References

Al-Safadi, Lilac A. E., 2009: Auto Classification for Search Intelligence. *World Academy of Science, Engineering and Technology* 49, 849–852 http://www.waset.org/journals/waset/v49/v49-150.pdf

Eysenbach, G. and Ch. Kohler, 2003: What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet. *AMIA Annual Symposium Proceedings* 2003: 225–229. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1480194/

Biography

Derek Sturdy is a knowledge management consultant at Tikit Ltd. Not everyone at Tikit would agree with everything in this article, which represents his views, not the company's. He also runs the Galilee Project, a not-for-profit research operation collating and verifying past climatic information for use in validating climate models. derek.sturdy@tikit.com.

Legal Information Management, 10 (2010), pp. 28–33
© The British and Irish Association of Law Librarians
doi:10.1017/S1472669610000277

Indexing of Free, Web-based Electronic Resources

Abstract: The internet provides access to a huge amount of information, and most people experience problems with information overload rather than scarcity. Glenda Browne explains how indexing provides a way of increasing retrieval of relevant information from the content available. Manual, book-style indexes can be created for websites and individual web documents such as online books. Keyword metadata is a crucial behind the scenes aid to improved search engine functioning, and categorisation, social bookmarking and automated indexing also play a part.
Keywords: indexing; internet; metadata; classification; automatic indexing

Introduction

Despite the success of search engines, human-created indexing remains important to enhance access to the most relevant information for searchers' needs. This article covers a range of access methods, focussing mainly on the areas in which human input is most significant, including website indexes (back-of-book style A–Z indexes to websites and web documents), metadata creation, social bookmarking, and classification, followed by a brief discussion of automated indexing. Searches for 'website indexing' also retrieve information about webwide search engines, but they are not discussed here.

1. *A to Z indexes*

Website indexing became important in the 1990s, as indexers, librarians and web managers experimented with different approaches for making the information they were providing on the internet more accessible. The tools for creating A–Z indexes have changed over time, from simple HTML coding to HTML Indexer and other

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press

specialist software. On the downside, website indexes appear to be an especially vulnerable part of a website, and many quality indexes that have been created over the last decade no longer exist.

A–Z indexes on the web range from simple alphabetical lists, often of names, to complex, carefully constructed book-style indexes that provide access to selected, important information. They often refer to the content of one site only, but may be used to provide access to a number of websites on related topics.

The advantages of using book-style indexes on the web is that they are familiar to users, they are easily browsable, and they provide guidance in the form of subheadings, cross references and so on. The problems with book-style indexes include the cost involved in their creation, and the need for continual updating if the content being referred to is changed. Research has also found that people do not always know what to expect from a website index, and often do not choose to use one (Barnum et al., 2004), so clear naming of the link (eg, 'A–Z index') or having the main index links on every page can be useful.

Creators of A–Z style website indexes employ a range of software products to create them. These include HTML Indexer[1], which uses embedded indexing to create updateable website indexes and Help files; XRefHT32[2], which is free, open source software that can be linked to a thesaurus and HTML/Prep[3], which is used to create website indexes from existing indexes. Indexers also use dedicated indexing software packages such as SKY Index, CINDEX and Macrex to create indexes which they can then format as appropriate. Unfortunately there is no generally-available single sourcing indexing program that makes it possible to create the index and output it in a variety of formats.[4] A–Z website indexes can be implemented with a Content Management System (Leise, 2002) but because this is not always straightforward, it may have contributed to the decline in the use of indexes on websites.

The following sections provide examples of A–Z indexes on the web.

Indexes not linked to online material

Indexes may be mounted on the web without providing links to the content being indexed. These indexes are functionally equivalent to print indexes, with the advantage of providing broad availability.

One example is the Dust Diseases Tribunal 'DDT Subject Matter Index'[5]. The Lawlink: Law Reform Commission New South Wales: Report 83 (1997) Partial Defences to Murder: Provocation and Infanticide index[6] also has no direct linking from index entries, but there are links to chapters of the book at the bottom of the page.

