Thesauri on the Web: current developments and trends

Ali Asghar Shiri and Crawford Revie

The authors

Ali Asghar Shiri is a PhD student and **Crawford Revie** is a Senior Lecturer, both in the Department of Information Science at the University of Strathclyde, Glasgow, UK.

Keywords

Information retrieval, Software, Internet, Indexes

Abstract

This article provides an overview of recent developments relating to the application of thesauri in information organisation and retrieval on the World Wide Web. It describes some recent thesaurus projects undertaken to facilitate resource description and discovery and access to wide-ranging information resources on the Internet. Types of thesauri available on the Web, thesauri integrated in databases and information retrieval systems, and multiple-thesaurus systems for cross-database searching are also discussed. Collective efforts and events in addressing the standardisation and novel applications of thesauri are briefly reviewed.

Electronic access

The current issue and full text archive of this journal is available at

http://www.emerald-library.com

Online Information Review
Volume 24 · Number 4 · 2000 · pp. 273–279
 · ISSN 1468-4527

Introduction

The rapid growth of networked information resources together with changes and innovations in the ways in which information is represented have necessitated a reassessment of the tools and techniques used for information management. World Wide Web technologies play a central role in refurbishing and redesigning information management tools.

Thesauri have played an important role in modern information storage and retrieval systems. While initial proposals to utilise thesauri focused on their ability to ensure consistent analysis of documents during input to information retrieval systems, they have increasingly become vital as aids to effective retrieval. Indeed, in the near future, it appears likely that thesauri will be used more during retrieval than at input (Milstead, 1998). The move to increasing use of thesauri as an aid to retrieval has expanded their functional span within information retrieval systems. As Aitchison *et al.* (1997) have noted, the role of the thesaurus is changing, but it is likely to remain an important retrieval tool.

This refocusing of the use of thesauri within information retrieval systems means that it is imperative that professionals take cognisance of the potential of thesauri as essential components of the largest information retrieval environment, namely the World Wide Web.

Why thesauri on the Web?

Although there are few operational information retrieval systems which have effectively incorporated thesauri as search and retrieval aids, we are witnessing an increased enthusiasm among thesaurus developers to make their thesauri available on the Web for potential applications. The reasons for this enthusiasm and the increasing availability of online thesauri are closely linked to key issues associated with the emergence of the World Wide Web, including:

- the colossal growth of information resources demanding their better subject identification;

Refereed article received 20 June 2000
Approved for publication 10 July 2000
This paper is based on a presentation delivered at the First Consultation on Agricultural Information Management (COAIM) conference, organised by FAO and held in Rome 5-7 June 2000.

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

- the migration of traditional information resources to the Web calls for more consistent subject approaches;
- an urgent need for resource description and discovery through reusing the existing information management tools such as controlled vocabularies;
- problems associated with quality of unstructured information retrieved from the Web;
- the need to provide users with knowledge structures such as thesauri for rapid and easy access to better organised information.

The emergence of vast amounts of information on the Web has created a new dimension in information retrieval research: Web-based information retrieval. As part of the attempt to apply information retrieval tools and techniques to manage information within this new medium in a more effective and efficient manner, the potential of thesauri to aid users' searching and browsing processes is being investigated.

Users, on the other hand, need good conceptual and semantic tools more than ever as they attempt to effectively organise the vast volumes of information available on the Web. The semantic structures provided by thesauri can play a part in both organising and retrieving Web information and knowledge resources.

Types of thesauri on the Web

Most thesaurus software producers are realising the potential of Web related programs such as HTML, JAVA and XML for thesaurus building and editing. Many of these producers have provided new versions of their traditional thesaurus management applications with more user-friendly Web interfaces and graphical displays together with navigation and browsing capabilities. These efforts in refurbishing traditional thesaurus applications have brought along thesaurus browsers and navigators.

Davies (1996) suggests that thesauri can be published over the Web in either static or dynamic form and the choice between the two methods of Web publication influences the format and organisation of the thesaurus.

