9 PERIODICALS INDEXING

Antoinette Kotze

Abstract

This chapter concentrates on the Index to South African Periodicals (ISAP) as an example of periodicals subject indexing database. It covers compilation, application and use by the National Library, selection and the function of its main indexing fields and discusses ways of fostering improvement. It also includes an addendum on the principles of periodical indexing.

Introduction

The Index to South African periodicals (ISAP) is an example of a periodicals subject indexing database. This chapter gives information on the following:

- ISAP as a product of periodicals subject indexing – a national indexing database
- how ISAP is compiled
- how ISAP is applied and used by the National Library and other database providers
- selection of information in non-scholarly periodicals illustrated with examples
- function of its main indexing fields
- improvements to ISAP

What is ISAP?

ISAP is a product of periodicals subject indexing that has become a national indexing database.

The Library Association of South Africa started ISAP in 1940, and it was continued by the then Johannesburg Public Library in 1945. In 1986 the then State Library acquired authority over ISAP in terms of a ministerial decision. The State Library was merged with the South African Library, Cape Town, on 1 November 1999 to form the National Library. As a product of periodicals subject indexing, ISAP represents a main function of the national bibliographic service as determined in the National Library Act 92 of 1998 (chapter 1, item 4(1)(b)(ii)).

ISAP does not reflect all legal deposit for South African periodicals, but comprises a comprehensive selection of mainly scholarly journals and technical and subject general periodicals. A limited number of popular magazines useful in schools or adult basic education are included.

As a subject index, ISAP provides access to information through natural language keyword terms, controlled thesaurus terms and abstracts, as well as a basic bibliographic

description to an article title, author and journal reference. Research and use value in the South African context form the basis for selecting subject terms for inclusion.

As a South African indexing tool, ISAP is a unique source to national periodical subject information.

How ISAP is Compiled

Contracts

The policy of the National Library in compiling ISAP is based on contracts with specialised institutions and qualified individuals.

Currently the Council for Scientific and Industrial Research (CSIR), University of South Africa (Unisa) Law Library and the National Institute for Theology and Religion (NITR) are institutions involved in indexing for ISAP. Professional subject expertise at indexing institutions can benefit ISAP by for example proposing additional periodicals for indexing that represent reliable research or new developments in a subject area. Institutions also represent a broader spectrum of institutional input in ISAP at national level.

Required qualifications for indexers include an acceptable library (or other subject) qualification with at least four years' experience in general indexing, classification and subject cataloguing. Experience in indexing of periodicals is preferred. Basic bibliographic cataloguing knowledge is required. Expertise in the use of indexing software has to be acquired. An indexer is required to have appropriate qualifications in subject areas of journals allocated for indexing or to acquire the necessary subject knowledge.

Qualified individuals undergo training in ISAP indexing and are considered for contracts with the National Library only on constant and regular delivery of trial work according to ISAP standards.

Standards, Rules and Guidelines

The following standards, rules and guidelines apply to ISAP:

- SAMARC minimum level (programmatically converted to MARC21)
- ISBD(S): International Standard Bibliographic Description for Serials
- AACRII cataloguing rules
- ISO 9115:1987 (E): Documentation – Bibliographic identification (biblid) of contributions in serials and books
- ISAP guidelines for indexing drawn up by the National Library and indexing institutions, and the ISAP thesaurus initiated by Johan van Wyk
- Q & A or Inmagic indexing program
- ISO 5963: Documentation – Methods for examining documents, determining their subjects and selecting indexing terms
- ANZI Z39.14: American National Standard for writing abstracts
- Library of Congress list of languages and language codes
- alphabet tables for transliteration of non-Roman and African languages

Standards for indexing electronic journals are not included because these journals are not indexed on ISAP as yet. The National Library is investigating this option.

Indexing Approach

Against the background of contracts, standards, rules and guidelines, indexers aim to index substantial and useful information according to subject. With selection of information, user needs feature prominently in the focus of ISAP.

Information in scholarly journals is fully indexed. Technical and subject general periodicals (e.g. *Computing SA*, and *African Security Review*) are selectively indexed, focusing on article-type presentations and appropriate subject material. Information that is useful for learners and neo-literates is selected for indexing in popular magazines (e.g. *Drum, Huisgenoot*).

The latter two groups of periodicals and magazines can include much information presented for a passing readers' market as opposed to essential information useful for research and recorded for this purpose. Information for a readers' market falls under ephemera to be excluded from ISAP.

A high priority is placed on indexers' professional insight and integrity with regard to selecting information for indexing. Regular contact with indexers is maintained to interpret new problems and reflect resulting decisions in the guidelines.

Periodical Selection

A broad spectrum of subject fields is covered by ISAP periodicals, for example natural sciences, humanities, education, medicine, technology, agriculture, theology and religion, public administration, and business and industry.

The list of periodicals inherited from the then Johannesburg Public Library (now the Greater Johannesburg Library Services) in 1945 is still in existence. Since 1986, it has been extended on a regular basis according to the focus and procedure for ISAP, accepted and developed by the National Library.

Focus

The aim pursued by ISAP periodicals has been to represent the following:

- a selection of periodicals published in South Africa, regarded as significant and useful
- important scientific, technical and subject general periodicals especially at scholarly level, appropriate for representation on ISAP
- a limited number of popular magazines with information relevant to the needs of schools and basic adult education

Extension of Periodicals Lists

It is procedure to extend the list of periodicals throughout a financial year. This ensures a continuous flow of periodical supply to indexers.

The list of periodicals will be extended for the following reasons:

- if a title is closed down
- to keep up to date with new periodical publications
- to follow up users' requests for additional titles
- to represent new areas in subject fields appropriate to the focus of ISAP

Additional Periodicals

Before being added to ISAP, periodicals are assessed against the following criteria:

- extent of support for a specific title (especially applicable with a survey on users' proposals for additional titles)
- need for representation on ISAP, considering the extent to which an additional periodical's subject area would complement existing coverage on ISAP (with the actuality of a subject area also taken into account)
- availability of funds

Proposal Procedures for Additional Periodicals

Users are invited to propose additional periodicals for indexing on the ISAP system. The following information is required to accompany such proposals:

- full periodical title and publisher
- brief motivation for its inclusion in ISAP (e.g. that it would complement existing subject coverage or support a user need)
- the first and latest issues of the relevant periodical (if it is not available at the National Library) or photocopies of the title, content and editorial pages of the named issues to give an indication of the scope of the periodical

Process of Compilation

Supply of Periodicals for Indexing

Periodical titles as stipulated in an indexing contract are supplied for indexing usually on a weekly basis, but also over longer intervals depending on availability of periodicals. The frequency of periodical publications could influence the regularity of periodical supply for indexing. Where late publication of current periodicals hinders periodical supply, retrospective issues of selected periodicals are sent instead to help maintain a regular flow of periodicals for indexing. It is mainly scholarly journals that are indexed retrospectively in this way.

Record Delivery, Quality Control and Supply of Records for Access

Indexed records are delivered on a weekly basis or over longer intervals, depending on periodicals received for indexing.

Printouts of newly processed records are checked weekly for quality control. Errors and corrections are discussed with indexers if necessary. Alternatively they receive copies of printouts to inform them of corrected minor errors – spelling mistakes and the like. It may

be necessary to resend records prior to acceptance. Corrections, other than the weekly ones, are dealt with on a continuous basis.

Following the weekly record processing and quality control, newly accepted and corrected records are loaded on the National Library's in-house ISAP database and on its Internet version, ISAPOnline, available at www.nlsa.ac.za/service_isaponline_about.html

At the same time, these records are supplied for licensed access to outside database providers such as the South African Bibliographic and Information Network (SABINET), National Inquiry Services Centre (NISC) and Potchefstroom University.

Claims and invoices for newly accepted records are paid on a monthly basis or as invoices arrive.

Application and Use of ISAP by the National Library and Other Database Providers

Access to ISAPOnline

The National Library allows non-subscribers gratis searching on ISAPOnline, where limited information is given in a record's periodical reference field. Subscribers to ISAPOnline have access to complete periodical reference information. The National Library awarded a prize of one year's free subscription for ISAPOnline to Vosloorus High School in Gauteng, for having the highest number of learners at the Careers, Education and Training Faire on 13-14 March 2003.

Support to Other National Library Projects and Functions

The ISAP database is a frequently used source for information searches undertaken for users in the National Library's reference and interlending sections. ISAP is made available to users conducting their own searches on online public access catalogue (OPAC) computers in the Reference Section. Complete articles to ISAP records are provided to SABINET users by the Interlending Section and to non-SABINET users by the reference sections in Pretoria and Cape Town.

ISAP supports the development of a complete periodical collection by identifying missing issues of existing periodicals and requesting substitute copies if necessary through the Periodical Section.

Support is given to the Legal Deposit Section by alerting them to the existence of titles that they have not yet acquired.

With ISAP's contribution, the National Library records South African documentary heritage for posterity.

Generation of Income through Use of ISAP by Other Database Providers and Support of Specific Research

The National Library earns some revenue through the selling of ISAP records or by royalties received. ISAP also indirectly stimulates income derived from information and document delivery functions.

ISAP records made available to SABINET, NISC and the University of Potchefstroom for their licensed use facilitate national and international access to ISAP. Users in the public and private sectors and at universities are reached in this way. ISAP is the most used indexing database on SABINET and the South African Studies CD-ROM published by NISC. The University of Potchefstroom has reported that ISAP stimulates the use of periodicals.

Research in specified subject areas was supported by the sale by special request of selected sets of ISAP records to Taung Agricultural College and the University of Pretoria Law Department.

Selection of Information in Non-scholarly Periodicals illustrated by Examples

Types of Periodicals and Selection of Information

Scholarly journals consist of formal articles accompanied by abstracts and often with keywords. These journals are indexed fully on ISAP and do not pose problems with the selection of information for indexing.

As mentioned above, non-scholarly periodicals (e.g. technical or other subject periodicals) are indexed selectively on ISAP. These periodicals can include much news and product information which are generally regarded as ephemera and excluded from ISAP. However, this kind of information may also include useful facts. These are assessed for inclusion against selection criteria based on practical experience acquired by ISAP.

Criteria for Selection of Information in Non-scholarly Periodicals

ISAP's main aim when it comes to selecting information in non-scholarly periodicals remains to record useful and substantial information applicable to South Africa.

Experience has revealed the following:

- The periodical's subject field or focus should be assessed for inclusion and use by its target market. For example, before including it in ISAP, a periodical focusing on computers but also offering information about human resources should first be weighed against the likely coverage in a periodical that focuses solely on human resources.
- Information presented for a passing reader market (to be omitted from ISAP) should be distinguished from information that could be useful for future reference and should therefore be included on ISAP.

The following can indicate information that is intended merely to attract the attention of readers (as opposed to information that is inherently useful for recording):

- when the information highlighted on the cover page does not even feature on the contents page
- when an item highlighted on the cover page is published as a news item (not as part of a full article)

Process of Selecting Information in Non-scholarly Periodicals

Indexers selecting information from non-scholarly periodicals are expected to consider these relevant points:

- Keep in mind the policy and aim of indexing. It is ISAP policy to omit news items. However, some news items may happen to contain supportive information. This illustrates the problem of vague and indistinct areas in the selection of information for indexing. One option used for handling such information is to refer to a relevant news item in the abstract of a record pertaining to an article – (or usual) – type entry covering a corresponding topic, as shown in Example 1. The news item topic should also be covered in the keyword field of such an entry. Example 1:

TITLE: Servers anchor desktop Linux plans : analysis.
[by/deur] Scannell, E.