The University of Auckland library has created the first index to festschriften on common law[7]. Festschriften are collections of articles written to honour a person or event. This index is not linked to full text content, but is linked to catalogue records at the University of Auckland library.

Websites that are simple lists

Website indexes that are simple lists include two created by the New South Wales Parliament:

- NSW Legislation Website[8].
- Parliament of New South Wales: Former Members Index A–Z[9].

Lawlink has created a number of list-style indexes, including some that provide access to the same information from different viewpoints. For example, you can access enterprise agreements through the name of the agreement (Lawlink 2009a) and also through the names of the participants to the agreements (Lawlink 2009b).

Some simple lists are alphabetised by a word that occurs in the middle of the title, eg, in the Lawlink Equal Time index (Lawlink 2009c). This can be problematic when more than one keyword in the article is relevant, eg, 'Homosexual discrimination – Discrimination and older gays; surviving aged care' is indexed at 'H' but not at 'D' or 'A'.

Manually created subject indexes to websites

Manually created subject indexes require more attention than simple lists, during both creation and maintenance. The website index of the Australian and New Zealand Society of Indexers[10] won the ASI Web Indexing SIG award in 2009[11]. It features an alphabet bar on every page for immediate access to the index, and is created using plain HTML.

Special challenges occur with bilingual sites, but can be managed, e.g., the Indexing Society of Canada (ISC) has a bilingual site index[12] with a column each for English and French entries.

The Province of British Columbia Hansard provides a subject index, members index, and business index for each session of parliament[13]. Indexes from previous sessions are available separately.

The Australian Law Librarians' Association (South Australian Division) has collaboratively created a subject index to South Australian Legislation[14] (originally funded by The Law Foundation of South Australia). Links are provided from the index to the legislation on the SA Attorney-General's website[15]. Index terms are based on a thesaurus which was compiled using terms from existing indexes such as the Index to Subject Matter of Victorian Legislation.

The BNA (Bureau of National Affairs) Labor Relations Reporter Index includes the LRR General Index and a Table of Cases for LRR Manuals[16]. It offers progressive display of the index, i.e., when you start examining the index only the main headings are displayed, but as soon as you select one of the main headings, the subheadings are opened up for display. This means you can quickly search for a main heading before browsing in more detail, especially important for legal and similar indexes which use classification and long lists of subheadings.

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press

Manually created subject indexes to books

Online books can also have their indexes transferred online. These are straightforward to maintain as the content does not continually change, although the links have to be created initially. One example is the index to the book Orders of Magnitude[17] which has been online for over a decade.

In some cases structure from the manual index does not transfer to the online version, although the direct linking ameliorates problems from this, e.g., the index to *Search User Interfaces* by Marti Hearst[18]. Periodicals can be indexed in a similar way to online books although, if they are ongoing publications, the index will require regular updating. Examples of long-term periodical indexes include *Rochester History*[19] and *Writer's Block*[20].

As well as standard alphabetical order, people have experimented with the use of maps to lead into content, e.g., the index to the *Handbook of Western Australian Aboriginal languages south of the Kimberley Region*[21].

2. Metadata

Metadata refers to data about data. It is a broad term, which in the context of indexing the web usually refers to the provision of keywords which reflect the subject content of web pages and make that content easier for searchers to find. Metadata is usually hidden 'behind the scenes' and works to enhance search without users necessarily knowing that it exists, although it can also be displayed to users for browsing. Document metadata is also important to lawyers. It refers to the metadata stored automatically or added by writers to word processed and other documents and is not discussed here[22].

Metadata keywords can be easier to create and maintain than a handcrafted A-Z index. This is because each metadata keyword can be selected in isolation from other keywords, and does not need to be changed to fit a specific index. Book-style indexing, on the other hand, is always fluid, and terms may need to be changed to better reflect the nature of a topic in context with the other topics being indexed. This is especially true with subheadings, which are all worded individually.