A range of approaches has been used to publish thesauri on the Web, resulting in thesauri with varying levels of sophistication. Web-based thesauri can be categorised as follows in terms of their publishing format and structure:

- thesauri in simple static text formats (ASFA Thesaurus);
- thesauri in HTML format but still static, without effective use of hyperlinks (Inforterm);
- thesauri in dynamic HTML format with fully navigable hyperlinks (MeSH);
- thesauri with advanced visual and graphical interfaces (Plumb Design Visual Thesaurus);
- thesauri in XML format (Virtual HyperGlossary).

This diversity of tools and techniques has both helped and hindered thesauri as aids to searching and browsing the Web. The best examples have used graphical techniques to provide users with access to the rich cross-references and multiple levels of thesaural relationships possible using these approaches.

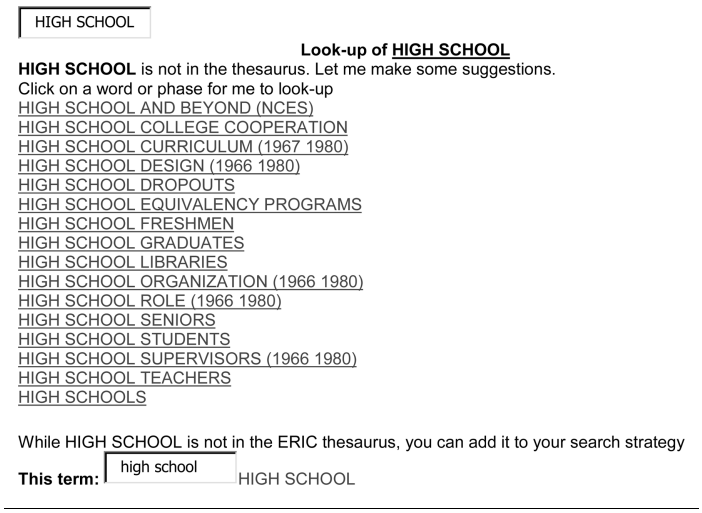
Web-based thesauri can also be divided into two general types in terms of their functionality and usability: standalone thesauri which are not a part of an information system (e.g. ASIS Thesaurus of Librarianship and Information Science) and those which are fully integrated into databases or information retrieval systems (e.g. the ERIC thesaurus, which is totally adopted and integrated within the ERIC database by SilverPlatter). References to a range of thesauri on the Web are provided at the "Controlled vocabularies resource guide" Website.

Thesauri as integral parts of databases

There are an increasing number of thesauri available on the Web covering a wide variety of subjects, formats, displays and languages. There are over 50 major thesauri covering around 30 subjects and the number is steadily increasing. The thesauri operating in connection with a database or an information retrieval system support different levels of user searching and browsing. Some major examples of thesauri that are fully integrated into databases are as follows.

The ERIC Thesaurus on the Web as implemented by the Assessment and Evaluation Clearinghouse provides sophisticated support in a sense that it attempts to provide the user with a suitable starting point and then to expand the query at different levels of thesaurus hierarchy in a user-friendly environment (see Figure 1). A useful feature of this implementation of the ERIC is that if the user enters a term, for

Figure 1 ERIC search wizard



example "high school", which is not in the thesaurus as a descriptor, the system informs the user and also provides several options of related terms to modify further browse and search entry.

Another example of a thesaurus which is explicitly used for both browsing and searching is the Humanities and Social Science Electronic Thesaurus (HASSET) developed and maintained by the University of Essex. HASSET was developed to allow subject access to the Data Archive, which is a specialist national resource containing the largest collection of accessible computer readable data in the social sciences and humanities in the UK. Using the HASSET subject thesaurus, users can extend their searches using different levels of hierarchy, e.g. broader or narrower terms. This makes the term selection process more flexible as users have access to both the thesaurus and the bibliographic database. HASSET also has different display formats with a variety of search and browse features.

The UK National Digital Archive Datasets (NDAD) which preserves and provides access to an emerging category of public record – computer datasets from UK government departments and agencies – is using a thesaurus whose construction is based on the UNESCO thesaurus. Searching the NDAD thesaurus will allow the user to find catalogues and administrative histories which relate to the subjects in the thesaurus.

The users can browse the thesaurus either hierarchically or alphabetically. Different thesaural relationships are nicely demonstrated and the user can easily choose among related, narrower or broader terms to be included in the search statement.

