (the previous year's runner-up; no longer on the web)
- Bowne Internet Solutions (www.bowneinternet.com/en/expublic/id.asp)
- Pre-Raphaelite Critic (www.engl.duq.edu/servus/PR_Critic/Reviews.html)

2000 AusSI Web Indexing Prize

The winner of the 2000 Web Indexing Prize (www.aussi.org/prizes/webindresults00.htm) was Patricia Kennedy for the Site Index: a Subject
Guide to the Queensland Environmental Protection Agency (now www.epa.qld.gov.au/site_information/site_index).

2001 AusSI Web Indexing Award

The Web Indexing Prize was changed in 2001 to a 'Web Indexing Award' (www.aussi.org/prizes/webindexawards.htm). Presentation of the award recognises a high-quality access tool for a single website, and is available to any web index from anywhere in the world so long as it is of a high enough standard. Recipients of the award are entitled to label their index as such for two years, after which they have to reapply. Sometimes the judges have made the award conditional on certain changes being made to the index. Current recipients of the award are:

- Fred Brown (*Writers' Block* magazine, www.writersblock.ca/common/index.htm)
- Fred Leise (PeopleSoft, www.peoplesoft.com/corp/en/indices/site_index.jsp)
- Patricia Kennedy (Queensland Environmental Protection Agency (www.epa.qld.gov.au/site_information/site_index).

APPENDIX 4: GLOSSARY

Words in **bold** indicate that there are separate glossary entries on those topics. This glossary is also available online at www.optusnet.com.au/~webindexing/WebsiteIndexing2Ed.htm.

Absolute addressing: the practice of including a complete **URL** in the link to a webpage: for instance *<A HREF=http://www.aussi.org>* is a link to an absolute address. Links to other websites are always absolute but links within a website may be absolute or **relative**.

Adobe Acrobat PDF, *see* **PDF**

Agent, *see* **Bot**

Anchor (Bookmark): an HTML anchor makes the location in the file at which it is inserted available as a target for a link. It is written in the format *<A NAME=AnchorName>…</A>*.

Automated categorisation: the use of computer software to categorise webpages. It can be done using rule-based methods, in which the system is gradually trained, or by fully automated methods. Taxonomies for categorisation can also be created automatically.

Back-of-book-style indexing: creation of a website index that looks and functions like a back-of-book index. It will usually be alphabetically organised, give detailed access to information, and contain index entries with subheadings and cross-references.

Bookmark, *see* **Anchor**

Boolean 'and': Use of the Boolean operator 'and' in a query means that all of the terms in the query must be present in a document for it to be retrieved. For example, 'automated and categorisation' means that a document must contain the term 'automated' and the term 'categorisation'.

Boolean 'not': Use of the Boolean operator 'not' in a query means that if the search term is present in a document, that document will not be retrieved. For example, 'bear not market' will *not* retrieve a document with the sentence 'Share prices have gone down in the bear market'.

Boolean 'or': Use of the Boolean operator 'or' in a query means that any one of the terms in the query must be present in a document for it to be retrieved. For example, 'categorisation and categorization' means that a document must contain either the term 'categorisation' or the term 'categorization'.

Bot (Agent, Robot): programs with some artificial intelligence that are sent to do a task in lieu of a real person. **Spiders** are one example. They run automatically and act autonomously.

Breadcrumb: link to all levels of the hierarchy above the current location, showing the route a searcher has taken, and the context of the current page. Breadcrumbs allow users to backtrack and to move up the hierarchy.
For example, *Rhinitis>Allergic rhinitis>Perennial allergic rhinitis (Hayfever).*

'Breadcrumbs' is based on the story of Hansel and Gretel, who dropped bits of bread to make a trail to help them find their way out of the forest. (Not that it helped them, as the birds ate the crumbs!)

Breadth: the number of navigation options available at each stage. A home page that provides links to 20 subsections has more breadth than one that says 'Click here to select a department'.

Cascading style sheet, *see* **Style sheet**

Categorisation: the use of **hierarchies** based on words rather than **notations**. Each topic is allocated to a group, and that group is allocated to a more general group, and so on. Searching typically involves moving from more general to more specific topics; for example, to search for information on *children's birthday parties* you might first select the option *Celebrations*, then *Birthdays*, then *Children's parties*. Category structures are fairly arbitrary and may vary widely from one site to another; on a different site you might select *Catering*, then *Parties*, then *Children's parties*, then *Children's birthday parties*, for example. By using techniques such as double posting and cross-referencing, categorised sites can provide for access from several different directions. See also **Automated categorisation; Taxonomy**.

Chunk: smallest unit of content that is used independently and needs to be indexed individually.

Classification: formal established classification schemes – for example, the Dewey Decimal Classification (DC) and Library of Congress Classification (LC) – that use a **notation** to describe classes of information.