ABSTRACT: *Reports from Infoworld(US)* that Linux's overall success is no longer measured by its desktop performance but rather on how it is being used as a mainstream server. Looks at developments taking place in Linux and how its performance in the desktop and server environment is likely to change in the future. **Also see news item "MS/Linux TCO controversy still rages" (p.11)**

IN: *Computing SA* (2003):23:6 p.20 Feb 17
(ISSN) 0254-2196: P11736:

KEYWORDS: SuSE; Linux; Servers; Operating systems; Red Hat

THESAURUS: Computer software; Communication; Business

Another option for ISAP is to make one compound entry comprising more than one corresponding short item, as demonstrated by Example 2. The number of compound entries and information included in compound entries should be limited to avoid unnecessarily cluttering information or obscuring clarity in content. Example 2:

TITLE: Shoden launches modular storage system; EMC Celerra NS600 integrates NAS, high-availability features: **storage & networking: technical**
ABSTRACT: Reports on the following announcements: Shoden Data Systems on the local availability of the Hitachi Freedom Storage Thunder 9500V Series storage systems, with details on specifications and application areas of and technology used in the storage devices. EMC on Celerra NS600 network server combining enterprise NAS operating system with CLARiiON storage architecture for high level information protection and ease of information management

IN: *Computing SA* (2003):23:6 p. 30,31 Feb 17
(ISSN) 0254-2196: P11736:
KEYWORDS: Hitachi Data Systems; Shoden Data Systems; Thunder 9500V; Storage systems; EMC; Celerra NS600; Networked storage; Information protection; Information management.
THESAURUS: Information storage and retrieval; Computer hardware; Technology

- Consult the contents pages for an idea of what the publisher regards as prominent information in the relevant market or industry for that periodical.
- Get an overview of the content as a whole and note how information is presented. This could assist in a process of 'information orientation' in order to select information for a periodical's target group or to fulfil the purpose of the database.
- Assess the usefulness of full reports and articles against selected appropriate short items to which references can be made in the abstracts of records for reports and articles that have been selected.
- Choose meaningful and necessary combinations of more than one short item in single compound entries. Information can be grouped according to form, such as a question-and-answer column, or to subject, where information by more than one author relates to the same or similar topic. Non-related information can also be grouped in compound entries when necessary, as long as the information is clearly identifiable and uncluttered.
- Subdivisions introduced by the publisher can be used in two ways to keep titles in compound entries clearly grouped or identified: 'Storage & networking: technical' is a subdivision found in *Computing SA* (2003).

In Example 2 it is used as a subtitle joining the individual short items.

In Example 3 it is used on its own as a title. In this case each listing of an individual short item in the Abstract is followed by the item's specific page numbering in brackets:

TITLE: Storage & networking: technical
ABSTRACT: Reports on the following announcements: Shoden Data Systems on the local availability of the Hitachi Freedom Storage Thunder 9500V Series storage systems, with details on specifications and application areas of and technology used in the storage devices **(p.30)**. EMC on Celerra NS600 network server combining enterprise NAS operating system with CLARiiON storage architecture for high level information protection and ease of information management **(p.31)**
IN: *Computing SA* (2003):23:6 p. 30,31 Feb 17
(ISSN) 0254-2196: P11736:
KEYWORDS: Hitachi Data Systems; Shoden Data Systems; Thunder 9500V; Storage systems; EMC; Celerra NS600; Networked storage; Information protection; Information management.
THESAURUS: Information storage and retrieval; Computer hardware; Technology

- Evaluate reports by non-South African authors for their relevance to the South African situation to decide how useful it would be to include them. Origin of such information should be acknowledged in the entry's abstract (see Example 1).

Function of the Main Indexing Fields on ISAP

Indexing fields for ISAP are based on the South African Machine Readable Cataloguing (SAMARC) minimum level. Computer programs have been written to convert to MARC 21, to enable ongoing electronic exchange of records.

The subject indexing fields for ISAP comprise keywords in natural language terms, broad thesaurus terms and an indicative abstract. Information in these fields is complementary because in each field the topic of an article can be described on another or further level.

The abstract can be used to indicate the article's focus or intended target group and provide information on how the article is presented, or on the article's relevance to previous or future articles. For further detail see chapter 13 on abstracting.

Natural language terms enable specific description of a topic according to terms used in an article. Additional appropriate terms that are equivalent or related to or descriptive of the article's topic are also allocated here to fully reflect different aspects of a specific topic covered in an article.

The ISAP thesaurus consists mainly of broad disciplinary terms. More than one broad term can be allocated to reflect the cross-disciplinary treatment of a topic.

Such a hybrid system of combining specific natural language keyword terms with broad disciplinary thesaurus terms facilitates a precise recall in searching, as illustrated in the following table.

Precise Recall Made Possible by Specific Keywords Combined with Broad Thesaurus Terms

KEYWORD	+THESAURUS	RESULT
Indexing		26
Indexing	+ Information services	9
Indexing	+ Computers	6
Indexing	+ Medicine	1

Bibliographic description fields include author, article title and periodical reference. The National Library shelf number for the periodical (P number) and its International Standard Serial Number (ISSN) are also made available in the periodical reference field, to facilitate correct periodical identification and expedient provision of complete articles available at the National Library by reference and document delivery services.

Conclusion

ISAP is a dynamic database that continues to develop according to users' needs and technological advancement.

The National Library has systematically worked on improving ISAP according to development plans, identified needs and criticisms or recommendations received. The list below includes examples of improvements.

Periodicals

- The list of 'Approved South African journals, assessed in the reporting year 2005', which is Appendix 4 to revisions of the Department of Education's 'Policy and procedures for measurement of research output of public higher education institutions', has been checked to identify the latest periodicals added to ISAP for more complete coverage of approved scholarly journals.
- In line with reflecting information in a developing democracy, South African published periodicals with a focus on Africa and applicable to situations in South Africa have been added (e.g. *African Finance Journal, African Energy Journal, African Human Rights Law Journal, African Journal of AIDS Research, South African Journal of African Languages, African Studies*).
- Useful and appropriate in-house journals are now included.
- Coverage on music and musicology was extended with titles such as *The South African Music Teacher, South African Journal of Musicology (SAMUS), Scenaria, Musicus* and *Ars Nova*. Other titles complementary to this subject area have also been added (e.g. *South African Journal of Music Therapy, South African Journal of Cultural History, Vuka, South African Theatre Journal (SATJ), Literator*).
- Results of a user survey in 1995 prompted the addition of more subject general periodicals and some popular magazines. Some of these periodicals cover information for developing communities and neo-literates. However, scholarly journals remain ISAP's main focus.

In consultation with the Gauteng Department of Education, periodicals used by school learners and their facilitators, and which could also benefit developing communities and neo-literates, were added to ISAP. A paper copy with records of this subset of educational periodicals dating from January 1997 to August 1998, titled *Edu-ISAP*, was introduced to schools in Gauteng as well as to some public libraries. But this project was discontinued because of lack of interest.

Frequency

- ISAP has been processed and updated on a weekly basis since so requested by SABINET and law librarians in 2002.

Specific Subject Terms

- The ISAP indexing fields described above illustrate the use of a hybrid system of natural language keywords combined with broad disciplinary thesaurus terms. This enables specific searching, as opposed to Library of Congress subject headings (LCSH) previously used and criticised as not specific enough.

Access and Visibility

- Apart from access to ISAP by other database providers referred to previously, the National Library provides Internet access to ISAPOnline at affordable prices in line with different levels of users. Non-subscribers may search ISAPOnline free of charge although access is limited, with incomplete periodical references.
- Promotional presentations on ISAP were delivered at the ISC Schools Library Conference (St Stithians College, Randburg) in 1998, the Community Library & Information Technology Symposium (Boksburg Community Library) in 2001, and at the Association for Southern African Indexers and Bibliographers (ASAIB) Indexing workshop (Unisa) in 2003.
- ISAP participated in exhibitions such as the Educating SA Exhibition at Gallagher Estate (Midrand, Gauteng) in 2000, the Pretoria Show in 2002, and the Careers, Education & Training Faire at Kyalami (Gauteng) in 2003.

Quality Control and Technological Advancement

Towards the end of 2003, the National Library accepted an offer by NISC to sponsor the use of their advanced NISCBase software for ISAP indexing. Training in the new software has started and preparatory work to run ISAP on the new system is under way. It is believed that the use of NISCBase software will greatly improve future content and quality control.

BIBLIOGRAPHY

Behrens, SJ. 1996. Recommendations for the improvement of national bibliographic services in South Africa. *South African Journal of Library and Information Science*, 64(2):81-85.

Bloomfield, M. 2001. Indexing – neglected and poorly understood. *Cataloguing & Classification Quarterly*, 33(1):63-75.

De Wet, LJ. 1955. Compiling the Index to South African Periodicals. *South African Libraries*, 22:93-96.

Ellis, D, Ford, N & Furner, J. 1998. In search of the unknown user: indexing, hypertext and the World Wide Web. *Journal of Documentation*, 54(1):28-47.

Lor, PJ & De Beer, JF. 1994. The State Library in the service of South African science and scholarship. *South African Journal of Science*, 90, August/September:462-466.

Mai, J-E. 2001. Semiotics and indexing: an analysis of the subject indexing process. *Journal of Documentation*, 57(5):591-622.

South Africa. 1998. The National Library of South Africa Act, Act 92 of 1998. Pretoria.

Von Beck, M. 1989. South African bibliographic control tools. *Mousaion*, 7(1):76-86.

Winter, JS. 1967. The Index to South African Periodicals. *South African Libraries*, 34(3):85-89.

9 ADDENDUM: PRINCIPLES OF PERIODICAL INDEXING

Marlene Burger

Abstract

The purpose of this addendum is to provide general information and guidelines for the indexing of continuing resources (serial publications). The focus is on periodicals (journals, magazines). It is a summary of sections from various sources such as Ashcroft and Langdon (1999), Beare (1999), Booth (2001), Lancaster (2003), Taylor (2004) and Wellisch (1995). These sources may also be used for the indexing of other serial publications such as newspapers, trade journals, consumer and special-interest magazines, yearbooks, annual reports, directories, and indexing and abstracting journals.

Introduction

A periodical (journal, magazine) can generally be defined as a continuing publication issued in successive parts with numerical or chronological designations. The medium may be paper, electronic or any other medium. There are two types of periodical index:

- individual indexes to individual periodicals
- broad indexes to a group of periodicals

In the first instance, the publisher of the periodical prepares an index, usually for a volume at the end of a year's run of the periodical. These indexes are prepared under the direction of the editor of the periodical. The approach to this kind of indexing can range from simple, uncontrolled vocabulary to a complex indexing system with a thesaurus.

Indexes of national or even world periodical literature are major enterprises that usually consist of large databases. Such indexes cut across many titles and specific subject areas. ISAP is an example of a national periodical index (see the first part of this chapter for detailed information).

While most books have one or two authors, periodical indexes will list thousands of authors. Periodical indexes are based on the same principles and have the same general objectives as book indexes; but because their scope is broader, they present a number of unique problems. For example, preparing a book index is a well-defined operation, with a beginning and an end. It generally focuses on a general topic, and it can usually be prepared entirely by a single indexer. But periodical indexes are open-ended projects, usually done by a number of indexers, covering perhaps years and with shifts in subject emphases and indexing objectives. Consistency, vital to quality indexing, becomes a challenge. Any one issue of a periodical may deal with a number of unrelated topics by several different

authors, written in different styles and aimed at different users. The periodical index must bring order out of these differences.

With the indexing of periodicals, there is very likely a policy guide that indicates what is to be indexed and what not. Such a policy is necessary in order to maintain consistency when many indexers are involved. The index should include the names of articles' authors and coauthors. Generally, every article should be indexed – these form the core of the periodical. The decision what to index beyond the articles depends on the nature of the periodical, its readers and the economics involved. Some periodicals have only the articles indexed and others may have book reviews, news and gossip columns, and advertisements indexed. The inclusion of such material depends on the degree of pertinent information in them. For example, some editorials contain significant material and often important research analysis, and should therefore be indexed. In general, conference programmes, letters to the editor and the like are not indexed.

A book index is finished and published together with the book (at the back of the book), but with some periodicals the finished index may not be produced until some time after publication of the issues to which it relates. It is also possible for a periodical to be indexed twice, namely currently (either at or close to the date of publication) or retrospectively (to cover material not initially included, or to reflect the needs of a new group of users).