The MeSH thesaurus of the National Library of Medicine is a controlled vocabulary which consists of more than 19,000 main headings. In addition to these headings, there are 103,500 headings called Supplementary Concept Records. The MeSH browser may be used to find descriptors of interest and see these in relationship to other descriptors. This vocabulary aid is designed to help quickly locate descriptors of possible interest and to show the hierarchy in which these descriptors appear. The browser displays virtually complete MeSH records, including the scope notes, annotations, entry vocabulary, history notes, allowable qualifiers, etc. One of the most interesting features of the MeSH browser is its vast and extensive terms and vocabulary entries for different levels of searching. The user can browse through a MeSH tree structure or choose main headings, qualifiers and supplementary concepts. It provides a vast array of entry terms for users to choose from.

One of the most interesting examples of thesaurus integration with bibliographic databases is the work of Ovid Technologies. If the user enters a term, the term is automatically mapped to the subject headings of the database. By clicking on a subject heading the user can browse the thesaurus-related terms and choose the relevant terms. One of the notable search options of the Ovid products is the "focus" option through which users can limit their search to those documents in which the subject heading is considered the major point of the article.

Bowker-Saur's LISAnet is the Library and Information Science Abstracts database on the Web which allows the searchers to have easy access to its thesaurus for searching the database. Searchers can browse and search the terms in the thesaurus and through an readily accessible "paste" button, send the descriptor(s) to main search template for searching the database. It helps the searcher to easily identify and select alternative search terms and to add them for query reformulation purposes.

Multi-thesaurus searching and browsing systems

This enthusiasm for using modern Web technology to publish thesauri on the Web has resulted in a growing number of thesauri and the need to think of thesauri interoperability as well as the need for accessing and using

different thesauri for search and retrieval. As Fidel (1992) has showed in her research, better quality and availability as well as support for multi-database searching are likely to increase the use of controlled vocabularies. This is happening as we move toward using different thesauri for cross-database searching. The use of the Resource Description Framework (RDF), which was proposed by the World Wide Web Consortium, can also be considered one of the major breakthroughs in Internet-based application of thesauri and proving the ground for thesauri interoperability. The RDF namespace concept allows the controlled usage of distributed vocabulary systems. It also provides a syntax (XML) for exchanging controlled vocabulary data with other applications and services (Koch, 1999).

Currently, a number of projects are underway to enhance the level of access to multiple thesauri on the Web. One of the recent developments in this regard is the CERES (Californian environmental resources evaluation system) Thesaurus project for multiple environmental thesauri.

The purpose of this project is to construct an integrated controlled environmental vocabulary, and to develop the tools to use it for creation of metadata and construction of queries. Currently, the CERES program is involved in development of a key component, a controlled vocabulary database and user environment that will: provide pick list(s) for selection of keywords, synonyms, and related concepts for use in metadata and queries; provide a hierarchical organisation of information to serve as a browsing structure in information discovery; and allow simultaneous browsing and comparison of terms as presented in multiple standard thesauri in use nationally and internationally to index environmental information.

Knowledgecite Library (KCL) is a new online service which makes use of a multi-thesaurus tool for cross-database searching. If the user enters a term, knowledgecite will list thesaurus terms that match the user words or the phrase (e.g. cross-references, related, broader and narrower terms or scope notes) of all thesauri built into KCL, or those in the discipline (e.g. social sciences) that the user chose. Besides these thesaurus terms there are also cross-references to other thesauri (Jacso, 1999).

Multi-thesauri management system with Web interface is also another new development in using multiple thesauri. This project aims to provide a means for searching distributed databases of alternative medicine produced in various countries. The thesaurus management system performs at two levels, both with a Web interface: a search site open to anyone who wants to do cross-thesauri searching and browsing and a thesaurus maintenance site for editing the thesauri. At the moment, an alphabetical display, index of all terms, hierarchical display, and rotated display are the options for uploading, downloading, printing, and browsing. Once a hierarchical list or an alphabetical thesaurus is uploaded, other lists/displays can be generated automatically. Cross-thesauri searching for particular terms is performed based on exact and partial matching. Search terms are directly linked to the entries in a thesaurus. Our next step is to implement database searching.

