Dagobert Soergel
College of Information Studies
University of Maryland
College Park, MD 20742-4345
Office:(301) 405-2037 Fax (301) 314-9145
ds52@umail.umd.edu www.clis.umd.edu/faculty/soergel/

Thesauri and ontologies in digital libraries

Tutorial

Part 1: Structure and use in knowledge-based assistance to users

Part 2: Design, evaluation, and development

Joint Conference on Digital Libraries (JCDL 2002)
Portland, OR, USA
July 14, 2002

Abstract

This introductory workshop is intended for anyone concerned with subject access to digital libraries. It provides a bridge by presenting methods of subject access as treated in an information studies program for those coming to digital libraries from other fields. It will elucidate through examples the conceptual and vocabulary problems users face when searching digital libraries. It will then show how a well-structured thesaurus can be used as the knowledge base for an interface that can assist users with search topic clarification (for example through browsing well-structured hierarchies and guided facet analysis) and with finding good search terms (through query term mapping and query term expansion — synonyms and hierarchic inclusion). It will touch on cross-database and cross-language searching as natural extensions of these functions. The workshop will cover the thesaurus structure needed to support these functions: Concept-term relationships for vocabulary control and synonym expansion, conceptual structure (semantic analysis, facets, and hierarchy) for topic clarification and hierarchic query term expansion). It will introduce a few sample thesauri to illustrate these principles. Lastly the workshop will give a checklist for evaluating thesauri.

Course objectives

Participants should appreciate the complexity of subject access and understand the problems that a thesaurus can help solve.

Participants should understand the principles of thesaurus structure.

Participants should be able to apply thesaurus structure to solving subject access problems.

Participate should be able to identify and evaluate thesauri suitable for a specific situation defined by a user community an by the collection of a digital library.

Brief biography of the Instructor

Dagobert Soergel holds an MS equivalent in mathematics and physics (1964) and a PhD in political science (1970), both from the University of Freiburg, Germany. He is Professor of Information Studies, University of Maryland, where he teaches courses in information retrieval, thesaurus development, expert systems, and information technology, and an information systems consultant. He has been a visiting professor at the universities of Western Ontario, Chicago, and Konstanz, Germany. Among other books, he has authored *Organizing Information* (1985), which received the American Society of Information Science Best Book Award, *Indexing Languages and Thesauri. Construction and Maintenance* (1974) and numerous papers. He has designed several thesauri, most recently the Alcohol and Other Drug Thesaurus (for which he chairs the advisory committee) and is developing TermMaster, a thesaurus management software package. In 1997 he received the American Society of Information Science Award of Merit.

Part 1. Outline

Part 2. Outline

Challenges for digital libraries

Improve retrieval effectiveness to handle the sheer mass of material

Provide unified access to materials in different media (esp. access to non-text materials)

Provide knowledge-based support for end users who access networked information without the benefit of an intermediary

Support creation and maintenance of personal or work-group information systems

Support information seeking as an integral part of problem solving, learning, and intellectual work

Support collaborative work:
Scholarly communication as computer-supported multi-party conversation

Thesauri, ontologies, taxonomies, ... must support these functions

Support information seeking as an integral part of problem solving, learning, and intellectual work

Help users to explore ideas in conjunction with exploring information

Support fine-grained retrieval and assimilation of information

Support processing of information along with or after retrieval

Support collaborative work

Make users full participants in the continuing improvement of information systems through feedback and other contributions

Establish linkages between people

Why thesauri:
A first look with examples

Problems

Vocabulary confusion

User orientation in a concept space

Queries illustrating these problems

Queries:

Synonym expansion and Hierarchic expansion

Query 1. Drug use by teenagers

Query 1.1 teenage* AND drug*

Query 1.2 Synonym expansion for teenage*

(teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND drug*

Query 1.3 In addition, synonym expansion and hierarchic expansion for drug*

(teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND (drug* OR substance* OR alcohol OR nicotine OR smoking OR cigarette* OR mari*una OR cocaine OR crack OR heroin)

Query 1.4 Query more narrowly focused

(teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school") AND (cocaine OR crack OR heroin)

Query 1.1. teenage* AND drug* (AltaVista)

- -.
About 30 documents match your query.

1. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS
What do parents want from their teenagers? Basically, parents want: To know your kids are alright and not in danger. To know your kids think you're OK...
http://www. ceida. net. au/info_centre/drug~myths/what_do. html - size 3K - 21-May-97 - English

2. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS
Better Ways of Communicating. Different points of view Communication is the key to resolving problems, if they exist. Or to finding out if they exist....
http1A~www. ceida. net. au/info_centre/drug~myths/better.html - size 9K - 21-May-97 - English

3. Testimony of Donna E. Shalala, Secretary of HHS on Teenage Drug Use
Testimony of Donna E. Shalala, Secretary of Health and Human Services on Teenage Drug Use. Testimony of. Donna E. Shalala. Secretary of Health and Human...
http://www.apa.org/ppo/shalala.html - size 15K - 13-Sep-96 - English

4. Statement of Senator Richard C. Shelby on Teenage Drug Use
Statement of Senator Richard C. Shelby on Teenage Drug Use. Statement of. U.S. Senator Richard C. Shelby. Before The. Senate Judiciary Committee Hearing..
http://www.apa.org/ppo/shelbyhtml - size 3K - 13-Sep-96 - English

5. Testimony of John P. Walters on Teenage Drug Use
Testimony of John P. Walters, President of The New Citizenship Project, on Teenage Drug Use. Testimony by. John P. Walters* President of the New...
http://www.apa.org/ppo/walters.html - size 28K - 13-Sep-96 - English

6. Drug Use Rises for Teenagers
Parent News for November 1996. Of Interest. Drug Use Rises for Teenagers. by Anne S. Robertson. A recent report released by the Parents Resource ...
http://ericps.ed.uiuc.edu/npin/pnews/pnewn96/pnewn96f.html - size 4K - 23-May-97 - English

7. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS

Query 1.2. Synonym expansion of teenager

(teenage* OR teen OR teens OR youth OR adolescent* OR kid* OR "high school")
AND drug*

About 249 documents match your query.