Metadata keywords are usually taken from a thesaurus (or other controlled vocabulary), so decisions about the terms to be used have already been made, although indexers need the skills to select the most appropriate terms. Cross references are created outside the system. The use of a controlled vocabulary helps to ensure consistent retrieval no matter what words are used in the pages (ie, it improves recall). Metadata also helps ensure that the most relevant pages rank highly (ie, it improves precision). It therefore overcomes some of the problems of free-text search. Creating a thesaurus from scratch is a time-consuming, specialist task. Pre-existing thesauri can be purchased or, in some cases, freely downloaded from the web. There is a comprehensive list at the Taxonomy Warehouse[23].

Because metadata is hidden, it is not always possible to point to sites that are using metadata, nor to examine what difference the metadata is making. There is, however, some research on the topic. For example, the Attorney General's Department of New South Wales has published a report (Lawzone 2009) on the impact of metadata on resource discovery on the internet.

There are a number of general and specialist metadata standards in use, including the Dublin Core (most widely used) and government standards such as AGLS (Australian Government Locator Service), which is based on Dublin Core. The Attorney General's Department of New South Wales publishes the Justice Sector Metadata Standard, which is based on AGLS, thus promoting compatibility with other services. Some of the issues specific to the creation of metadata for legal materials are the need to include jurisdiction; to make reference to the Act or Regulation covered by the secondary material; and to indicate whether the information is primary (Acts, case law) or secondary (guides to the law). Details of the Justice Sector Metadata Standard, along with an example, are provided on their website[24].

A good example of browsable metadata is Montague Institute Review. The A–Z index to the site[25] is dynamically updated from metadata derived from a thesaurus and stored in a database.

The Flare Index to Treaties[26] is a searchable database of basic information on over 1,500 of the most significant multilateral treaties from 1856 to the present, with print information and, if available, a link to the online publication. The index is published by the Institute of Advanced Legal Studies, University of London. Technical information on the index database is available in an article by Whittle and Clinch (2009).

The Australian Domestic & Family Violence Clearinghouse Good Practice database[27] provides a central access point for shared information about domestic violence interventions. The database is searchable by keyword, program title, and record ID. In addition, you can browse by sub-categories of Target Group, Service Types and Interventions. When you select an entry for a project it shows the keywords that have been allocated to that program. This helps with search refinement, as the user can then browse the additional keywords in programs that have been retrieved in return for their original search. Hits are sorted by jurisdiction.

3. Classification, categorisation, portals and taxonomies

As well as indexes and metadata which give access to specific topics, many content providers use classification or categorisation to allow users to browse through collections of related information.

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press

In the early days of the internet a number of websites attempted to provide classified access using library classifications such as the Dewey Decimal Classification. Most of these projects have not continued, but BUBL LINK[28] still provides access to topics by classification number (as well as alphabetically). Information on Law is available from 340 to 349.

Portals/directories appear to be relatively common in the legal field. Some examples are:

- InfoLaw (UK)[29]
- Lawlinks (UK)[30]
- Legal Resources in the UK and Ireland maintained by Delia Venables[31]
- Find Legal Answers (New South Wales)[32]
- FreeAdvice (US)[33].

The Public Library of Law (PLoL) is the largest free law library in the world, and allows searching for information that is scattered across many different websites[34]. There are also many national websites that provide access to legal materials from around the world[35] and from specific countries[36]. Davis (2001) has written on the processes involved in the development of AustLII.

4. Social bookmarking/ Folksonomies

Social bookmarking (tagging) brings the power of the people to indexing the web, as people tag sites of interest to them, which may help other users find those sites. The lists of tags are somewhat like an index, but lack features such as cross-references and subheadings, and often lack consistency. I tag website indexes I find with the tags website_indexes, website_indexes_names and so on (see del.icio.us.com/GlendaIndexer/website_indexes).

Other social bookmarking sites include Technorati[37] (for blogs) and Citeulike[38] (for articles). When searching for content relevant to indexing, I have had occasional serendipitous finds, but have not yet found them a good way to do a thorough search.

5. Automated indexing

Automated indexing has been used alongside and instead of manual indexing for a number of years. It involves the analysis of text using computer algorithms. Analysis of terms in their context enables the identification of concepts as well as simple keywords. Losses of work for indexers in some areas have been replaced by newer roles in other areas, so it is difficult to predict the balance for the future, although high labour costs and increasing technological sophistication suggest that the role of manual indexing may decrease.