Collaborative filtering: personalisation technology that uses recommendation engines to extract trends from the behaviour of website visitors and use that information to present suggestions to searchers. Amazon.com uses collaborative filtering to recommend books on the basis of purchases by other people with apparently similar interests.

Concordance: a (usually alphabetical) list of words from a book or website indicating the locations at which they occur. A concordance differs from an index in that no attempt to filter the source material or sensibly collate the information has been made.

Content management system (CMS): system for the creation, modification, archiving and removal of information resources from an organised repository. Includes tools for publishing, format management, revision control, indexing, search and retrieval.

Controlled vocabulary: a list of terms to be used in indexing (or cataloguing); often a **thesaurus** or **synonym ring**. Use of the same list by all indexers enhances consistency. Most libraries use the *Library of Congress Subject Headings* as a controlled vocabulary for cataloguing books and other library items.

Cost-per-click listing (CPC), *see* **Pay-per-click listing (PPC)**

Crawler, *see* **Spider**

Cross-reference: a ***See* reference** or ***See also* reference** leading the user from one part of the index to another.

CSS, *see* **Style sheet**

Database: a collection of records about individuals. Each record is made up of a number of fields relating to different characteristics of the individual. Many websites and web indexes are generated from databases.

Depth of hierarchy: the number of levels in the navigation hierarchy to the most specific topics. A site where you select 'amphibians' then 'frogs' is shallower than one where you select 'animals', 'vertebrates', 'amphibians', and 'frogs' then 'green tree frogs'.

Depth of indexing: the number of entries and their specificity. A deep index will give direct access to all the topics that have been dealt with in the text. A shallow index will cover major and general topics, but will not index minor topics.

Dialog box: a box into which users of a computer application can enter information.

Directory: a collection of evaluated links to websites, usually categorised by subject. Many **search engines**, such as Yahoo and Google, have associated directories. When directories are limited to information on a specific subject or discipline they are often called **subject gateways**.

Distributed authoring: content creation by people distributed throughout an organisation, not by a centralised group of web specialists or writers. With distributed authoring there is often an expectation that subject **metadata** will also be created by authors. This is distributed indexing.

Document: Any item (not necessarily on paper) that can be indexed or catalogued.

DTD (Document Type Definition): schema specification method for XML documents. A DTD is a collection of **XML** markup declarations that define the structure, elements and attributes that can be used in a document that complies with that DTD. By consulting the DTD a parser can work with the **tags** from the **markup language** that document uses. DocBook is an example of a DTD often used with technical documentation to enable sharing and reuse.

Ebook/Electronic book: standalone document intended for on-screen reading on a PC or a handheld device, either a dedicated 'reader' or a general purpose Personal Digital Assistant (PDA).

Editorial results: **search engine** hits dependent on content and not influenced by payment.

Embedded indexing: indexing method in which tagged index entries are inserted into document files. **Tags** are used to bracket blocks of text and to show headings and subheadings for index entries. Tagged index entries are not seen in the printed version, but can be compiled by software to make an index. If parts of the document are removed or rearranged the tagged index terms go with them. The index can then be recompiled to give an updated version. Embedded indexing is

more time-consuming than normal indexing, but is efficient for documents that change often, or are not complete when indexing starts.

Facet: grouping of concepts of the same inherent type, for example, processes, disciplines, people, materials, places, and times.

Faceted metadata classification: breaking subjects into standard component parts (facets) and presenting these to users as search options. A topic such as wine might be divided into the facets such as *country of origin, variety* and *price*. In the best faceted search systems the user is provided with feedback about the number of terms retrieved at each stage.

False drop: document that is retrieved by a search but is not relevant to the searcher's needs. False drops occur because of words that are written the same but have different meanings (for example, 'squash' can refer to a game, a vegetable or an action).

Field searching: ability to limit a search by requiring that the search term is present in a specific 'field' (category of data) in the record. Field searching is often done with categories such as author and date that are common to most records.

Filing order: rules used for ordering (sorting) index entries. When a computer performs the sequencing it is often called *sort order*. [67]

Gateway, *see* **Subject gateway**

Global navigation: generally applicable navigational links (for example, Search; Site Map) available from all pages of a website.

Granularity: level of detail at which information is viewed or described. The more granular an access tool, the smaller the chunks of information it leads to. An index linking to specific paragraphs is more granular than a table of contents or site map linking to specific pages.

<HEAD> section: The <HEAD> section of an HTML document is placed at the top of the page between an opening tag, <HEAD>, and a closing tag, </HEAD>, and contains metadata about the document itself, not the content that will be displayed on the page. It is followed by the <BODY> section.

Hierarchy: a series of ordered groupings moving from broader general categories to narrow specific ones. In a web directory you may only see one level of the hierarchy at a time. When you select a topic you are then shown the options at the next level. *See also* **Taxonomy**; **Thesaurus**.

Hit highlighting: highlighting of the words in a results list which resulted in a document being retrieved by the search.