Formats

In this section we concentrate on formats other than printed periodicals.

Ashcroft and Langdon (1999) report an annual rate of increase of about 30 per cent in the number of electronic periodicals, with more than 7 000 titles currently available (please note that this was in 1999). These figures include periodicals on the Internet or CD-ROM or provided by other electronic means. In the academic and research fields in particular, there has been a noticeable increase in the number of periodical titles being published electronically as well as, or instead of, in print. For some titles, the electronic content exactly matches that of the printed version, but for others provision may be of selected articles only, or just a contents list. A smaller number, referred to as e-journals, are available online only. Some titles make the full articles freely available on the Internet, whereas some provide abstracts for free consultation, and others are available only to subscribers or after online registration with the producers.

Publishers may increasingly decide to economise by no longer producing printed indexes and instead directing searches to the online versions of their periodicals, which can be searched in various ways. Indexes may be provided that are similar to the printed versions, but without locators such as page numbers: searchers just click on a heading (author or subject) in the index for the relevant article to be displayed on the screen. Another approach is to use full-text searching by keywords, although using natural language for searching has pitfalls, for example occurrence of words with the same spelling but different meanings (e.g. 'printer' – the equipment or the person?), the existence of more than one word to represent a concept (e.g. dassie, rock-rabbit, Cape hyrax), and variant spellings (e.g. archaeology or archeology). Large subject-orientated databases containing articles from different periodicals can similarly be searched by title word and text word. Indexers

could play a more active role in the design of databases like these, by supplying advice on the most effective methods of providing access for various kinds of user (Booth 2001:198). Users have the same indexing needs, regardless of the entities' physical format. The main issue, however, is quick access from multiple entry points (such as in full-text searching).

Indexes to websites containing the texts of periodicals vary in form. In some, the user has first to select the periodical title from a list of items produced by the same publisher and is then shown a list of year dates to choose from. Selecting a date produces a list of individual issue dates or numbers, and subsequently the contents list of the chosen issue. In others a letter of the alphabet is selected, giving a list of titles starting with that letter; selecting a title gives year dates, issues and contents successively. Author and keyword searches are often possible, either by calling up a list or browsing through it or by entering the name or word in a 'dialogue box' and commanding a search.

Some periodicals, newsletters and bulletins appear in audio form only, or on film and video, designed for users with particular listening and viewing facilities. The index needs of users of these periodicals are the same as those of users consulting printed entities.

For an example of a local electronic periodical index, consult the *Mousaion* index on ISAP.

Indexing Levels

This section is based on Booth (2001:202-207).

In the area of subject indexing, periodical indexing and book indexing differ most. Book indexes aim to represent in fine detail the whole subject content, looking at individual pages, paragraphs and sentences, and including headings containing words that may not appear in the text but that are likely to be sought by users. Indexers of periodicals tend towards general representation or summarisation of each article, perhaps indexing the overall subject and a small number of topics, sometimes referred to, particularly in electronic media, as 'metadata' or 'metatopics'. Indexing of this kind, characterised by less detailed representation of the topics in an article, can resemble the subject indexing (classification) of entities for a library catalogue. Indexing requires great skill in denoting the whole significance of a complete article within a small set of index headings. Some periodical indexes provide no subject index, including instead merely an index of article titles alphabetically arranged according to the first (significant) word of each title. This is of little use to someone searching for subject information, as few users have in mind the exact title of an article when they are searching for information on a topic. Such indexes can complement a proper subject index, but cannot replace it. Another variation on this is to use the article title as a subheading to the chosen keyword or keywords, which is an improvement but still not very helpful.

In the indexing of any kind of entity, the best practice is to provide a heading that represents the topic most directly, so that users can depend on finding what they need at the point of access, not having to consider under which broader heading the topic may be included. For example, an entity on comets should be indexed as 'comets' rather than 'solar system' or 'solar system: comets'. An index consisting of a succession of 'classified' heading sequences, in which sought terms are used merely as subheadings under broader headings, with no possibility of direct access, is unhelpful and should be avoided.

Words from Titles

Sometimes indexing terms are taken directly from the title of the article, but only when the title represents the actual content of the article. Otherwise the quality of the index may be poor because the essence of the articles is not captured and valuable information may be omitted. In many instances titles are poor indicators of content, for example an article about fossilised specimens of dinosaurs, bearing the title 'Locked in rock' would not be helpfully indexed by using the keyword ('rock') from the title. Full titles are sometimes used as a basis for indexes, with the words manipulated automatically – see chapter 10 for information on KWIC and KWOC indexes.

Words from Abstracts

Abstracts usually appear at the head of articles. Before using words from an abstract, the indexer must make sure that the abstract is of good quality and suitable length, and competently written so as to accurately convey the essence of the article. When relying on abstracts as generators of indexing terms, indexers must be careful since the indexing is not done directly from the original articles (unless the indexer is also the abstractor). However, since the indexer is not relying on words from the title only, abstracts provide the opportunity to add as headings any relevant sought terms that are not featuring in the text. See chapter 13 on abstracting.

Words from the Text

Using words from the text provides better representation since the indexer must work through the whole article, reading it attentively, concentrating on sections such as the introduction where the aims and background of the article are described, the methodology used and the outcome (e.g. research results).

Words from Controlled Vocabularies

Using a standard vocabulary such as a thesaurus aimed at a specific subject field provides control, continuity and consistency over long periods of time. There is then no doubt about which term and which spelling of the term should be used. If the thesaurus is updated regularly by adding new topics and inserting suitable cross-references, it will continue to reflect the current and past scope of the periodical. Some thesauri include against each term the date when it was first used, or the bibliographic source from which it was taken. In other instances an obsolete or no longer preferred term carries the date when its recommended status was removed, as well as the term(s) by which it has been replaced. See chapter 14 on thesaurus construction.

Authority Files

If no pre-existing list is used for indexing, it is good practice to build up an authority file of terms used for indexing. This way indexers have a record of terms already in use and

can be sure that they are always used in the same form and that similar headings follow the same pattern. Any names (personal, corporate, geographical) that are included in the index should also be included in the authority file, supported by cross-references from all other forms.

Authors' Keywords

Authors of articles are often asked to provide keywords for their articles to be used in the index. Such keywords may be useful, but they should not be relied on as the sole basis for the index. Authors not conversant with indexing techniques may tend to 'word spot' or provide unsuitable word forms. In the end it is still the indexer's responsibility to ensure that each article is represented adequately with inclusion of relevant cross-references to link related terms, variant spellings and different words for the same topic.

Classification and Coding

Periodical articles can also be indexed by using a system of classification or coding in which topics are represented by alphabetical or numerical codes, or by a combination of both. Well-known classification systems such as the *Dewey decimal classification, Library of Congress classification* or *Universal decimal classification* may be used, or a customised system for the particular periodical, covering its specific subject field and vocabulary. When using such a system, the indexer must identify the indexable topics at the appropriate level in context of user needs, and convert them into the class notations or codes that most suitably represent them (as is done for entities in libraries). Because the notations and codes from the well-known systems are internationally understood, they can be used in indexing and abstracting services to bring together articles in different languages and from a variety of countries.

Guidelines for Indexing

Similar to book indexing, the periodical indexer must be familiar with the indexing process, how to construct entries (for names, scientific vs popular terms, etc.), and how to present entries (this differs from those for a book index and are discussed further on). The following are a few general guidelines:

- Read the title.
- Read the headings in the article.
- Read the abstract/summary that often appears at the head of the article.
- Take a look at the bibliography.
- Scan-read the article.
- Allocate indexing terms (either by using the natural language of the text or by using a controlled vocabulary).
- Take a look at the indexing terms once they have been allocated and decide whether they will guide the user to the article – do they represent the content of the article?
- Work according to the indexing policy of the periodical.

Keep in mind that the existing index (if there is one) to the periodical will indicate what is expected regarding choice of terms and presentation of entries.

It is not possible to discuss or illustrate all the types of periodical index entries an indexer may encounter. The following are some general examples only. It will be worth your while to take a look at a number of periodical indexes to familiarise yourself with the different characteristics.

Examples

The main characteristics of a periodical index are that they may contain certain abbreviations, the locators are made up of different elements, and have author/title information. Presentation differs from index to index, but must be consistent within any one index.

Typical abbreviations used are

A = advertisement

E = editorial

F = feature

R = review

Such abbreviations are not used in all periodical indexes. The abbreviations are considered helpful to the user. For example:

pyramids 61-62(E), 95-105(F), 134(R)

Valley of the Kings 3(A), 25-29(F)

Periodical indexes are presented in different ways, for example subject entries with appropriate titles of articles indented underneath (which may or may not contain the author(s), authors' names with the titles of their articles indented underneath, subject entries with more specific subject entries indented underneath, etc.). It also depends on whether the index is compiled for a single periodical only or for a database containing many periodical titles. In the last instance, the title of the periodical is also included in the index entry.

The following examples of periodical index entries may be useful for identifying the component elements. Note the use of capital letters, italics, bold, et cetera (these may differ from index to index). Each index will have its own rules. These are just examples and are not prescriptive.

Subject Index Entry with Titles of Articles

In this entry the names of authors appear in brackets, followed by the volume number and date of the periodical. Issue numbers and page numbers are excluded. The titles of the articles are not presented alphabetically, but chronologically. This is an entry for an index to a single periodical title:

ARCHAEOLOGY

Digging up the past (Evans, B.C.T.) 1/95

A new era dawns (Smith, B.) 1/95

Ancestors? (Clancy, C.K., Tlhako, E., Botha, W. and Makhanya, M.) 2/96

Egypt's most famous mummy (Hawass, Z.) 4/98

Author Index with Titles of Articles

Here the articles are listed alphabetically under the authors' names:

Botha, W. *see* Clancy, C.K., Tlhako, E., Botha, W. and Makhanya, M.
Clancy, C.K.
    Cradle of humankind 5/99
Clancy, C.K., Tlhako, E., Botha, W. and Makhanya, M.
    Ancestors? 2/96
Evans, B.C.T.
    Digging up the past 1/95
Hawass, Z.
    Egypt's most famous mummy 4/98
    Nefertiti 5/99
    Tomb robbers 4/98
Makhanya, M. *see* Clancy, C.K., Tlhako, E., Botha, W. and Makhanya, M.
Smith, B.
    A new era dawns 1/95
Tlhako, E. *see* Clancy, C.K., Tlhako, E., Botha, W. and Makhanya, M.

In the example of Clancy and three other authors, one may also use 'and others' or '*et al.*' in the cross-references: Botha, W. *see* Clancy, C.K. and others 2/96, or Botha, W. *see* Clancy, C.K. *et al.* 2/96.

Cumulative Index

The locators in a cumulative index to a single periodical (covering, say, five years) must include sufficient date and volume information to guide the user to the correct issue. The following index entries are alphabetically arranged according to title and include year, volume, issue and page numbers. Authors are excluded:

Ancestors? 1996 2(3):34-40
Clay tablets and their hidden messages 1997 3(1):11-18
Cradle of humankind 1999 5(3):52-62
Digging up the past 1995 1(1):50-58
Egypt's most famous mummy 1998 4(4):22-31
From clay to papyrus 1996 2(3):15-18
Museums as social institutions 1997 7(1):43-47
Nefertiti 1999 5(4):71-75
A new era dawns 1995 1(6):10-18
Tomb robbers 1998 4(2):42-45

Index Covering Several Periodical Titles

In an index covering several periodical titles, full details of the containing periodical will be indicated, including title, volume number, issue number, page numbers. Authors are excluded and may appear in a separate author index. For example:

Ancestors? 1996 *Archaeological Studies* 2(3):34-40

Dinosaur finds. 2001 *Journal of Palaeontology* 4(4):29-35

Evolution vs creation. 2000 *Bulletin of the SA Archaeological Society* 3(1):69

All the examples in this section illustrate that periodical indexes differ to quite an extent. It cannot be stressed enough that the indexer must work according to the rules (policy) for a particular index.