GenThes or General Thesaurus Browser is another example of a facility for cross-references among thesauri which is able to handle several heterogeneous, multilingual thesauri. It is currently used by several environmental catalogue systems on the Web (Nikolai, 1999).

The Unified Medical Language System (UMLS) Metathesaurus is one of the largest thesauri in the area of medicine and related fields. It contains information about biomedical concepts and terms from many controlled vocabularies and classifications used in patient records, administrative health data, bibliographic and full-text databases and expert systems. The 2000 edition of the metathesaurus includes more than 730,000 concepts and 1.5 million concept names from over 50 different biomedical vocabularies, some in multiple languages. The metathesaurus is used in a wide range of applications including: linking between different clinical or biomedical vocabularies; information retrieval from databases with human assigned subject index terms and from free-text information sources; linking patient records to related information in bibliographic, full-text, or factual databases; natural language processing and automated indexing research; and structured data entry. The UMLS Metathesaurus is available to licensees via ftp, Web interface, and applications program interface (API) from the UMLS Knowledge Source Server. It is also available on CD-ROM by request.

The Consumer Health Terminology (CHT) Thesaurus provides consumers with online access to more than one million medical phrases and terms. The CHT Thesaurus is derived from Lexical's Metaphrase®

Online Information Review
Volume 24 · Number 4 · 2000 · 273–279

Thesaurus, which in turn is based upon the UMLS® Metathesaurus®, a compendium of controlled medical vocabularies maintained in part by Lexical under contract to the National Library of Medicine.

The CHT Thesaurus contains approximately one million medical definitions, phrases and terms, which have been inter-related over the past ten years. The CHT Thesaurus is especially useful to consumers because it includes more than 14,000 WellMed-developed consumer and lay terms that are mapped into these other vocabularies.

Using the CHT Thesaurus, search engines will be able to refine common search, providing users with the most appropriate information on the first try.

SNOMED, the Systematized Nomenclature of Medicine, is another comprehensive, multiaxial nomenclature created for the indexing of the entire medical record, including clinical findings, etiologies and interventions. As a reference terminology, SNOMED allows for consistent gathering and transmitting of detailed clinical information, retrieval of information for disease management or research and performance of outcome analysis for quality improvement.

Thesauri application in subject-based information gateways

Koch (2000) defines subject gateways as:

> Internet-based services which support systematic resource discovery. They provide links to resources (documents, objects or services), predominantly accessible via the Internet. Browsing access to the resources via a subject structure is an important feature.

Subject access through some kinds of knowledge structures like thesauri and classification systems is one of the significant features of good quality subject gateways. These quality-controlled subject gateways have established procedures for selection and content description of Web pages and also use thesauri for careful and consistent resource description.

Recently, several subject-based information gateways have been developed on the Web which use thesauri for indexing and retrieval of Web pages and Web sites. The following are some examples:

- Art, Design, Architecture and Media information gateway (Art and Architecture thesaurus);
- Engineering Electronic Library, Sweden (Engineering Information's EI thesaurus);
- Organising Medical Networked Information (Medical Subject Headings (MeSH) thesaurus);
- Social Science Information Gateway (HASSET thesaurus).

These subject gateways use thesauri to manually or automatically index Web pages and provide structured and more consistent subject access for browsing and searching the Web pages.

The Social Science Information Gateway (SOSIG), in its advanced search mode, provides a thesaurus search interface through which the user can enter a term and browse through and choose from the narrower, broader and related terms. The user can also send the query directly to SOSIG catalogue for searching the Web pages indexed by the term(s).

NKOS workshops: key discussions on thesaurus applications on the Web

The Networked Knowledge Organisation Systems (NKOS) workshops were initiated in 1997 within the ACM Digital Library Conference to address the issues of creating interactive knowledge organisation systems, including thesauri, over the Internet. The first workshop focused on using thesauri for searching and generating metadata. The use of thesauri in digital libraries, thesaurus-based metadata, subject vocabulary in distributed search environment and automated classification of networked information were among the main issues discussed in the workshop. This workshop aimed to gather information on research and products relating to the use of thesauri as metadata content tools.