HTML: hypertext markup language. The majority of webpages are made up of ordinary text 'marked up' with instructions in HTML which determine how the text is displayed by the user's browser: for instance, the HTML code 'Huey, <B>Dewey</B> and <I>Louie</I>' appears in a browser as 'Huey, **Dewey** and *Louie*'. HTML is also used to display graphics and define links to other sites and locations. *See also* **XML**.

Hypertext link, *see* **Link**

Indented style index: indented indexes start each subheading on a new line, indented under the main heading. For example:

names
    indexing rules for 41-42
    keyword searching and 5

Index entry: record in an index, consisting of a **main heading** and any associated locators, **subheadings**, and **cross-references**. This means the whole 'metadata' example below is *one* entry. When indexers charge by the entry they usually define each cross-reference or **locator** as an entry, meaning the 'metadata' sample below would contain six entries, made up of one cross-reference and five locators.

metadata, *see also* thesauri
    Dublin Core 15, 33-37
    misspellings useful in 14
    website structure derived from 99-101, 105

Indexing: often used to refer to the automatic selection and compilation of 'meaningful' words from a website into a list that can be used by a **search engine** to retrieve pages. This list is more properly called a **concordance**. As this procedure involves no intellectual effort indexers distinguish their own work by calling it intellectual indexing, manual indexing, human indexing, or **back-of-book-style indexing**.

Information architecture: design of the structure of information systems, particularly websites and intranets, including labelling and navigation schemes.

Information foraging: seeking information according to its adaptive value. Information foraging theory analyses trade-offs in the value of information gained against the costs of searching based on the analogy of 'foraging for wild food'.

Information scent: visual and linguistic cues that indicate to a searcher whether a website has the information they seek, and help the searcher navigate to the required information. Information scent is a component of **information foraging**.

Instantiation: the electronic or physical manifestation of a resource.

Internet: a global electronic communications system allowing public access to email, newsgroups, chat and the **web**.

Intranet: a local network with restricted access that uses some or all of the same systems and software as the Internet.

Keyword: a) In the **search engine** section keywords are words that are used to search for a topic. Also called 'search terms'. b) In the **metadata** section, keywords are subject metadata terms.[68]

Keyword searching: typing significant words and phrases that relate to a topic into a **search engine**. For example, to find information about your pet, Gerby, you might type the keywords *gerbils* and *sand rats*. If you wanted scientific information you could try the scientific terms *Gerbillus*, *Tatera*, *Taterillus*

gracilis and so on. If you needed to find more general information you could broaden your search with the terms *domestic animals* and *pets*.

Legacy data: data stored on older computer systems or in older formats that remains behind as the legacy of outdated technologies. It can be difficult to integrate into newer systems.

Link: a block of text or a graphic appearing on a webpage, which a user can click with the mouse pointer to cause an event to happen. This usually involves being taken to another webpage or another part of the same page.

Live file: the copy of an electronic document that is currently being worked on, for example, by a writer or indexer. All changes must be made to the live file. If an indexer worked on one copy of a document, and an editor on another, the changes made by one of them would have to be incorporated into the document worked on by the other. (Live in a different context means that the file has been loaded onto the web and made available to users).

Local link, *see* **Relative addressing**

Local navigation: links that are specific to a section of a website, compared with **global navigation** which is available from all parts of a site.

Locator: the part of an **index entry** that tells the user where to look for information. In a book index locators are usually page numbers (but can also be references to items, paragraphs and so on). In a website index they are direct links to the information. The links can be the **main heading** or **subheadings**.

Main heading: heading at the beginning of an **index entry**, either used alone or modified by **subheadings**. The main heading is an entry point into the index. (**Cross-references** are the other entry points).

Markup language: a way of depicting the logical structure or **semantics** of a document and providing instructions to computers on how to handle or display the contents of the file. **HTML, XML** and **RDF** are markup languages. Markup indicators are often called **tags**.

Metadata: structured data about data, which may include information about the author, title and subject of web resources. Metadata is added in the **<HEAD> section** of the webpage or is stored in a database. It is available for searching but is not displayed on the page.

Multi-purposing, *see* **Single sourcing**

Namespace: a closed set of names or a place where a **schema** (set of names) is stored. Namespaces are identified via a **URI** (for example, a **URL**) and are a mechanism to resolve naming conflicts. Within a given namespace all names must be unique, although the same name may be used with a different meaning in a different namespace.

Navigation, *see* **Supplementary navigation; Taxonomy**

Notation: code used in formal **classification** schemes. In the Dewey Decimal Classification the notation *993* refers to the history of New Zealand, and the notation *994* refers to the history of Australia.

Ontology: specification of a conceptualisation of a knowledge domain. An ontology is a **controlled vocabulary** that describes objects and the relations between them in a formal way, and has a grammar for using the vocabulary terms to express something meaningful within a specified domain of interest. The vocabulary is used to make queries and assertions. Ontological commitments are agreements to use the vocabulary in a consistent way for knowledge sharing. Ontologies can include glossaries, **taxonomies** and **thesauri**, but normally have greater expressivity and stricter rules than these tools. A formal ontology is a controlled vocabulary expressed in an ontology representation language.