Automated indexing using software such as packages from Autonomy is used by many legal companies for the management of unstructured data (Autonomy 2009), including automated taxonomy generation. Recommind is another program used in law firms for enterprise search to index email, people finding and so on, as well as straight document retrieval (Alber 2007).

Automated book indexing

While automated indexing has been used with large-scale collections such as intranets, it has generally proven ineffective for standalone products such as single books, where much smaller 'content chunks' are being indexed. Kluwer Belgium has taken a new approach to this problem, dividing their content into small, reusable 'content units' and classifying content at acquisition (i.e., when it is created as a 'chunk' of content) rather than when all of the content has been gathered into one product. Indexing keywords are derived from a multilingual French and Flemish thesaurus covering terms for laws, regulations, jurisprudence, and comment[39].

Conclusion

The Declaration on Free Access to Law made by legal information institutes of the world (http://www.worldlii.org/worldlii/declaration) affirms the importance of access to legal information for the promotion of justice and the rule of law. The range of legal indexes on the web enhances access to this material.

Footnotes

[1]http://www.html-indexer.com

[2]http://publish.uwo.ca/~craven/freeware.htm

[3]http://www.levtechinc.com/ProdServ/LTUtils/HTMLPrep.htm

[4]Proprietary systems have been created to do this, but even with those it is not an easy process. For example, NRMA Insurance in Australia used an online Help system to create advice for customer consultants in Help format, HTML for the intranet, and print for those without computer access.

[5]http://www.lawlink.nsw.gov.au/lawlink/caselaw/ll_caselaw.nsf/pages/cl_ddt_subject_index

[6]http://www.lawlink.nsw.gov.au/lrc.nsf/pages/R82INDEX

[7]http://www.library.auckland.ac.nz/databases/learn_database/public.asp?record=iclf

[8]http://www.legislation.nsw.gov.au/maintop/scanact/sessional/NONE/0

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press

[9]http://www.parliament.nsw.gov.au/prod/parlment/members.nsf/V3ListFormerMembers
[10]ANZSI; www.anzsi.org
[11]http://www.web-indexing.org/award/2009.htm
[12]http://www.indexers.ca/siteindex.html
[13]http://www.leg.bc.ca/hansard/hansindx/38th3rd/2007-subjmhds.htm
[14]http://www.alla.asn.au/sa/sisal/index.html
[15]http://www.legislation.sa.gov.au
[16]http://www.bna.com/lrr/lrrindx.htm
[17]http://www.informationuniverse.com/ordersmag/orders.htm
[18]http://searchuserinterfaces.com/book/sui_index.html
[19]http://www.rochester.lib.ny.us/~rochhist/indexa.htm
[20]http://www.writersblock.ca/common/index.htm
[21]http://coombs.anu.edu.au/WWWVLPages/AborigPages/LANG/WA/section1.htm#1.4
[22]The following websites discuss legal issues to do with document metadata: http://www.legalethics.com/?cat=2, http://www.mrblaw.com/CM/TechnologyAndTheTrialLawyer/Technology-And-The-Trial-Lawyer-What-It-Is.asp, http://www.llrx.com/features/metadata.htm.
[23]http://www.taxonomywarehouse.com
[24]http://info.lawaccess.nsw.gov.au/lawaccess/lawaccess.nsf/pages/jsms
[25]http://www.montague.com/Public/indexes.htm
[26]http://ials.sas.ac.uk/library/flag/introtreaties.htm
[27]http://www.austdvclearinghouse.unsw.edu.au/good_practice.html
[28]Bulletin Board for Libraries, http://www.bubl.ac.uk
[29]http://www.infolaw.co.uk
[30]http://www.kent.ac.uk/lawlinks
[31]http://www.venables.co.uk
[32]http://www.legalanswers.sl.nsw.gov.au
[33]http://www.freeadvice.com
[34]http://www.plol.org/Pages/Search.aspx
[35]http://www.worldlii.org
[36]e.g., www.austlii.edu.au
[37]http://technorati.com
[38]http://www.citeulike.org
[39]http://www.mondeca.com/index.php/en/success_stories/kluwer_belgium_groupe_wolters_kluwer

References

All websites were accessed in December 2009.