The second workshop in 1998 was entitled "Application of terminology and classification tools for digital collection development and networked-based search". It focused on practical aspects of using thesauri in networked environments. Data models to support interactive use, intellectual property models, structural issues relating to terminology, including thesauri and XML, were among the topics addressed. In 1999, the NKOS workshop discussed standards for distributed thesauri on the Net. The application of thesauri to the Resource Description Framework (RDF), thesauri and metadata at Microsoft, and

multi-thesauri management systems with Web interfaces were among the issues for discussion.

This year, the NKOS 2000 workshop focused on ontology and related projects to develop a language for describing and exchanging ontologies as well as projects on constructing and maintaining special controlled vocabularies and terminologies for networked environments.

All the NKOS workshops have been significant events in defining the application of thesauri within the new information environment of the World Wide Web.

In search of a standard for electronic thesauri

The growing number of Web-based thesauri has stimulated discussion on the need to review existing standards for thesauri. The workshop on "Electronic Thesauri: Planning for a Standard" which was held in November 1999 and sponsored by the National Information Standards Organisation (NISO), American Society of Indexers (ASI), and the Association for Library Collections and Technical Service (ALCTS) was an attempt to address the issue of standards as they relate to thesauri. The workshop aimed to investigate the desirability and feasibility of a standard for electronic thesauri which:

- speaks to criteria and/or methods for generating thesauri by machine-aided or automatic means;
- shows semantic relationships among terms, as aids to text and information analysis and retrieval;
- supports a variety of electronic thesaurus displays;
- supports interoperability protocols, structures, and/or semantics applicable to thesauri.

Vocabulary mapping, management and interoperability were among the issues highlighted in the workshop. The need for flexible electronic thesaurus displays to ease and augment indexing and retrieval was also discussed. XML and RDF were put forward as widely supported formats to be useful for different browsing tools. It was pointed out that Web browsers are not thesaurus-aware and existing metadata formats make little use of thesauri. To address these issues it was suggested that special emphasis should be placed on the use of thesauri within metadata.

The workshop concluded that a standard for thesauri is needed and that this standard should provide for a broad group of controlled vocabularies such as classifications, taxonomies and subject headings. It also emphasised that the primary concern should be with interoperability rather than with construction or display.

Both this standards workshop and the NKOS events clearly demonstrate the need for and significance of thesauri as knowledge organisation tools. Moreover, one of the positive results of such joint efforts has been the emergence of a group of researchers and developers who are working toward creating and maintaining such knowledge organisation structures. This should provide good groundwork for future joint thesaurus projects and research initiatives.

Issues and areas of concern for the future

Due to the lack standards for publishing thesauri on the Web, a variety of formats, structures and features are currently in use. This causes problems concerning thesauri interoperability, reusability and shareability. There is an urgent need to examine semantic and syntactic tools, formats, and standards, used by Web-based thesaurus publishers and to find ways in which these aspects can be harmonised or integrated.

Many Web-based thesauri are not fully embedded as search and browse aids in databases, information retrieval systems and emerging Web search engines. These tools can be effectively utilised by search engines for more consistent and unified resource description and discovery.

Web-based thesauri can also be considered as tools for query formulation, refinement and expansion and can help users define their information needs more precisely and clearly. However, efforts are required to assess the extent to which these tools can contribute to more effective and reliable information retrieval within the context of the Web.

One of the major areas of concern relating to Web-based thesauri is the lack of evaluation studies which address their usability and functionality. At present, the growing number of thesaural interfaces with different searching and browsing features have not been subjected

to evaluative analysis. While traditional thesauri have been difficult to consult even for indexers, modern thesaurus browsers and navigators promise to become easy to use thanks to the Web-related technologies. Availability and accessibility of thesauri, particularly on the Web, has opened new doors for information users of all kinds and levels including scientists, researchers, and lexicographers. Therefore, it is of paramount importance to look into the variety of thesaurus consultation behaviours and to take account of users' preferences and perspectives in order to ensure their satisfaction. Research is also needed to shed light on how and for what purpose different types of people use thesauri and what value these tools can add to the information searching process in emerging information environments.