Page number, *see* **Locator**

Pageless index: electronic index in which **index entries** link directly to the text they refer to rather than listing page numbers for the user to find.

Paid inclusion: payment for inclusion of a site in a **search engine's editorial listings**, without an artificial boost in ranking.

Paid placement (Advertising): listing in **search engine** results where advertisers pay for a guaranteed high ranking, usually dependent on specified **keywords** being used in a search. These listings are usually segregated from **editorial results** and labelled to indicate that they are ads. Also known as 'pay for placement', 'pay for performance', or **pay-per-click listings (PPC)**. The last two terms refer to the usual method of payment, which is based on the number of times the link is selected ('clicked') by a user. See also **sponsorship**.

Paid submission: payment for guaranteed consideration of a site for inclusion in a **directory**.

Pay-per-click listing (PPC): **search engine** advertising in which payment is based on the number of times the website is selected (clicked) from the results list. Is used in **paid placement** advertising and **paid inclusion**. Also known as cost-per-click listings (CPC).

PDF (portable document format) file: a way of displaying documents in the form in which they will be printed. Used when for legal or other reasons an exact copy of a printed document must be made available electronically. PDF files are displayed and printed using Adobe Reader software.

Pick list: list from which a computer user can select terms. Usually found in a menu, form or **dialog box**.

Precision: the relevance to the searcher of the items that are retrieved. If a search retrieves one hundred documents of which ninety-five are very relevant, that search has high precision.

RDF: a formal data model from the **W3C** using **XML** for the description of web resources using machine readable metadata. It has potential for use in the **semantic web**.

RDF schema (RDFS): defines a set of metadata properties (for example, 'Creator') that can be associated with resources.

Recall: the proportion of relevant information that is retrieved by a search. If a search only retrieves one hundred relevant documents out of three thousand that are available, that search has low recall. If it retrieves all the available documents on the topic, it has high recall.

Recommendation engine, *see* **Collaborative filtering**

Relative addressing (Local links): linking to another page on the same website through a local address rather than a URL: for instance, a link back to the homepage might take the form *<A HREF=/home.htm>Home page</A>*.

Repurposing, *see* **Single sourcing**

Robot, *see* **Bot**

Run-on (run-in) style index: run-on indexes list all subheadings in sequence, separated by punctuation marks such as semicolons. For example:

> names: indexing rules for 41-42; keyword
> searching and 5

Schema: a description of the structure and rules a document must satisfy for an **XML** document type. Includes the formal declaration of the elements that make up a document.

Search engine: server that 'indexes' webpages, stores the results, and uses them to return lists of pages which match users' search queries. *See also* D**irectory**; **Indexing**.

Search log: record of searches performed.

Search term, *see* **Keyword**

***See also* reference:** directs index users to related topics that could be consulted in addition to the topic they are currently at: for example, 'beds, 26, *see also* cots'

***See* reference:** a way of indicating to a user that they should look elsewhere. A *see* reference may point to two or more locations: for example, 'rodents, *see* mice; rats'. The choice of which terms to use and which to refer from depends on the language of the material being indexed and the target audience.

Semantic web: project of the **W3C** in which automated methods based on quality **metadata** are envisaged to replace much human searching of the web. Relies on **ontologies**, **XML** and **RDF**.

Semantics: meaning. If a computer understands the semantics of a document, it understands the meaning, rather than just interpreting a series of characters.

Single sourcing: using one content repository to generate documents in different formats. The content only needs to be written and maintained in one place, but can be output in formats such as **HTML** and RTF (rich text format) as required. Also known as multi-purposing. Repurposing refers to the sequential output of content in different formats using different software tools.

Site indexing, *see* **Website indexing**

Site map: Overview of the navigational structure of a website, acting like a Table of Contents, and used to orient users and show them the scope of the site. Site maps can be textual or visual. Usually each location is an active link, enabling a user to move directly to that section. Site maps can also be important sources of links for search engine spiders to follow.

Sort order, *see* **Filing order**

Specificity: narrowness of terms. 'Maternity leave' is more specific than 'parental leave', which is itself more specific than 'leave'. Book indexers normally aim to use a term with the same specificity as the information being indexed, although users often search with broader terms.

Spider (Crawler, Web crawler): bot that visits publicly accessible websites following all links it comes across collecting data for **search engine** 'indexes'. A spider discovers new sites and updates information from sites previously visited. A spider can also be used to check links within a website.

Sponsorship: sponsored ads are normally located in a separate boxed and labelled section at the top of a **search engine's** results list. See also **paid placement**.

Stemming: expansion of searches to include plural forms and other word variations.