Teamwork

As with most cumulative indexes, indexing is done on a cooperative basis – employing more than one indexer (mostly freelance indexers). This implies that each member of the team must understand and observe the rules of the indexing policy. There must also be the opportunity to exchange information and discuss problems. See the section under 'Contracts' regarding ISAP.

BIBLIOGRAPHY

Ashcroft, L & Langdon, C. 1999. The case for electronic journals. *Library Association Record,* 101:706-707.

Beare, G. 1999. *Indexing newspapers, magazines and other periodicals.* Sheffield: Society of Indexers. (Occasional papers on indexing; no. 4).

Booth, PF. 2001. *Indexing: the manual of good practice.* München: Saur.

Lancaster, FW. 2003. *Indexing and abstracting in theory and practice.* 3rd ed. Champaign: University of Illinois, Graduate school of Library and Information Science.

Storage & networking: technical. 2003. *Computing SA,* 23(6), 17 February.

Taylor, AG. 2004. *The organization of information.* 2nd ed. Westport, Conn: Libraries Unlimited.

Wellisch, H. 1995. *Indexing from A to Z.* 2nd ed. New York: Wilson:347-354.

10 AUTOMATIC INDEXING

Peter Underwood

Abstract

Using a computer to process a text that is already in machine-readable form so that index terms are allocated to its content without direct human intervention can be described as 'automatic indexing'. There is, however, a significant gap between the systems currently available and this objective. The author explores this in some detail but concludes that it is a development of the future and that the human element is still a vital component in indexing.

'Drowning in information but thirsting for knowledge' represents the paradox of our times. The development of Information and Communication Technologies (ICTs) has changed the way in which many aspects of the 'information game' are played but has not affected its fundamental principles. The problem those principles seek to solve was articulated by Plato in the *Meno* (Stanza 80e): '[A] man may neither seek what he knows or what he does not know, because he does not have to seek what he knows, for knowing it, he need not enquire, nor may he seek what he does not know for then, not knowing it, he knows not what he is seeking.'[1] Despite significant progress in the last 20 years in developing computer systems that can rapidly retrieve text as sentences, words and word stems, the essence remains a gamble between author and potential reader. If both use the same word to describe the same concept, then the retrieved document *may* satisfy the reader.

Note the emphasis on 'may': the problem is more subtle than simple matching of single words or, even, sets of words. The capacity of a language to encapsulate fine distinctions means that even texts with a high degree of overlap between words used by the author and those used by the searcher can still fail to capture what it is the searcher is looking for. How we impute and recognise meaning is the subject of investigation by psycholinguists and it is clear that the development of strong theories and robust techniques is still some years away. Unless there is a significant breakthrough in understanding and modelling the complexity of textual communication and the association of meanings, it will not be for this generation to stand, as they might on the deck of the Starship 'Enterprise', and command the computer to perform a complex analysis resulting in retrieval of only a few highly relevant documents.

1 I am grateful to Jennie Underwood for this translation and for first drawing my attention to the significance of Plato's comment.

Yet this is a reasonable aim. Indexing by human labour is expensive and labour-intensive. It is also notoriously difficult for humans to agree on 'aboutness' and what words should be used to describe it (Blair 1990). Even more difficult is to ensure that the words used by searchers correspond to those used by indexers (Lancaster 1991:10-14).

Mechanising index production would clearly have benefits in speeding up production, replacing human with machine time – and thus reducing costs of production – and ensuring greater consistency. However, the present limitations of our understanding of psycholinguistics still means that the rules by which computers produce such indexes are unlikely to result in indexes where there is a perfect match between what is represented and what is sought. Lancaster has suggested that it may be more fruitful to concentrate on assisting searching rather than seeking greater consistency of indexing (Lancaster 1991:74-75).

Using a computer to process a text that is already in machine-readable form so that index terms are allocated to its content without direct human intervention can be described as 'automatic indexing'. There is, however, a significant gap between the systems currently available and this objective. 'Allocation to content' requires an understanding of a text, imagination and a mastery of language in order that the meaning of the content is expressed. A commonly used approach has been to use the computer to create concordances, or lists of words, which the searcher then explores.

'Word association' prompts the searcher to link target words together, using logical (sometimes called 'Boolean'[2]) operators, into a search statement. The role of the computer is firstly to scan texts and create an 'inverted file' which associates each word in the file with positions in the texts. Its second use is to match each word in the search statement against the inverted file and to identify texts that have words in common. These can then be retrieved and scanned for relevance. By suppressing very common words with low subject significance, such as 'and', the inverted file is kept within manageable proportions. Because the searchable word list reflects only words in the texts, the technique may be described as 'derived' or 'extractive' indexing.

Computers are excellent tools for creating and managing such lists, so that the addition of a document to the database results in the automatic inclusion in the inverted file of the words it contains. Computers are also highly suitable for searching the inverted file, matching the pattern of a sought word against the patterns contained in the file.

Automatic indexing is the creation of index terms from a text or image by computer, without intervention by a human indexer. With the widespread use of computers to store text, a potential for confusion has arisen regarding the indexing function. 'Text searching' is the scanning of a text to match terms decided by the searcher; it is sometimes called 'natural language searching' to contrast it with 'searching by index' where the search for terms is conducted in a file separate from the text and the text and index are linked by a locator such as a page number. The text can be a complete work (often called 'full text'), or some representation of it, such as an abstract or title. Many electronic sources of information,

2 Named after George Boole (1815-1864), Professor of Mathematics at Queen's College, Cork. The diagrams sometimes used to depict logical combinations, 'AND', 'OR' and 'NOT' were development of the English logician, John Venn (1834-1923).

held in databases or accessible over the Internet or World Wide Web, are available for full text search. The low cost of computer storage and the speed and power of the processors have meant that this approach is both feasible and economical to provide.

Maintaining the distinction between indexing and text searching is crucial if one wishes to consider effectiveness, however. The assumption that text searching can invariably offer 'better' results than searching based upon indexes arises because it is the whole representation of the text that is available for searching rather than terms considered, through some process of selection, to be significant indicators of content. The argument is based on the presumption that searchers can best describe and recognise what will be significant. It also begs the question 'What is "better"?' Once a text has been transferred to a form suitable for computer storage, a text searching approach to location of relevant items shifts most of the costs of searching and shifts *all* of the effort required to the searcher. This is certainly better for the provider of a source of information and may even give the illusion to the searcher that this is a better product – but it is not necessarily borne out by experience.

Since the 1950s, there has been a considerable volume of research on the effectiveness of methods of indexing and comparison with other approaches such as text searching. In addition, practical experience in using databases and searching the World Wide Web and other Internet channels has yielded contradictory, and sometimes puzzling, results. It cannot be safely concluded, for example, that either approach is, of itself, superior even for the fundamental task of all information retrieval, that of indicating to the searcher highly relevant documents in response to a search.

The fundamental task is, however, only one aspect of how 'effectiveness' might be measured and assessed. A full assessment must take into account the costs of input, costs of storage and production and costs of searching, and compare these with the effectiveness of the search. Assessing effectiveness is also fraught with difficulties: the aim of searching is to retrieve *relevant* documents and a minimum of irrelevant documents; but, surprisingly, even experts in a subject area are not consistent in their judgments of what constitutes a document that is an acceptable answer to a search (Cleveland & Cleveland 1990:147). 'Relevance' is also a complex concept. At one level, one may judge the relevance of a document in terms of how well the search terms used to retrieve it match the contents of the document; on another, there is a judgment on how useful the document may be to the searcher. This aspect of 'pertinence' is obviously much more subjective and temporal than relevance per se, and takes into account the effort that may have to be expended on locating the document, as well as its intrinsic worth in solving the searcher's problem. The speed of a computer search may provide the illusion that the results are also going to be speedily assessed. Practical experience of searching on the World Wide Web should disabuse one of this assumption, however: a search retrieving several thousand apparent 'hits' is hardly effective if it takes several hours to assess the relevance of the items found, even assuming that the searcher is persistent enough to evaluate the whole set.

Text searching approaches to information retrieval depend on documents being available and accessible on a computer system. The provision of information in the form of databases is the most prominent example and it is common to find full text searching facilities provided as a means of information retrieval. It is also quite common to find that

the records in such databases are enriched with subject descriptors assigned by an indexer, or added automatically, from an index vocabulary. Research on index languages has explored the utility of searching using assigned index term systems, compared with natural language text searching. The results of early studies suggested that the assignment of index terms did not convey any advantage over natural language approaches and this finding was confirmed by further experiments. Later work has resulted in some modification of this view, especially when large databases and full-text systems have been considered: the problem of overwhelming the user with retrieved documents is recognised as being significant. Because of the richness of natural language, it is highly probable that searches of full-text systems will retrieve many documents but will also retrieve many that are irrelevant, or only partly relevant.

Despite this consideration of effectiveness, the cost of manually assigned indexing may eventually serve as a deterrent to their production, especially when sources such as electronic books and journals become more common. 'Bundling' and electronic text with software capable of natural language searching of the full text is likely to be a cheap option, attractive to the publisher because it is fast to produce and update and superficially attractive to the purchaser because it appears to offer unlimited possibilities of searching. The 'value-added' component that can be supplied by a skilled indexer may be relegated or diminished in importance *unless* publishers and purchasers can be persuaded of the virtue and power of assigned indexing, even as an adjunct to natural language searching. The desirable product is a text that can be searched using both means: index terms assigned by an indexer after inspection of the document combined with natural language searching ability, when required.

The starting point for any process of automatic indexing is input of the text to the computer. The text is allocated a unique document identifier by which a copy of it can be subsequently located in computer storage. The characters of the text are then analysed to create a list of lexemes, or word-like units such as chemical formulae, that it contains. Each lexeme or unit is linked back to its parent document using the document identifier, this creating an 'inverted file'. Lexemes derived, or extracted, from the text provide the raw material on which automatic indexes are produced.

The earliest applications of this approach were the indexes compiled by Hans Peter Luhn (1896-1964) and Herbert Marvin Ohlman in November 1958, when the International Conference on Scientific Information (ICSI) took place in Washington, DC. Luhn produced indexes for the conference programme using his 'Keyword in Context', or 'KWIC' technique. Ohlman demonstrated a similar approach, which he called 'Permuterm' indexing, and used it to produce an index to the proceedings of the conference. The impact of the 'information explosion' following from the upsurge in economic, technological and scientific activity after World War II was the driving force behind efforts at improving documentary control. Speedy production of indexes and searching aids was considered essential if the growing volume of scientific material was to be made quickly accessible. In addition, by applying simple rules the lack of consistency implicit in human indexing would be avoided. Luhn (1957), Ohlman (1957) and Phyllis B. Baxendale (1953), who developed techniques that identified topic sentences in documents, were pioneers in the field of applying computers to the task of mechanising the production of indexes, though many others contributed to

the search for techniques that could contain the 'information explosion' and aid searchers to find relevant documents. The KWIC index based on the titles of documents is the most familiar product of automatic indexing from this period.

At the core of these approaches is the concept of deriving keywords from some part of the text of a document by machine manipulation. Statistical analysis identifies very commonly occurring words, such as prepositions, coordinators and articles that are low in subject content, but needed for syntax; these are allocated to a 'stop list' that excludes them from indexing. Such approaches may be described as 'concordance-like' and, although they can be rapidly produced, they do not provide means of searching other than through the words used by the authors of the documents processed for the index. In subject domains where the vocabulary is more limited and standardised than might be encountered in, for example social sciences, the degree of consonance between searcher and author may be high, thus obviating the disadvantages of the purely derived approach.

The derived index approach may be summarised as a quick and economic approach to index generation with no human intervention, where the costs and effort have been shifted from the creation step to the searching step of the index production and use. It is the searcher who has to display some flexibility of mind to think of the range of terms that authors may have used to describe a sought phenomenon. Despite this, and taking account of the costs of human intervention in indexing, Lancaster has suggested that it may be more fruitful to concentrate on assisting searching than seek greater consistency of indexing (Lancaster 1991:74-75). A familiar aspect of this problem is that of seeking relevant documents on the World Wide Web using a search engine.