Alber, John. (2007) Search at the foundation of the enterprise *LLRX.com* 27 August 2007, http://www.llrx.com/features/searchenterprise.htm.

Autonomy. (2009) Autonomy strengthens position as the de facto standard for global law firms. Washington, DC: Autonomy press release, http://www.autonomy.com/content/News/Releases/2009/0826.en.html.

Barnum, Carol, et al. (2004) Index versus full-text search: a usability study of user preference and performance *Technical Communication* 51(2) pp. 185–206.

Davis, Madeleine. (2001) Building a global legal index: a work in progress. *The Indexer* 22(3).

Lawlink. (2009a) Lawlink: Justice & Attorney General: Industrial Relations Commission: Index to enterprise agreements – clause 44C – http://www.lawlink.nsw.gov.au/lawlink/irc/ll_irc.nsf/pages/IRC_procedures_legislation_awards_index_all_having_effect_as_ea.

Lawlink. (2009b) Lawlink: Justice & Attorney General: Industrial Relations Commission: Index of participants to awards, contract determinations and enterprise agreements – clause 44C – http://www.lawlink.nsw.gov.au/Lawlink/irc/ll_irc.nsf/pages/IRC_procedures_legislation_awards_participant_index_all).

Lawlink. (2009c) Lawlink: Justice & Attorney General: Anti-Discrimination Board: Index of Equal Time articles, http://www.lawlink.nsw.gov.au/lawlink/adb/ll_adb.nsf/pages/adb_et_index.

LawZone. (2009) LawZone: a new way of searching: improving community access to legal information on the Internet, http://info.lawaccess.nsw.gov.au/lawaccess/lawaccess.nsf/b302b703a1328b7e4a2565e8002658dc/c8603d9cdb84688c4a256bdc000b5b32/$FILE/metadata.pdf.

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press

Leise, Fred. (2002) Improving Usability with a Web-site Index. www.web-indexing.org/article-leise.htm or http://www.boxesandarrows.com/view/improving_usability_with_a_website_index. Originally published July 15, 2002 on Boxes and arrows (www.boxesandarrows.com).

Whittle, Steven and Clinch, Peter. (2009) FIT for purpose – the new FLARE Index to Treaties. *LLRX.com* 19 April 2009, http://www.llrx.com/features/flareindex.htm.

Further reading

All websites were accessed in December 2009.

American Society of Indexers Web Indexing SIG, http://www.webindexing.org/web-index-examples.htm.

Browne, Glenda and Jermey, Jonathan. (2004) *Website indexing: enhancing access to information within websites* 2nd edn. Adelaide: Auslib Press 2004. Also for sale as an e-book through RMIT Informit (http://www.informit.com.au), as a PDF at http://www.webindexing.biz and as a PDF or print-on-demand book from Lulu (http://www.lulu.com).

Hedden, Heather. (2007) *Indexing specialties: Web sites*. Medford, NJ: Information Today, in association with American Society of Indexers.

Lamb, James. (2006) *Website Indexes: visitors to content in two clicks, or website indexing with XRefHT32 freeware*. Available as a print-on-demand book from http://www.lulu.com.

Stephenson, Mary Sue. (2005) Indexing resources on the WWW, www.slais.ubc.ca/resources/indexing/database1.htm.

Biography

Glenda Browne has been an indexer of books, journals, databases, online help and websites since 1988. She has written or co-written three books – *Website Indexing* (2nd ed), *The Indexing Companion* and *The Indexing Companion Workbook: Book Indexing* (www.webindexing.biz). Glenda has taught indexing to indexers and editors through institutions such as Macleay College and Macquarie University. She has also taken on many committee roles in ANSZI (Australian and New Zealand Society of Indexers) over the last 20 years.

glendabrowne@optusnet.com.au

https://doi.org/10.1017/S1472669610000277 Published online by Cambridge University Press