Style sheet: a block of text in which one or more formats for webpage display are defined. This may include redefinitions of standard formats such as <H1> or new formats specific to that page or site. Style sheets may be embedded in a particular webpage or stored as a separate text file to which some or all of the webpages on a site are linked. Where several style sheets are linked to one page, the order in which they are named determines which ones take precedence in the case of conflicting definitions. These are called cascading style sheets (CSS).

Subheading/subentry/subdivision: headings that follow a **main heading** to modify it. In the index sample below, metadata is the main heading, and 'Dublin Core' and 'misspellings useful in' are subheadings.

> metadata, *see also* thesauri
> Dublin Core 15, 33-37
> misspellings useful in 14
> website structure derived from 99-101, 105

Subject gateway: a **directory** limited to a specific subject area such as *education*, or *Tasmania*. Sometimes called a 'portal'.

Subsite: a distinct section of a website that might warrant its own navigational systems**.**

Supplementary navigation: information access methods separate from the basic site structure or browse navigation. *See also* **Back-of-book-style indexing; Site map.**

Synonym ring/list: sets of synonyms. If someone searches using one synonym from a set (ring), the other words or phrases in the set are also included in the search.

Table of contents, *see* **Site map**

Tag: a piece of text that describes the **semantics** or structure of a unit of data (element) in **HTML, XML** or other **markup language**. **Tags** are surrounded by angle brackets (< and >) to distinguish them from text. 'Tags' is also used to describe the code indicating index entries in **embedded indexing.**

Target, *see* **Anchor; URL**

Taxonomy: controlled vocabulary used primarily for the creation of navigation structures for websites. Often based on a **thesaurus**, but may have shallower **hierarchies** and less structure, for example, no related terms. See also **Categorisation**.

Thesaurus: a structured list of approved subject headings (preferred terms) showing the relationships between them. The relationships include broader (parent) terms, narrower (child) terms, and related terms. The thesaurus also shows terms that are *not* to be used in indexing (nonpreferred terms) with references to the terms that should be used instead (for example, 'automobiles, *see* cars'). See also **taxonomy**. (According to the NISO standard on thesaurus construction, the plural of 'thesaurus' can be 'thesauri' or 'thesauruses'. We used 'thesauruses' in the first edition of this book, but have yielded to reviewer preferences and the interests of brevity.)

Three-click rule: the three-click rule suggests that if a user has to click more than three times to find the information they are looking for they will give up the search.

Topic map: tool for representation of model-based data on the web for enhanced access. Topic maps are based on topics, associations and occurrences. In comparison with **RDF**, topic maps are developed separately from the documents they refer to.

Unlimited aliasing, *see* **Synonym lists**

URI (Uniform Resource Identifier): unique identifier of the location of a resource. In many cases the URI will be a **URL** (that is, a website address, for example: http://www.aussi.org).

URL: Uniform Resource Locator – the address of a webpage. For example, http://www.aussi.org (also written as www.aussi.org).

Usability: efficiency with which a **user** can perform required tasks with a product, for example, a website. Usability can be measured objectively via performance errors and productivity, and subjectively via user preferences and interface characteristics. Web design features that affect usability include navigation design and content layout.

User: also known as visitor, participant, actor, searcher, employee, customer, and client.

Visualisation: graphical presentation of information, often dependent on **categorisation** or clustering techniques to bring out patterns in the information.

W3C/World Wide Web Consortium: an international consortium of companies which develops specifications, guidelines, software and tools for the web using open standards to ensure interoperability. It is the chief standards body for HTTP and **HTML**. The W3C was founded in 1994 by Tim Berners-Lee, the original creator of the web.

Web: a vast collection of files accessible to the public through the **Internet**, viewed through a browser, and connected by hypertext **links**. Also known as the World Wide Web, or W3.

Web crawler, *see* **Spider**

Web indexing: a) **search engine** indexing of the world wide web; b) creation of **metadata**; c) organisation of web **links** by category; d) creation of **website indexes**.

Website indexing: the creation of a **back-of-book-style index** to an individual website (or subsite, web **document** or **intranet**).

XML: a relative of **HTML** that specifies not only the appearance, but also the type of material on display. For example, names can be given the XML **tags** <first-name> and <last-name> making for more flexible searching and display. **RDF** is written using XML.

ENDNOTES

All websites were accessed in December 2003 unless otherwise noted.

[1] Published in 1923 by Grant Richards Ltd., St. Martin's Street, London. Now available via Project Gutenberg Australia (http://gutenberg.net.au/plusfifty.html) in text or zipped format. The title is derived from a quote from Sir Thomas Browne (no relation), who wrote 'That immortal essence, that translated divinity and colony of God, the soul.'

[2] Evans, Joanne. 'Exploring Bright Sparcs: creation of a navigable knowledge space.' *Cataloguing Australia*. v.25 n.1/4 1999, p.156.

[3] Murphy, Tom. 'Exploring fiction and poetry through indexing'. *The Indexer* v.23 n.4 October 2003.