Several strategies have been developed as a means of assisting searchers. Truncation (also known as 'stemming') allows for searching on a portion of a word; 'leftmost' truncation is commonly used, suppressing the word ending so that singulars and plurals of words that follow the standard format can both be matched. 'Rightmost' truncation allows for the suppression of prefixes, while internal truncation, or 'wildcards', allows for suppression or substitution of letters within words. These latter forms of truncation are often useful for specialised forms of searching, such as for molecular formulae. In each case, the result of truncation is to increase the number of matching documents during a search; however, this will almost certainly also increase the number of irrelevant documents also retrieved. The more extensive the truncation, the more documents will have to be scanned by the searcher and the greater the likelihood that many will have to be rejected as irrelevant. Weak truncation tends to be the most effective because it balances the ability to suppress usually insignificant differences such as singular and plural word forms, with retention of enough of the word stem to provide discrimination.

The list of words in the document can be further processed to analyse their distribution. By noting the frequency with which each word appears, some judgment can be exercised on whether it is likely to be significant as a descriptor of document content. Words or terms appearing infrequently are unlikely to be good descriptors of content, whereas those appearing with high frequency may not be good discriminators. Candidate words or terms can then be chosen from within a suitable frequency range. Such a list is, at least, an aid and starting point for an indexer but is still likely to include many words that are syntactic rather than descriptive. Automatic indexing systems need to be able to

recognise word forms and functions to highlight those that are most likely to connote subjects. Reliance on word frequency is also not effective once many texts have been added to the database: what becomes important is the discriminating power of a word or term within the database – that is, across the corpus of documents that it contains – rather than within a document. The discriminatory power of a word or term is also sensitive to context: what may be a common word or word form in one type of document will be uncommon in another. The discriminatory power of a word may be improved by assigning it a weight, such as being a major or minor descriptor – but this again raises the question how to recognise the appropriate thresholds to make such a choice. Simple reliance on computer analysis of word frequency is thus unlikely to produce indexes that are effective for searching.

Syntactic analysis has been explored as a step beyond word frequency approaches. It is an attempt to recognise the form and function of a word, or group of words, in a phrase. By 'parsing' a sentence according to rules that define parts of speech, the computer processing yields a series of candidate phrases that are deemed to describe it. This is presently an expensive exercise both to develop and to implement, and there is a lack of evidence that the information retrieval is sufficiently improved to justify the attendant costs. It remains an important area for further research.

Automatic indexing applied on a large scale must be regarded as a development of the future, allied with greater computer power and a much deeper understanding of how human beings recognise, weigh and assign meaning to ideas and documents. For the present, computers can provide a useful function by rapid production of simple extractive indexes and natural language searching. Human indexers still have an important role to play in providing indexes to conventional publications and, arguably, to enriching the retrieval potential from databases and other electronic sources of information.

BIBLIOGRAPHY

Baxendale, PB. 1953. Machine-made index for technical literature: an experiment. *IBM Journal of Research and Development*, 2(9):354-361.

Blair, DC. 1990. *Language and representation in information retrieval*. New York: Elsevier.

Cleveland, DB & Cleveland, AD. 1990. *Introduction to indexing and abstracting*. 2nd ed. Englewood, Colo: Libraries Unlimited.

Lancaster, FW. 1991. *Indexing and abstracting in theory and practice*. London: Library Association.

Luhn, HP. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309-317.

Ohlman, H. 1957. *Permutation indexing: multiple-entry listing by electronic accounting machines*. Lexicon, Mass: Rand Corporation, System Development Division.

11 WEB INDEXING

Madely du Preez

Abstract

The meaning of web indexing (or online indexing) is explained, with advice on how to index a website. Some web indexing decisions and questions are explored, focusing on indexing a single website for navigational and retrieval purposes. A brief explanation of HTML (HyperText Markup Language) is given to introduce meta-tags as a means of controlling how search engines index a website. This is followed by a brief discussion on the description and keyword meta-tags and their creation.

Introduction

Indexing, as in 'indexing the Internet', is an approach to information discovery in a way that lends itself to effective navigation across distributed repositories (Denenberg 1996:1). *Information discovery* means locating objects of interest when the population of objects from which to choose is potentially widely distributed. Objects may be aggregated into collections, and organised thematically, but may also be distributed across the Internet – even though the collection appears logically cohesive because the metadata structure associated with the collection is supporting coherent navigation.

The term 'web indexing', or 'online indexing', covers a wide range of processes and products, from putting a book index online to attempting to create an organised index of the entire Web. The most common web indexes could include static indexes, online publications (e.g. online newsletters), combination print-online indexes, website indexes, site maps, web directories, selective web indexes, indexes of web indexes, keyword indexes and controlled vocabularies/thesauri. Usually, though, web indexing simply involves the creation of hyperlinks within documents that are accessible on the Web (Maislin 2003:3).

Denenberg (1996:1) distinguishes between two broad web-indexing models: a *global* index versus independent *local* indexes. Global indexes are created by search engines like Yahoo and Google or indexes for an intranet where each department, section or branch of the organisation to which the intranet belongs is responsible for its own website. Since the current model of the Web is a single, logical agglomeration of documents, with logical flat indexes of its entire content, global indexes are also physically distributed. A local index is an index created for a specific website, such as the departmental or branch website mentioned above. This index will not be distributed across the Internet in the same way as global indexes.

Website indexing (Rowland 2000b:1) involves two basic skills: indexing and website design as well as the willingness and resourcefulness to invent new ways to present information. Browne and Jermey (2001:47) add an understanding of the structure of information on the Web to these skills. The best form of a website index is highly dependent on the type of website being indexed and the needs of the audience for whom the index is being created (Rowland 2000a:1). As a website index aims to improve access to information on the website, the success of the index can be measured by how quickly users can get in and out of the index page and find the desired information (Maislin 2000:40).

The purpose of this chapter is to explain what web indexing (or online indexing) is and advise on how to index a website. Some web indexing decisions and questions will be examined. The focus will be on indexing a single website for navigational and retrieval purposes. A brief explanation of HTML (HyperText Markup Language) will be given to introduce meta-tags as a way of controlling how search engines index a website, followed by a brief discussion on description and keyword meta-tags. This will include an example on the creation of meta-tags.

Web Indexes

Indexing the Web involves three different kinds of indexing: a book style of hard-coded index links within a website to serve as a navigational tool, subject trees of reviewed sites, and indexes created by search engines (Denenberg 1996:1). The most important characteristic that indexes on websites and the Internet share with print indexes is the goal of any good index: 'to bring together all the items on a similar subject which are separated in the book [or website] itself' (Broccoli & Van Ravenswaay 1999:37). The scope, structure and design of a web index is however different from traditional indexes. Cleveland and Cleveland (2001:220) explain that a book index usually points to information in one file (the text of the book) and periodical indexes point to information in several thousands of files (numerous journals). Web indexes however point to many millions of files, at many levels, existing in every niche and corner of the world.

Web indexing could, however, also refer to the use of meta-tags where keywords are assigned to a web document to assist search engines in correctly indexing the document in their indexes.

The term 'index' has developed a broader, more general meaning on the Web and could essentially be a collection of Uniform Resource Locators (URLs) designed to take users to other websites that have either been collected manually or by search engines like Yahoo. Such indexes usually reflect little analysis of the websites they list and these lists are more like catalogues than traditional indexes.

All indexes have page numbers, unlike the World Wide Web. Embedded indexing techniques, however, do not require the indexer's awareness of page numbers (Maislin 2000:43), which makes it an ideal technique to index web documents. The indexer merely creates hyperlinks to the nearest section titles (or the specific point in the document where the term occurs) instead of page numbers to aid navigation among documents (or within a specific document) that are accessible on the Web (Maislin 2000:45).

Indexing Pitfalls

Website indexes can enhance search success, but are costly to implement and, depending on the experience of the indexer and the time available for indexing, are of variable quality. There will inevitably be a time delay between the creation of the web page and the addition of links to it in a search engine's index. These delays can be reduced if the page is reviewed for indexing before being loaded to the Web (Browne & Jermey 2001:28-29).

It is important to bear in mind that online or Internet users might find it fun but also diverting to surf the Web. Those users who make use of indexes are generally looking for specific information, and it is the indexer's task to help them find it. Lathrop (2003:3-4) cautions that readers are more easily frustrated when searching an online index because they do not have the advantage of spatial orientation, which they are familiar with in printed books. The sheer size of the online document may cause them to think scanning the document for the information they want would simply be an exercise in futility. A well-written, comprehensive index increases customer satisfaction and reduces costly product support time because it makes a product easier to use.

Web Indexing Decisions and Questions

A number of indexing decisions and questions need to be made and asked at the beginning of a web indexing project.

What Are the Boundaries?

Will information be added to this website, like a corporation's intranet where new memos, presentations and budget reports are added to the site which would constantly change the indexing? (Wright 1997:3). If so, it could be advisable to create a thesaurus or a terminology-orientated ontology to assist in building a structured index of the website (Desmontils & Jacquin 2003:1). In this context, an ontology is a 'set of concepts each one represented by a term (a label) and a set of synonyms of this term, and a set of relationships connecting these concepts by the specific/generic relationship' (Desmontils & Jacquin 2003:[3]).

Web Authoring Tools

These are software packages that build raw information into usable online resources. These tools operate in several ways and one would have to determine the orientation of the project (Wright 1997:4).

How Will the Index be Displayed?

It is important to know how the index will appear to the user – how many levels of index will appear? When the user clicks on an entry will a list of topics relative to the applied term appear, or is there only a one-to-one correspondence between index entry and topic? Are there any controls over the sorting of the entries? How will cross-references be put in, and what will happen if the user clicks on them? How will the user get to different sections of the index: by clicking a button or typing a letter? (Wright 1997:6-7).

How Does the Indexing Get Into the Files?

Closed systems force the indexer to work within a closed software package, where only that tool can be used to input entries into the system. This would usually require the indexer to operate at the client's worksite so that their software and files can be accessed. Open authoring tools will however allow the indexer to import and export indexing into and out of its file structures, usually with spreadsheet-like data files (Wright 1997:7).

What Kinds of Files are Included in the Project?

Online indexes can lead a user to a topic, to photographs or sounds, a menu of choices the user needs to make, to a website, or the index connection can start a macro running, opening files or formatting text. Wright (1997:8) advises one to ask about the structure of the project and its menu screens if they are used. She explains that menu screens are like table-of-contents screens that present a user with choices, which then lead to specified topics. Another type of topic structure one needs to ascertain concerns decision trees and whether any further help screens are going to be offered.

Who is Inserting the Indexing and Maintaining It?

Who will be maintaining the indexing efforts, and how will that process work? Expanding indexes usually require a master indexing list of used terms to be maintained, or a complete thesaurus (Wright 1997:9).

Time Frames

- Various factors, such as authoring tool problems, debugging glitches and editing difficulties, could complicate the completion time for a web indexing project – working with a known tool that has been used before will reduce the work schedule while unknown and new tools need more time for testing and compiling (Wright 1997:9).

Equipment Needs

It is advisable to determine the necessary kinds of equipment before beginning the project, especially when working off-site. It is very important to ensure that the computer involved can handle the tools to avoid memory or incompatibility problems. Basic equipment needs include (Wright 1997:10)

- a computer with enough memory and storage
- modem or network card
- authorisation and passwords to access files
- ability to FTP or have remote access if off-site
- authoring tool(s) installed correctly with all subsidiary software and utilities
- capability to run authoring tools without conflict

Availability of Files

When working in-house it is important that one has access to the authoring tools and the files to ensure sufficient time to complete the project. When working in stand-alone indexing software, and then embedding or inserting the indexing into the authoring tool, it is important to get print-outs or copies of the files to work from. The indexer should also consider how files will be transported (Wright 1997:11).