As noted earlier, some of the metadata formats on the Web are utilising thesauri for consistent subject assignment, yet it is not clear how much improvement thesaurus-based tagging will contribute to resource description and discovery. Furthermore, the feasibility of applying more structured and well-established thesauri onto different metadata formats is an issue for further investigation.

Concluding remarks

Electronic thesauri have been around for more than three decades and have greatly aided the process of information organisation. The advent of the Web together with recent developments in the application of thesauri as retrieval rather than organisation tools have brought about a new generation of thesauri. These tools are finding their way into a variety of Web-based information organisation and retrieval environments. Examples of their use include the application of thesauri in metadata, Web site/page indexing, Web accessible databases and Web search engines. Further research is needed to evaluate thesauri usability, functionality and effectiveness as distributed knowledge organisation and retrieval tools.

References

Aitchison, J., Gilchrist, A. and Bawden, D. (1997), *Thesaurus Construction and Use: A Practical Manual*, 3rd ed., Aslib, London.

Art, Design, Architecture and Media Information Gateway, http://adam.ac.uk/

Californian Environmental Resources Evaluation System (CERES) Thesaurus http://ceres.ca.gov/thesaurus/thesaurus_tool.html

Consumer Health Thesaurus http://www.cap.org/html/public/thesaurus.html

Controlled vocabularies resource guide http://www.fit.qut.edu.au/InfoSys/middle/cont_voc.html

Davies, R. (1996), "Publishing thesauri on the World Wide Web", *Advances in Classification Research Proceedings of the 7th ASIS SIG/CR Classification Research Workshop Held at the 59th ASIS Annual Meeting*, Baltimore, MD, October, pp. 44-54.

Engineering Electronic Library, Sweden http://eels.lub.lu.se

ERIC search wizard http://www.ericae.net/scripts/ewiz

Fidel, R. (1992), "Who needs controlled vocabularies?", *Special Libraries*, Vol. 83, Winter, pp. 1-9.

Humanities and Social Science Electronic Thesaurus (HASSET) http://dasun1.essex.ac.uk/services/zhasset.html

Jacso, P. (1999), "Savvy searchers do ask for direction", *Online & CD-ROM Review*, Vol. 23 No. 2, pp. 99-102.

Koch, T. (1999), "Automatic classification and content navigation support for Web services. DESIRE II Cooperates with OCLC", June 2000 www.oclc.org/oclc/research/publications/review98/koch_vizine-goetz/automatic.htm#Zthes

Koch, T. (2000), "Quality-controlled subject gateways: definitions, typologies, empirical overview", *Online Information Review*, Vol. 24 No. 1, pp. 24-34.

LISAnet (Bowker-Saur) http://www.lisanet.co.uk

Medical Subject Heading (MeSH) browser http://www.nlm.nih.gov/mesh/MBrowser.html

Milstead, J. (1998), "Thesauri in a full-text world", in Cochrane, P.A. and Johnson, E. (Eds), *Visualizing Subject Access for 21st Century Information Resources, Proceedings of the 1997 Annual Clinic on Library Applications of Data Processing*, Graduate School of Library and Information Science, University of Illinois, Urbana-Champaign, IL, pp. 28-38.

Multi-thesauri management system http://circe.slis.kent.edu/mzeng/thesaurihome.html

Networked Knowledge Organisation Systems (NKOS) workshops www.alexandria.ucsb.edu/~lhill/nkos

Nikolai, R. (1999), "GenThes: a general thesaurus browser for Web-based catalogue systems", www.computer.org/proceedings/meta/1999/papers/49/rnikolai.html

OMNI: Organising Medical Networked Information http://omni.ac.uk/

Ovid Technologies http://demo.ovid.com/demo/ovidWeb/ovidWeb.cgi

Social Science Information Gateway (SOSIG) http://www.sosig.ac.uk

Systematized Nomenclature of Medicine (SNOMED) http://ncvhs.hhs.gov/990517t5.htm

UK National Digital Archive Datasets (NDAD) Thesaurus http://ndad.ulcc.ac.uk/search/thesaurus.htm

Unified Medical Language System (UMLS) Metathesaurus http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Workshop of electronic thesauri: planning for a standard http://www.niso.org/thesau99.html