[4] Site temporarily available 10 December 2003.

[5] Chulick, Mary Ann (presented by David Ream). 'The mysteries of fiction indexing'. *The August Indexer: proceedings from the 1999 International Conference of the Australian Society of Indexers*. Melbourne, Vic: Australian Society of Indexers, 2000, pp. 48-54.

[6] For a detailed discussion see: Robertson, James. 'Metrics for knowledge management and content management, February 2003, www.steptwo.com.au/papers/kmc_metrics/index.html.

[7] Published as a chapter in Carroll, John M., Ed., *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction*, Cambridge, MA: MIT Press, 1987, pp. 80-111.

[8] Rosenfeld, Louis and Morville, Peter. *Information architecture for the World Wide Web*. Cambridge, Mass: O'Reilly, 2002, p133.

[9] Rosenfeld and Morville. *ibid.* p123.

[10] Fetters, Linda. 'A book-style index for the web: the University of Texas Policies and Procedures website'. *The Indexer* v.21 n.2 October 1998, pp. 73-76.

[11] Rowland, Marilyn. 'Plunging in: creating a web site index for an online newsletter'. In Brenner, Diane and Rowland, Marilyn, eds. *Beyond Book Indexing: how to get started in web indexing, embedded indexing, and other computer-based media*. Medford, NJ: Information Today, 2000.

[12] Ryan, Christine Nelsen and Henselmeier, Sandra. 'Usability testing at Macmillan USA'. *Key Words* v.8 n.6 November/December 2000, pp198-202.

[13] Matthews, P. and Bakewell, KGB. 'Indexes to children's information books'. *The Indexer* v.20 n.4 October 1997.

[14] Jörgensen, Corinne and Liddy, Elizabeth. 'Information access or information anxiety? – an exploratory evaluation of book index features'. *The Indexer* v.20 n.2 October 1996.

[15] Olason, Susan. 'Let's get usable! Usability studies for indexes'. *The Indexer* v. 22 n.2 October 2000, pp. 91-95.

[16] Maislin, Seth. 'Ripping out the pages'. In: Brenner, Diane and Rowland, Marilyn, eds. *Beyond Book Indexing: how to get started in web indexing, embedded indexing, and other computer-based media*. Medford, NJ: Information Today, 2000, pp.43-57.

[17] Wittmann, Cecelia. 'Subheadings in award-winning book indexes: a quantitative evaluation' *The Indexer* v.17 n.1 April 1990.

[18] Olason, Susan, *op. cit*.

[19] Ryan and Henselmeier, *op. cit*.

[20] Jörgensen and Liddy, *op. cit*.

[21] This works fully in Internet Explorer, but only highlights the first reference when using Mozilla as the browser.

[22] Ryan and Henselmeier, *op. cit*.

[23] Jörgensen and Liddy, *op. cit*.

[24] For a discussion of information architecture software in general, see Peter Morville's article 'Software for information architects' (19 February 2001, http://argus-acia.com/strange_connections/current_article.html). Although some of the examples are outdated, the general categories remain important.

[25] In this book the leading 'http://' has been removed to make URLs shorter and less likely to wrap around to the next line. These URLs are still functional. Where the URL has a different beginning, for example, 'https://' it has been kept in full.

[26] Also published as: Wyatt, Michael. 'Software review: CINDEX for Windows 1.0 and SKY Index 5.1 (Professional edition).' *Australian Society of Indexers' Newsletter* v.22 n.8 Sept 1998, pp71, 74-79 and v.22 n.9 Oct 1998, pp.81, 84-87; updated as 'Review of CINDEX for Windows 1.5 and SKY Index Professional 5.1 (revision).' *Australian Society of Indexers' Newsletter* v.24 n.9 October 2000, pp.81, 84-87.

[27] Browne, Glenda. 'Automatic indexing and abstracting.' Paper given at the Indexing in the Electronic Age Conference, Robertson, NSW, 20-21 April, 1996. Published in *Online Currents* June 1996 and in *LASIE* v.27 n.3 September 1996, pp. 58-66. Also at www.aussi.org/conferences/papers/browneg.htm.

[28] Hodge, Gail M. and Milstead, Jessica L. *Computer Support to Indexing*. Philadelphia, PA: NFAIS, 1998. Reviewed in *Australian Society of Indexers' Newsletter* v.23 n.2 March 1999. Review also at www.aussi.org/anl/1999/02mar/computer.htm.

[29] Information on Macrex provided by Gale Rhoades from the Macrex Support Office, North America in a note to the Index-L discussion group on 2 July 2000.

[30] I indexed the first edition of this book using SKY Index and then embedded the index entries in the text so they would be available if we revised the book. We have changed and added so much content, however, that the thought of updating the index rather than starting from scratch was daunting, and the book has been totally re-indexed. While embedded indexing is the obvious solution for many documents, it is a very difficult job to pick up the threads of an old index and try and weave them into a new one.