Trial-run Indexing

Consider doing a few sample entries, and then do a trial run of the compiled index to ensure correct and smooth running. Avoid entries that could cause the computer to hiccup: sorting problems, special characters, odd punctuation, cross-references, special jumps to animations or odd topics, things that might not work as planned (Wright 1997:11).

Fees for Online Projects

Wright (1997:11) advises charging an hourly rate for web indexing projects unless one is working for a group that has established its procedures and used the authoring tools several times before.

Indexing as a Navigational Tool

One needs to consider two indexing processes when indexing a website or a web document. The first is indexing as a navigation tool where the purpose of the index is to transport the user between various points of interest within a website or number of websites. The second process is to assign keywords or indexing terms using meta-tags to control how search engines index the website for retrieval purposes.

An entry in a book index contains much information. Consider the following single entry:

Multiple indexes, 49, **66-67**, 137

Users know that the bulk of the information on multiple indexes start on page 66 and can assume that page 49 provides an introductory comment, while page 137 casually mentions multiple indexes since it comes completely isolated from the other page numbers.

An online or web entry would look something like this:

Multiple indexes, *,*,*

The user no longer knows which entry is the most important and can no longer understand how these documents fit into the entire collection of documents. Neither is it possible to know the lengths of the documents themselves or to determine the order of the entries. Maislin (2000:44) explains that this loss of page numbers and of global context presents the indexer with a fundamental handicap to writing a good index.

Most of the other dilemmas facing indexers are related to layout (Maislin 2000:44). In the main authoring software designers pay little attention to the need for indexes or the indexing process, since these are often merely added as fancy enhancements. Some do not allow for sub-sub-entries, cross-references or special types of references, nor do they allow for locator-specific formatting. HTML (HyperText Markup Language) also does not allow for indentation.

The following sections will examine some ways of accomplishing these goals and overcoming some of the online environment challenges.

Rating Index Entries for Importance or Relevance

There is no intrinsic way to present ranges of information on a website. Large quantities of information are also split up due to *document chunking* (the breaking up of documents into small units). As with books, indexers can write locator text that explicitly communicates context by making use of a section title or a heading to provide the locator. The locator text is however the web indexer's only opportunity to provide context. It is advisable rather to use descriptive headings so that the user can confidently guess what these sections contain. For example (hyperlinks used instead of page numbers and represented by underscoring):

Maintenance, in web indexing
    Markup language
    Mediation of discussing list disputes
    <META> tags
    Multiple indexes

Maislin (2000:45) is of the opinion that the best way to improve usability is to avoid ambiguity. Even one-word locators like 'Marketing' and 'Networks' are unambiguous to most users.

Visual cues, such as colour and enhanced font styles, can be added to online or web indexes to make the index more meaningful. More important entries can be presented with larger lettering while links to definitions can be coloured red. However, these should be used sparingly and only for clearly defined purposes. Note that an index with a cornucopia of line sizes and colours looks unprofessional.

Book indexes are arranged alphabetically. Maislin (2000:46) suggests using a sort order that makes more sense to the user in an online or web index. Possible sorting schemes include *chronology*, *importance* and *page order* (for printed books converted to online documents).

The online or web indexer can also rate entries by displaying those that meet a pre-defined threshold of importance in bold – a binary rating system. But this system is not as effective when applied together with subjective thresholds like 'importance' or 'relevance', because the user does not necessarily understand that threshold. The only advantage of a binary rating system is that it is easy to implement.

A discreet rating system with three to five levels is better, the most intuitive being the 'asterisk rating system' where the additional * represents slightly greater importance or relevance.

Communicating Relationships between Entries

Cross-references are more complicated online or on the Web, but are essential in a stand-alone index since they communicate the relationships between index entries. *See* references direct users from terms that are not used to terms that are, for example 'Unisa, *see* University of South Africa'. *See also* references are used with terms that do have links, to suggest additional search possibilities. *See also* references are usually reciprocal, and one would expect to find the reference '*Haemophilus influenzae, see also* Meningitis' as well as 'Meningitis *see also Haemophilus influenzae*'. Note references that contain a scientific name have to be in italics next to an italic *see* or *see also*. When both terms in a reference are in italics, the *see* or *see also* is written in roman font (Browne & Jermey 2001:41).

Maislin (2000:49) gives the following tips on creating cross-references:

- *See* cross-references should be used carefully because of the difficulty users have with navigating long online indexes. Rather post index information at both locations if two main-level terms are effectively identical.
- *See* references can become a technique to manage users' word choices and educate them towards the preferred vocabulary set. This can be very important if the index is enhanced by a search engine, especially as search engines do not handle synonyms or related terminology.
- *See also* cross-references should be presented as early as possible, like at the top of a sub-entry list as opposed to the bottom. This should help users understand the document structure.
- *See also* references could also be used to point to other locations within the index, appropriate reference lists within the document, transitory web pages and websites outside the documentation.
- Use circular *see also* references in a hyperlinked text so that users can retrace their steps.
- 'Related Topics' paragraphs in the documentation work similarly to *see also* cross-references.

An interesting alternative to multiple cross-references is the creation of multiple indexes. This approach can make a website more comprehensive without overloading it with text. For example:

Engineering Institutes in South Africa, by discipline:
Acoustical | Aeronautics | Chemical
Civil | Electrical | Electronic
Industrial | Mechanical | Mining | Structural
Clicking on Electrical will display the second index with the names of the different Electrical Engineering Institutes in South Africa.

The creation of multiple hyperlinks for a single locator remains the most obvious technique to demonstrate relationships between index entries: 'Multiple indexes, 49, 66-67, 137' could now be presented using a multi-lined locator approach or simple numbers when images are presented: 'Pictures of Table Mountain, 1, 2, 3'.

The following is an example of an HTML index. Note that this layout is reproducible using standard HTML definition lists. Entries are sorted either alphabetically or by importance, depending on the context. Qualifiers could be added while most sub-entries get their own line. The lists of cross-references are combined horizontally, with semicolons between them:

Professional organisations in South Africa
  Association for Consulting Engineers
  Council for Professional Engineers
    See also Engineering Institutes in South Africa, by discipline: Acoustical; Aeronautics; Chemical; Civil; Electrical; Electronic; Industrial; Mechanical; Mining; Structural
  Information Science interest groups
    See also Association of Southern African Indexers and Bibliographers; ASAIB; Interest Group for Bibliographic Standards; IGBIS; LIASA interest groups; Research and Training Interest Group; RETIG; South African Online Users' Group; SAOUG
  South African Institute of Architects
  South African Institute for Library and Information Science (SAILIS): *See* Library & Information Science Association of South Africa; LIASA
  South African Teachers' Association

When choosing a design for an index, one should always ensure that enough space is allowed for qualified links, enough vertical space for entries and sub-entries, a visible white space between topics and subtopics, clear and unambiguous lists of cross-references, and no conceptually overlapping sub-entries (Maislin 2000:51).

Developing a Visual Architecture and Navigation System

Rowland (2000b:1) finds it necessary for one to know basic HTML commands before starting to index or at least have an understanding of what HTML is and how it works, as well as know how to construct an index in plain HTML. This will help to format the index page more precisely and it will also help identify and solve HTML problems. (HTML will be discussed later.) Maislin (2000:51) reckons it is not uncommon for an author or indexer to spend more time designing and editing the format than actually writing the index or adding context to the locators.

One way to solve the problem of having a single web page with thousands of lines, where a user has to scroll down to the second last entry in the index, is to divide the index into various pages. The most common solution is to break such a long index into pages by letter. A navigation bar of letters, with each letter linked to one of the index pages is placed at the top of each page:

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

Letters could also be grouped together or divided further. It is however very important to be consistent. The navigation bar could also be repeated at the bottom of each index page, to enable the reader who wants to jump ahead to do so without having to scroll back to the top of the page. The navigation bar (along with the Next and Previous buttons) could also be placed in a frame that does not scroll, at the top, bottom or side of the page (Maislin 2000:52).

Indexing for Internet Retrieval

The World Wide Web lacks a systematic structure by which an information seeker can find desired information in the minimum time. Broccoli (2003:1-2) explains that many search engines do free-text searches. This means that the engine is looking for every occurrence of a word within the text of a page and a document could thus be retrieved even if the word used in the search appears once in the whole document. The solution lies in assigning keywords and setting search engines to search only keywords, developing a custom-designed thesaurus based on the terminology used on the website as well as within the specific industry or profession, and setting the search engine to recognise synonymous terms.

The assigned keywords are then entered into the header of the website using meta-tags. Meta-tags are thus a means to control the process of web indexing (Alimohammadi 2003:238).

What are Meta-tags?

Meta-tags are non-displaying, or hidden, HTML tags that may provide site owners and authors with a degree of control over how a web page is indexed. The term 'meta-tag' actually arises from the term 'metadata' which means data about data that is structured to describe an information object or resource. As such it may describe the content of a web page (Alimohammadi 2003:238). Meta-tags have become a design element used by webmasters and indexers to help search engines to better acknowledge information-poor pages (Sullivan 1998:3). Website designers or indexers use them in the indexing process of a website by choosing keywords describing the content of the site. They also provide a mechanism for identifying information that should be included in the response headers for an HTTP (Hypertext Transfer Protocol) request and can be extracted by servers and clients for use in identifying, indexing and cataloguing documents.

The Concept of HTML

HTML, the computer language used to create web documents, is written using tags, which are commands written inside angle brackets: <and>. Rowland (2000b:2) explains that all tags are written in pairs, with opening and closing tags surrounding the affected text. The opening and closing tags are identical except that the closing tag contains a forward slash symbol (/). In <B>Association of Southern African Indexers and Bibliographers</B>, <B> is the opening tag and </B> the closing tag. When the document is viewed in a browser,

one would not see the tags, but will see the formatting: 'Association of Southern African Indexers and Bibliographers'.

It does not matter whether the 'B' is typed in the upper or lower case. HTML editors might automatically change all tags to upper case. It is therefore best to be flexible.

All web pages have two sections: a HEAD section and a BODY section. The information about the web page goes into the HEAD section, including the title and any information one would use to describe the web page with, such as the information included in meta-tags, for example the keywords.

The title for the ASAIB website will appear as follows in the HEAD section:

```
<HEAD>
<TITLE>Association of Southern African Indexers and Bibliographers</TITLE>
</HEAD>
```

Remember that this title appears on the top of the browser, and not on the page itself; the title that will be displayed on the web page appears in the BODY section of the HTML document.

Each page in a website should have its own set of keywords and descriptions. It is however desirable to use the name of the website as the title for all the pages, along with subtitles for specific pages. This will make the web page more identifiable to both users and search engines (Rowland 2000b:3).

Different kinds of meta-tags are available to identify properties of a document and assign values to those properties. The title, keyword and description meta-tags are the most important ones that are used by search engines for indexing purposes. Title and author metadata can often be generated automatically from web pages. Subject keywords (words describing the subjects of web pages) and descriptions (abstracts or summaries of the content for display by search engines) usually have to be generated by people (Browne & Jermey 2001:8). Alimohammadi (2003:239) explains that meta-tags should always be placed at the head of the HTML document, just after the <TITLE> element, between the actual <HEAD> tags and before the <BODY> tags. The following example illustrates what the meta-tags for the ASAIB website could look like:

```
<HEAD>
<TITLE>ASAIB</TITLE>
<META NAME = "author" CONTENT ="Marlene Burger">
<META NAME ="description" CONTENT ="Official web site for Association of Southern
African Indexers and Bibliographers (ASAIB). It includes the personal details for
the executive committee members as well as the Association's training courses,
conferences and publications. It also includes a directory of freelance indexers who are
members of ASAIB">
<META NAME = "keywords" CONTENT = "book indexing, periodical indexing, indexing
workshops, freelance indexers directory, indexing conferences">
<HEAD>
```

Note that keywords can also be included in the <TITLE> tags. One could for example include a business slogan or a few keywords with the name of the company in the <TITLE> tag.

Description Meta-tag

The description meta-tag will be displayed when a user enters a query and the search engine retrieves some of the relevant websites. Alimohammadi (2003:240) advises indexers to use a description meta-tag if there is no descriptive paragraph at the start of the document, but to omit it if there is one. It is particularly useful for documents with a small amount of text, such as one consisting mainly of photographs.

Search engines usually take the first two lines of text, or the first number of characters (the exact number of characters is determined by the search engine) found on a website, and use them as the description. Those lines might not necessarily clearly describe what the website is about. The description included in the description meta-tag should be short, concise and to the point, but should not be so compressed that it cannot appropriately reflect the contents.

Keywords Meta-tag

The keywords meta-tag helps search engines to categorise websites and allow people to find web pages more quickly. Without this meta-tag, search engines will choose words for the site from the title and text of the site and these might not be the words or terms that best describe it. It is important to use only those keywords that the user might type in a search engine to try to find the website. One should also think of acronyms, synonyms, related words or word combinations that could be used. It is advisable to keep keywords in lower case since most users do not capitalise most proper nouns when searching on the Internet. Names are traditionally capitalised and should be treated this way when used as keywords (Du Preez 2002:119).

Be aware that most search engines do have limits on how many keywords are viewed, so make sure those that are used are as concise and as specific as possible (Alimohammadi 2003:241).

There is the problem of misuse with the keywords meta-tag. It is called 'spamindexing', 'spamdexing' or 'spamming'. Misuse includes repeating keywords over and over or using words that really have nothing to do with the website but are of interest to people, for example 'sex'. Keep in mind that although the misuse of keywords will increase traffic to the website, hate mail will also increase. Most larger search engines now also counter spamming strategies by ignoring the entire tag (Alimohammadi 2003:241).

Conclusion

The purpose of online indexing or web indexing was explained throughout this chapter. Attention was paid to indexing pitfalls when embarking on an online or web indexing project. Some web indexing questions and considerations were also answered. Indexing

both as a navigational tool and for information retrieval (IR) was then discussed. These included discussions on HTML coding and the use of meta-tags to assist search engines in accurately indexing the website.

BIBLIOGRAPHY

Alimohammadi, D. 2003. Meta-tag: a means to control the process of web indexing. *Online Information Review*, 27(4):238-242.

Broccoli Information Management. 2003. *Web site indexing*. Available: www.bim.net/website_indexing.html (accessed on 4 December 2003).

Broccoli, K & Van Ravenswaay, G. 1999. Web indexing, anchors away, in *Beyond book indexing*, edited by D Brenner and M Rowland. Phoenix: American Society of Indexers:37-42.

Browne, G & Jermey, J. 2001. *Website indexing: enhancing access to information within websites*. Adelaide: Auslib.

Cleveland, DB & Cleveland AD. 2001. *Introduction to indexing and abstracting*. 3rd ed. Englewood: Libraries Unlimited.

Denenberg, R. 1996. *Structuring and indexing the Internet*. Available: http://www.loc.gov/z3950/agency/papers/italy.html (accessed on 4 December 2003).

Desmontils, E & Jacquin, C. 2003. *Indexing a web site with a terminology oriented ontology*. Available: http://www.sciences.univ-nantes.fr/irin/indexGB.html (accessed on 4 December 2003).

Du Preez, M. 2002. Indexing on the Internet. *Mousaion*, 20(1):109-122.

Indexing the web. 2003. American Society of Indexers. Available: http://www.asindexing.org/site/webindex.shtml (accessed on: 4 December 2003).

Lathrop, L. 2003. *Consideration in indexing online documents*. Available: http:///www.bwa.org/articles/considerations_in_indexing_online_documents.htm (accessed on 4 December 2003).

Maislin, S. 2000. Ripping out the pages, in *Beyond book indexing*, edited by D Brenner and M Rowland. Phoenix: American Society of Indexers:43-57.

Maislin, SA. 2003. *What exactly is 'online indexing'?* Available: http://www.taxonomist.tripod.com/indexing/paperless.html (accessed on 4 December 2003).

Rowland, MJ. 2000a. <META> tags, in *Beyond book indexing*, edited by D Brenner and M Rowland. Phoenix: American Society of Indexers:71-76.

Rowland, MJ. 2000b. Website indexing: prepared for the Web Indexing Workshop, presented at the June 1999 Annual Conference of the American Society of Indexers and updated July 2000. Marisol Productions. Available : http://www.marisol.com (accessed on 4 December 2003).

Sullivan, D. 1998. *How to use meta tags*. Available: http://searchengingwatch.com/webmasters/article.php/2167931

Wright, JC. 1997. How to index online. *The Indexer*, 20(3):115-126. Also available: http://www.mindspring.com/~jancw (accessed on 4 December 2003).

12 METADATA AND INDEXING

Ansie Watkins

Abstract

This chapter introduces the reader to the concept of metadata and its purpose, and gives an overview of the various types of metadata. The importance of standards is emphasised, with a brief discussion of the most frequently used standards with regard to metadata and the inter-operability of systems. The Dublin Core as a widely used metadata standard is discussed in some detail. Finally, the steps of creating metadata during a digitisation project are discussed.

What is Metadata?

Librarians have been creating metadata since the earliest times, when they documented lists of items such as clay tablets, parchment scrolls and handwritten manuscripts in their collections. The term 'meta' originates from a Greek word meaning 'with, next, after, alongside' and, more recently, in Latin or English it refers to something beyond nature. Metadata can be described as data about other data, in this case information resources. It describes the characteristics of these information resources that can be in various formats but are mostly associated with networked information resources on the Internet. The activity of creating metadata may be compared with cataloguing and indexing that take place in libraries, archives and museums.

Why Metadata?

The purpose of metadata is to describe the content of an information resource (descriptive metadata). Apart from content description metadata is also used for administrative control, security, personal information, management information, content rating and resource identification.

To be effective for organisations that interact, metadata must be structured and consistent in its application. The ideal is therefore that a metadata standard is widely applicable. Metadata provides a link between the creator and the user of the information. It adds value to networked electronic resources. It enhances the retrieval process by enabling potential users to find relevant information easily and to exclude irrelevant information from their search results.

The term aims increasingly at the identification, location and description of these resources and to facilitate inter-operability of various electronic collections, such as digital images

of artworks, electronic texts of literary works, digital collections of research materials, electronic books and journals and many others. These resources are created by several different organisations, such as publishers, graphic designers, photographers, universities and digital libraries.

A metadata record consists of a set of predefined elements that represent specific attributes of an information resource and each element can have more than one value. This record is associated with an electronic resource that can be in various formats. The set of elements may exist as a separate record that is linked with the electronic resource or as part of the electronic resource itself.

How do the Concepts 'Metadata' and 'Indexing' Relate?

Indexing of content description is important in any digital project. Creating an index of values to describe digital resources is a vital component of the digitising process. Effort required to produce an index varies, depending on the nature of objects and depth of retrieval required. At some point in the process, human intervention is needed to assign values to the index terms. The index can be built after the conversion of an object into digital form.

Assigning index terms will add value to searching and retrieval of digital resources. All the traditional indexing practices and principles also apply in the case of indexing digital resources. These index terms will be included in the elements of the metadata structure that will be used to describe the electronic resources. The metadata structure is pre-defined and this process will be discussed later on in this chapter.

As in the case of 'traditional indexing' it is also necessary to decide on an indexing standard and format for the indexing of digital objects. Consistency is important to achieve valid data. Several indexing models have been developed for indexing digitised objects. The most commonly used indexing standards are MARC cataloguing and the Dublin Core metadata scheme.

Different Levels and Types of Metadata

One may distinguish different types of metadata, all of which perform different functions but which are often interrelated. Descriptive, structural and administrative metadata may be created at collection or item level.

Levels of Metadata

Collection-level Metadata

Collection level metadata refers to data about collections of information resources. It is used for a holistic description of a collection, such as archival collections or journals.

Item-level Metadata

This level of metadata describes individual items within a collection, for example photographs or correspondence in an archival collection, and articles in an electronic journal.

Different Types of Metadata

Descriptive (Content) Metadata

This type of metadata refers to the bibliographic and subject description of an entity. This type is the most familiar to us, for example MARC, Dublin Core, TEI (Text Encoding Initiative) and EAD (Encoded Archival Description).

Structural Metadata

Structural metadata is information that is relevant to the presentation of the digital object to the user, describing it in terms of navigation and use. Typical examples of structural metadata elements are

- unique identifier
- content type (text, digital image, file formats – XML, jpg, pdf)
- data file size
- structural divisions (which indicate various parts of the document, e.g. letters in a diary)

The University of California, Berkeley, defined an excellent example during the Making of America II project.

Administrative Metadata

Administrative metadata allows the repository to manage its collection, and includes the following:

- data related to the creation of the digital image (date of scan, resolution, etc.)
- data that can identify a version or edition of the image and help determine what is needed to view or use it
- ownership, rights and reproduction information

Some metadata elements may be both structural and administrative and may be used for similar purposes. Administrative metadata is critical for long-term management. Without well-designed administrative metadata, a file may be unrecognisable and unreadable within a decade. Examples of the elements in administrative metadata include

- source type – to identify the material from which the digital file was created
- source physical dimensions
- source ID – a local catalogue for a book or accession number for a special collections item
- scanning date
- scanning resolution
- owner of the copyright
- copyright date
- distribution restrictions

Metadata Standards

Literally hundreds of metadata standards have been created and are used throughout the world. The reason for this is to accommodate information resources that exist in several diverse

forms. To be effective, metadata must be structured. Another requirement for metadata to be effective is that it should be widely applied in a consistent manner. The goal of the creator of metadata should be to choose a popular and persistent standard that offers inter-operability with other metadata databases and one that is simple enough for public use.

A few examples of the most frequently used metadata schemes that became international standards for descriptive metadata have been mentioned above. The TEI and EAD will be briefly described later in the chapter. The Dublin Core will be discussed in more detail to illustrate how metadata is created and applied in practice.

Administrative, technical and structural metadata are more often uniquely defined by the institutions who created or are hosting collections of digital resources and aimed to meet their specific requirements. These elements are not necessarily displayed online because there may be no need in this regard. It is, however, important that once these metadata sets have been defined they should be consistently applied. The use of controlled vocabularies can enhance the consistent retrieval of relevant resources.

Something about Syntax

Data will not be usable unless the encoding scheme understands the semantics of the metadata scheme. Apart from the metadata sets there are therefore also a few other standards related to metadata that should be mentioned, namely those of syntax. This refers to the language or semantics in which the metadata set has been encoded. Encoding plays an important role in the creation of metadata. It differs according to the type of document. Examples of international standards in this regard are as follows.

Standard Generalised Markup Language (SGML)

SGML is a non-proprietary language for describing highly structured information. It is very useful for detailed encoding of information in databases and also for the conversion between different formats. It is rather complex, requires a high level of technical expertise to use it effectively and 'raw' SGML is not supported by web browsers.

Hypertext Markup Language (HTML)

HTML is the standard text formatting language for documents on the Internet and much simpler to use than SGML. It consists of text files encoded with tags that tell the computer how to format text on the screen, but is not recommended for highly structured information. HTML tags can also be used to encode metadata. *HTML Meta-Tags* is an initiative of meta-tagging web pages in order to make them searchable by means of search engines. Consult the following websites for more details about meta-tags:

How to use HTML
Meta_Tags http://searchenginewatch.com/webmasters/article.php/2167931
A Dictionary of HTML
Meta-Tags vancouver-webpages.com/META/**metatags**.detail.**html**

eXtensible Markup Language (XML)

eXtensible Markup Language (XML) is a subset of SGML that has emerged as a widely used international text-processing standard. It is as powerful and flexible as SGML, but not as complex and also supported by web browsers. It is selected by many international initiatives as an encoding standard and is used for several metadata schemes.

The following Internet links can be consulted for more detailed information on XML:

eXtensible Markup Language (XML)
http://www.w3.org/XML/
Sgml/Xml Based Metadata Initiatives
http://www.ifla.org/II/metadata.htm#sgmlxml

Resource Description Framework (RDF)

The RDF is a basic language, developed by the Worldwide Web Consortium (W3C) for the writing of metadata and processing of it on the Internet in a flexible way. Therefore various institutions can define and use metadata according to their needs, but also according to international standards. XML is used to write RDF tags.