[31] Site not available 10 December 2003.

[32] Fetters, Linda. 'A book-style index for the web: the University of Texas Policies and Procedures Website'. *The Indexer* v.21 n.2 October 1998.

[33] Browne, Glenda. 'Something wiki this way comes'. *Online Currents*. v.17 n.2 March 2002.

[34] Chulick, Mary Ann (presented by David Ream), *op. cit*.

[35] The images of the AusSI website index are from the 2001 edition of this book, as we have access to nonpublicly available source images from that time. The current index is similar in construction but is indexed by another indexer.

[36] Walker, Dwight. 'Web indexing prize'. *Australian Society of Indexers' Newsletter* v.22 n.1 Jan/Feb 1998, p.7.

[37] Rosenfeld and Morville. *op. cit.*

[38] If this address gives the message 'You are not authorised to view this page', you can read a previous online draft without figures dated 1996 at http://dlis.gseis.ucla.edu/research/mjbates.html. Also published in *JASIS* v.49 November 1998, pp. 1185-1205.

[39] Lakoff, George. Chicago: University of Chicago Press, 1987.

[40] Browne, Glenda. 'Automatic categorization' *Online Currents* Part 1: v.18 n.1 January/February 2003, pp. 17-19, 22; Part 2: v.18 n.2 March 2003, pp. 7-8, 10-11; Part 3: v.18 n.3 April 2003, pp.9-10, 12-14; 'Brochure engineering and buzzword marketing about automatic categorization and taxonomy creation' v.18 n.3 April 2003.

[41] Maislin, Seth. Personal communication. Also discussed in 'Earning online trust.' *The Indexer* v.22 n.1 April 2000, pp.29-30. On 4 November 2000 a Google search for *millennium* retrieved about 3,420,000 sites, a search for *millenium* retrieved about 1,240,000 sites and a search for *milennium* retrieved 7,510 sites.

[42] Walker, Alan. 'Building an Australian thesaurus: Indexing Australian historical photographs.' *Cataloguing Australia* v.19 n.3/4, 1993, pp.268-279.

[43] For example, on 24 November 2003, a 'Basic' search for 'barbecue' retrieved 27 records, which were displayed immediately. A 'Subject' search for 'barbecue' retrieved a hit titled 'St. James' Roto-barbecue' which was listed under the heading 'SelectWks', while a 'Subject' search for the plural 'barbecues' brought up the thesaurus display shown in Figure 20 with both of the terms 'barbecues (cookers)' and 'barbecues (events)' listed under the heading 'SelectWks'. There were 144 hits for 'barbecues (events)' and 6 for 'barbecues (cookers)', indicating the increased recall available using a thesaurus. This search also indicates that searching using singular or plural terms retrieves different types of results.

[44] Stumm, Deborah. 'When is a forest fire a bushfire?: Towards an Australian pictorial thesaurus.' *Cataloguing Australia* v.25 n. 1/4, 1999, pp.140-146.

[45] Deacon, Prue. 'Is classification needed to supplement subject indexing in metadata for a web gateway?' Proceedings of the 2nd International Conference of the Australian Society of Indexers 27-29 August 1999. Melbourne: Australian Society of Indexers, 2000; Deacon, Prue. 'Simplicity vs Structure: which way for the Dublin Core?' Cataloguing Australia v.25 n1-4 March/December 1999. Papers from the 13th National Cataloguing Conference, pp. 36-43; and Deacon, Prue. 'Changes in the *Health and Ageing Thesaurus* and reindexing in HealthInsite'. *Australian Society of Indexers' Newsletter* v.27 n.10 November 2003. Presented at: Indexing the World of Information: International Conference of the Australian Society of Indexers, Sydney, 12-13 September 2003.

[46] This site requires the Java plug-in. An alternative site that is not being updated is www.slais.ubc.ca/courses/arstlibr512/winter2000/database1.htm.

[47] American Society of Indexers list of online thesauri (www.asindexing.org/site/thesonet.shtml), 'Classification schemes and thesauri on-line' (www.fbi.fh-koeln.de/fachbereich/labor/Bir/thesauri_new/indexen.htm), 'Controlled vocabularies, thesauri and classification systems available in the WWW' (www.lub.lu.se/metadata/subject-help.html), 'Compendium of thesauri' (www.darmstadt.gmd.de/~lutes/thesauri.html), Multites (go to

www.multites.com and select 'Web Thesaurus'), 'Thesauri of the United Nations Educational, Scientific and Cultural Organisation' (UNESCO, www.ulcc.ac.uk/unesco/index.htm), and wordHoard (www.mda.org.uk/wrdhrd1.htm).

[48] Site viewed November 2003. Is apparently moving to a permanent domain, but is not currently available on the web.

[49] The definition of ontology is derived from documents by Tom Gruber (www-ksl.stanford.edu/kst/what-is-an-ontology.html),Victor Lombardi (www.noisebetweenstations.com/personal/essays/ metadata_glossary/metadata_glossary.html) and Search Tools (www.searchtools.com/info/classifiers.html).

[50] The definition of ontology is derived from documents by Tom Gruber (www-ksl.stanford.edu/kst/what-is-an-ontology.html),Victor Lombardi (www.noisebetweenstations.com/personal/essays/metadata_glossary/metadata_glossary.html) and Search Tools (www.searchtools.com/info/classifiers.html).

[51] The XML examples in this article were adapted from the following tutorial: Altenburger, Anitta, 2001, 'Topic Maps. Bond University', http://topicmaps.bond.edu.au/tutorial1.

[52] Biezunski (www.infoloom.com/tmsample/bie0.htm) gives the example of the Spanish letter 'll', which sorts in Spanish after the letter 'l'. Thus words starting with 'lo' file before words starting with 'll'. This can be arranged by using appropriate sort names.

[53] Has to be viewed in an XML-aware browser such as Internet Explorer. In Mozilla the XML code is displayed.

[54] For more details see Jonathan Jermey, 'Enquire within on everything; getting questions answered on the Internet', *Online Currents*, v.18 n.2, March 2003.

[55] Apparently the section of the Inktomi FAQ that states that Inktomi does not crawl frames is out-of-date (http://hotwired.lycos.com/webmonkey/01/23/index1a_page3.html).

[56] Some clients also reported a decline in ranking when they dropped out of the program. Since the report was based on a small sample number (30) and anecdotal evidence it is hard to draw definite conclusions.

[57] 'Selection criteria for quality controlled information gateways: work package 3 of Telematics for Research project DESIRE (RE1004). Version 1.1 May 1997' (www.ukoln.ac.uk/metadata/desire/quality). Other useful information from the DESIRE project is at www.desire.org.

[58] To learn more read one of the following books, or take a course in indexing (course providers are listed in Appendix 1. The most commonly used indexing texts are Wellisch, Hans. *Indexing from A to Z.* 2nd ed. Bronx, New York: HW Wilson Co, 1995; Mulvany, Nancy. *Indexing books.* Chicago: University of Chicago Press, 1994; and Booth, Pat. *Indexing: the manual of good practice*. London: K.G.Saur, 2001.

[59] *Anglo-American Cataloguing Rules* 2nd ed. 1988 revision. Eds Michael Gorman and Paul W. Winkler. Chicago, Illinois: American Library Association, 1988.

[60] Information and documentation – Guidelines for the content, organization and presentation of indexes. Australian/New Zealand Standard. AS/NZS 999:1999 (ISO 999:1996).

[61] Moncrieff, Lynn. 'Indexing computer-related documents'. In: *Beyond Book Indexing: how to get started in web indexing, embedded indexing, and other computer-based media.* Brenner, Diane and

Rowland, Marilyn, eds. Medford, NJ: Information Today, in association with the American Society of Indexers, 2000, p.23.

[62] For a discussion of current rules about filing 'the', and a proposal that it should be considered more important in filing, see: Browne, Glenda. 'The definite article'. *The Indexer*, v.22 n.3 April 2001, pp. 119-122.

[63] Walker, Dwight. 'AusSI Web Indexing Prize'. *The Indexer* v.20 n.1 April 1996.

[64] Walker, Dwight. 'AusSI Web indexing prizewinners'. *The Indexer* v.20 n.3 April 1997.

[65] Walker, Dwight. 'Web indexing prize 1997'. *The Indexer* v.21 n.1 April 1998.

[66] Walker, Dwight. 'Web indexing prize 1998'. *The Indexer* v.21 n.3 April 1999.

[67] The terms **filing** and **sorting** are often used interchangeably. Wellisch (*op.cit*) describes **filing** as the most general term, relating to the arrangement of all graphic signs. **Alphabetization** refers specifically to the sequence of letters in words, phrases and abbreviations. **Sorting** is used when the arrangement of graphic signs is performed by a computer. It includes filing of symbols, numbers, and letters, including separate sequences for uppercase and lowercase letters (p.133).

[68] Keyword is defined in Wellisch (*op.cit*.) as '**keyword**. 1. A *word* occurring in the *natural language text* of a *document* or its *surrogate* that is considered significant for *indexing* and *information retrieval*. 2. Any word not on a *stop list*, occurring in a verbal segment of a document or in a *title, abstract* or *subject heading* assigned to it.' Keywords are used as *access terms* in keyword indexes such as *KWAC, KWIC*, and *KWOC indexes*' (p39). Keyword is defined in *Beyond Book Indexing (op.cit)* as '**keyword**. A word that has special meaning within a specific context. In <META> tags, keywords are terms that describe the webpage and may be used by search engines to retrieve the pages' (p137).

SUBJECT INDEX

This is a detailed index to subjects referred to in the text; authors have not been indexed. Page numbers in *italic* refer to figures. Word-by-word filing has been used. This index was created by Glenda Browne.