More information on the RDF developments, including technical documentation, is available at *W3C Metadata area*: http://www.23.org/Metadata/ (Dublin Core Usersguide Glossary http://library.csun.edu/mwoodley/dublincoreglossary.html)

Examples of Metadata Standards

As mentioned before several metadata schemes are used throughout the world to describe different types of information resources. A few well-developed examples will be discussed in this section.

Text Descriptors: Text Encoding Initiative (TEI)

The Text Encoding Initiative is an ongoing international research project that has developed an extremely flexible set of guidelines, mostly for describing academic electronic texts, such as literature and reference works. It may be used for a single work or a collection of single works.

A TEI document consists of a bibliographic header and the text. Both these components are encoded in great detail. XML is increasingly being used as the encoding language, but many TEI documents are still in SGML. TEI documents can be created by means of XML editors and indexed by means of search and display software.

An explanation of how to use the TEI Guidelines to create electronic text will take up a reference source on its own. Therefore more detailed information on the TEI may be accessed on the following website:

The Text Encoding Initiative
http://www.tei-c.org/P4X/

Encoded Archival Description (EAD) for Archival Finding Aids

The EAD had its origin at the University of California, Berkeley. It was developed as a non-proprietary standard for machine-readable archival finding aids, namely inventories, registers, indexes and other documents created by archives, libraries, museums and repositories to support the use of their holdings:

- The information in a finding aid describes, controls and provides access to other information – not itself.
- It preserves and enhances the current functionality of existing inventories.
- The standard is intended to facilitate interchange and portability.
- The needs of public users, curatorial and reference staff were given priority in the standard's design.

The finding aid consists of the following segments:

- information about the finding aid itself
- a 'front matter' or title page
- the actual finding aid
- hierarchically organised information
- adjunct information that facilitates the use of the papers by researchers (e.g. a bibliography)

The EAD header is based on the TEI header and consists of

- EAD identity
- file description
- profile description
- revision description
- footer

Other Examples of Metadata Standards

These are as follows:

- music (Standard Music Description Language)
- images and objects descriptors (Categories for the Description of Works of Art, Consortium for the Computer Interchange of Museum Information [CIMI])
- classification (CyberDewey)
- numeric data (Standard for Survey Design and Statistical Methodology Metadata [SDSM])
- geospatial data (Content Standards for Digital Geospatial Metadata)
- MARC (Machine Readable Cataloguing)

Inter-operability of Metadata

The most important requirement for success of metadata standards is that it should be inter-operable and widely applied in a consistent manner. Metadata harvesting and metadata crosswalks will be discussed as two concepts that are associated with the inter-operability of metadata that will enhance accessibility and availability of networked electronic resources.

Crosswalks

A metadata crosswalk is a table that maps relationships between two or more metadata formats. It supports the ability of search engines to search effectively across different databases; in other words it promotes inter-operability.

The Library of Congress has prepared a useful article on MARC to Dublin Core Crosswalk that can be viewed at http://www.loc.gov/marc/marc2dc/html

Metadata Harvesting

The Open Archives Initiative (OAI) is an example of how metadata can be used to add value to electronic information resources. It develops and promotes inter-operability standards that aim to facilitate the efficient dissemination of content by means of a process that is known as metadata harvesting. A harvester collects metadata from various repositories. A repository is a server that is accessible for collection metadata. See http//www.openarchives.org/documents/FAQ.html

Dublin Core Metadata

What is the Dublin Core?

The Dublin Core metadata standard is a very effective element set for the description of a wide range of networked electronic resources. It is an example of descriptive metadata, but also contains elements that can serve as structural or administrative metadata. The Dublin Core also recommends a wide range of controlled vocabularies that can be applied to various elements.

See http://dublincore.org/documents/usageguide/

Advantages of the Dublin Core

The Dublin Core is often criticised for not making provision for all the types of information that people would like. But its purpose is to merely serve as a lowest common denominator for describing electronic resources. The advantages of the Dublin Core are as follows:

- It is usable and flexible.
- Its semantics are simple enough to be understood by a wide range of users without intensive training.
- The elements are easily identifiable.

- It is not intended to replace other resource descriptions, but rather to complement them. Other important metadata such as archival and accounting data were deliberately excluded to keep it as simple and usable as possible.
- It is mostly syntax-independent to support the widest range of applications.
- All elements are optional, but allow each site to define which elements are mandatory and which are optional.
- All elements are repeatable.
- The elements may be modified in limited and well-defined ways through the use of specific qualifiers, such as the name of a thesaurus used in subject qualifiers.
- It can be extended to meet demands of more specialised communities.
- Elements can be added for site-specific applications and it could be mapped to a more controlled system such as MARC.
- Dublin Core has received wide acceptance among the electronic information community and is seen as the unofficial Internet metadata standard.

See 'An Introduction to Metadata' at http://www.itb.hu/fejlesztesek/meta/hgls/core/Background/An_Introduction_to_Meadata.htm

The Dublin Core Element Set

The Dublin Core standard includes two levels: Simple and Qualified. Simple Dublin Core consists of 15 elements. Qualified Dublin Core has an additional element, namely Audience. It also contains qualifiers that describe the elements in more detail and may be useful during the retrieval process.

The Dublin Core elements can be defined in three categories:

- Elements related to the content of the resource:
  - title
  - description
  - subject
  - source
  - language
- Elements related to intellectual property:
  - creator (person/organisation responsible for intellectual or creative content)
  - publisher
  - contributor
  - rights
- Elements related to the physical manifestation:
  - date
  - type
  - format
  - identifier

Each element is optional and repeatable.

Dublin Core Metadata Template

Dublin Core Element Set Version 1.1- Reference Description http://dublincore.org/documents/dces/

The following table defines the various elements with brief comments on how they are used and also on standards and controlled vocabularies that may be used to qualify the elements. See the Dublin Core website for more information.

Elements	Definition	Comments
Title	A name given to the resource	Typically, a Title will be a name by which the resource is formally known
Creator	Person/Organisation responsible for creating the intellectual or creative content of the resource	Examples of a Creator include a person, an organisation or a service. Typically, the name of the Creator should be used to indicate the entity.
Subject	The topic of the content of the resource	Typically, a Subject will be expressed as keywords, key phrases or Classification Codes that describe a topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme.
Description	An account of content of the resource	Description may include (but is not limited to) an abstract, table of contents, reference to a graphical representation of content, or a free-text account of the content.
Publisher	An entity responsible for making the resource available to public	Examples of a Publisher include a person, an organisation, or a service. Typically, the name of a Publisher should be used to indicate the entity.
Contributor	An entity responsible for making contributions to the content of the resource	Examples of a Contributor include a person, an organisation, or a service. Typically, the name of the Contributor should be used to indicate the entity.
Date	A date associated with an event in the life cycle of the resource	Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601 [w3CDTF] and follows the YYYY-MM-DD format.
Type	The nature or genre of the content of the resource	Type includes terms describing general categories, functions, genres or aggregation levels for content. Recommended best practice is to select a value from controlled vocabulary (e.g. the working draft list of Dublin Core Types [DCT1]) To describe the physical or digital manifestation of the resource, use the format element.

Elements	Definition	Comments
Format	The physical or digital manifestation of the resource	Typically, Format may include the media-type or dimensions of the resource. Format may be used to determine the software, hardware or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (e.g. the list of Internet media types [MIME] defining computer media formats).
Identifier	An unambiguous reference to the resource within given context	Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. For example formal identification systems include the Uniform Resource Identifier (URI) (including the Uniform Resource Locator [URL]), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN).
Source	A reference to a resource from which the present resource is derived	The presence of the resource may be derived from the Source resource in whole or in part. Recommended best practice is to reference the resource by means of a string or number conforming to a formal identification system.
Language	A language of the intellectual content of the resource	Recommended best practice for the values of the Language element is defined by RFC 1766 [RFC1766] which includes a two-letter Language code (taken from the ISO 639 standard [ISO639]), followed optionally, by a two-letter Country code (taken from the ISO 3166 standard [ISO3166]). For example, 'en' for English, 'fr' for French, or 'en-uk' for English used in the United Kingdom.
Relation	A reference to a related resource	Recommended best practice is to reference the resource by means of string or number conforming to a formal identification system.
Coverage	The extent or scope of the resource	Coverage will typically include spatial location (a place name or geographical coordinates), temporal period (a period label, date or date range) or jurisdiction (e.g. named administrative entity).
Rights	Information about rights held in and over the resource	Typically, a Rights element will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright and various Property Rights. If the Rights elements is absent, no assumptions can be made about the status of these and other rights with respect to the resource.

The Nordic Metadata Project offers an online template for creating Dublin Core metadata. If you use the metadata created by this form and follow the examples, term lists and recommendations, your HTML documents will carry high quality metadata. Please consult the website for more details: http//www.lub.lu.se/cgi-bin/nmdc.pl

Creating Metadata

Process of Creating Metadata during a Digital Project

Metadata may be deployed in various ways, such as

- embedding metadata in a web page – metatags in HTML encoding
- separate HTML document linked to the resource it describes
- in a database linked to the resource

Descriptive metadata for digitised documents are often found in an external database. In other words the index terms and their values are stored in a separate database from the images. Once images have been created and stored in a database, discovery and retrieval become critical functions. Indexing each image is necessary to ensure the location of images during a search. The digital resource can only be useful if it contains metadata to allow users to navigate the digital file.

The steps in creating metadata during a digitising project can be summarised as follows:

- Select a suitable descriptive metadata scheme.
- Establish administrative metadata.
- Decide on structural metadata.
- Decide on technical standards for digital images, electronic texts, file formats, software tools, and so on, to create digital resources and metadata templates.
- Establish the naming convention for a set of filenames for unique identifiers and match with corresponding record.
- Set filenames, attribute fields (elements), order, checklist.
- Develop indexing protocols to ensure consistency.
- Decide on controlled vocabularies if applicable.
- Develop a quality control system.
- Develop metadata template to create metadata records.
- Train staff to create digital resources (if applicable), apply indexing principles and use the templates and other software tools as needed.
- Create digital resources.
- Create metadata records by assigning values to the pre-defined elements.
- Run the automated indexing program to enable the search engine to retrieve the records during an online search.
- Create or customise a user-friendly interface that will enable clients to search and view electronic resources online.

- Evaluate the system.
- Market the system.
- Teach clients to use the system.

Tools for Working with Metadata

Several software tools are available for the following:

- creating metadata
- creating/change of templates
- automatic extraction/gathering of metadata
- converting between formats
- integrated (tool) environments
- formatting and reviewing tools
- XML tools
- publishing tools
- searching tools

Tools can be downloaded from

Metadata UKOLN Software tools: http://www.ukoln.ac.uk/metadata/softwaretools/

Dublin Core Metadata Initiative (DCMI) Tools and software: http: dublincore.org.tools

Tools for working with metadata: http:///www.csc.noaa.gov/metadata/metatools.html

The IFLA website offers a comprehensive and detailed list of metadata resources: http://www.ifla.org/II/metadata.htm

Cataloguers Toolbox: http://staff.library.mun.ca/staff/toolbox/standards.htm

Conclusion

The Internet is an extremely dynamic environment. Therefore the optimisation of access to Internet resources offers a challenge to librarians and information workers. They should become proactively involved in the description of resources from which their clients will benefit. The ongoing increase in the creation and use of networked electronic resources will impact on the demand for metadata. Metadata will become increasingly important as we develop library systems that are inter-operable.

Although the few standards that were discussed here are well tested and widely used, this should not prevent librarians from exploring new opportunities in the field of metadata. Information resources are continually created in new or improved formats. Metadata standards should keep up with new developments. The guidelines in this chapter are based on resources prepared by international initiatives in the field of metadata. References as indicated throughout the text are constantly updated and should be consulted on an ongoing basis.