1. Adolescent Drug Abuse Treatment Outcome
Adolescent Drug Abuse Treatment Outcome. Executive Summary. This is a report on the evaluation of an inpatient adolescent drug abuse treatment program in..
http://www. cbc. med. umn. edu/~andy/drugabuse/adoltx. htm - size 3K - 28-Sep-96 - English

2. Poll finds parents overestimate communication with kids on drugs
03/03/97 - 07:26 PM ET - Click reload often for latest version. Poll finds parents overestimate communication with kids on drugs. NEW YORK - Most parents..
http://cgi.usatoday.com/elect/eq/eq17&htm - size 2K - 21-May-97 - English

3. Albany Youth Futures shows kids alternatives to drugs, alcohol/TITLE>
http://www *indreg.com/9-11-96/FEATURES/feature5.htm - size 5K - 13-Sep-96 - English*

4. IPRC Version - Keeping Youth Drug-Free - Exercise #3
Re-posted by the Indiana Prevention Resource Center at Indiana University Indiana's RADAR Network State Center. Exercise 3. Building Social Skills. Offer..
http://www. drugs. indiana. edu/pubs/radar/keeping/exer3. html - size 2K - 28-Jun-96 - English

5. Online NewsHour: Teen Drug Use Doubling -- August 20, 1996
THEY'RE NOT SAYING "NO" AUGUST 20, 1996. TRANSCRIPT. Two new and deeply troubling reports have just been released showing that drug abuse among 12 to 17...
http://web-crOl.pbs. org/newshour/bb/health/august96/teen_dru g~ab use_8-20. html - size 16K -10-Sep-96- English

6. Kmart: HOTNEWS/Kmart Kids Race Against Drugs Race Results
Kmart Kids Race Against Drugs. And the winner is... On Saturday. January 18. Jamie Barreiro of Port St. Lucie, FL, Joshua Brown of Willingboro, NJ and ...
http://www.kmart.com/hotnews/hotnews.stm. size 7K - 21-May-97 - English

11. OMH-RC Database Record: Drug Abuse Among Minority Youth: Methodological Issues Office of Minority Health Resource Center Database Record. When available, information on where these materials may be obtained has been listed below...
http://wwwomhrc.gov/mhr2/docs/95D2315.htm - size 3K - 1-May-97 - English

Query 1.3. Plus synonym and hierarchic expansion of "drug*"

(teenage* OR teen OR teens OR youth* OR adolescent* OR kid* OR "high school")
AND (drug* OR substance* OR alcohol OR nicotine OR smoking OR cigarette*)
About 409 documents match your query.

1. Smoking is NOT for kids!
We believe smoking is for adults only. We therefore require that you be at least 18 years of age in order to view this site. Click below to enter the...
http://www.smokers.org/ - size 820 bytes - 20-Apr-97 - English

2. Adolescent Drug Abuse Treatment Outcome
Adolescent Drug Abuse Treatment Outcome. Executive Summary. This is a report on the evaluation of an inpatient adolescent drug abuse treatment program in..
http://www cbc. med. umn. edu/~andy/drugabuse/adoltx. htm - size 3K - 28-Sep-96 - English

3. Poll finds parents overestimate communication with kids on drugs
03/03/97 - 07:26 PM ET - Click reload often for latest version. Poll finds parents overestimate communication with kids on drugs. NEW YORK - Most parents..
http://cgi.usatodaycomielect/eq/eq]7&htm - size 2K- 21-May-97 - English

4. Albany Youth Futures shows kids alternatives to drugs, alcohol/TITLE>
http..//www. indregcoml9-11-96/FEATURESfeature5.htm - size 5K - 13-Sep-96 - English

5. IPRC Version - Keeping Youth Drug-Free - Exercise #3
Re-posted by the Indiana Prevention Resource Center at Indiana University Indiana's RADAR Network State Center. Exercise 3. Building Social Skills. Offer..
http://www. drugs. indiana. edu/pubs/radar/keeping/exer3. html - size 2K - 28-Jun-96 - English

6. Smoking still increasing among teens
Despite a chorus of ignorance one woman wanted to dance... To all of those people who say that national role models are a thing of the past, I want to...
http://www.bascchusgamma.org/bb_october/staff_view. html - size 5K - 11-Oct-96 - English

7. Online NewsHour: Teen Drug Use Doubling -- August 20, 1996
THEY'RE NOT SAYING "NO" AUGUST 20, 1996. TRANSCRIPT. Two new and deeply troubling reports have just been released showing that drug abuse among 12 to 17...
http://web-cr0l.pbs. org/newshour/bb/health/august96/teen_drug_abuse_8-20. html - size 16K -10-Sep-96- English

8. KCEOC SUBSTANCE ABUSE/YOUTH PROGRAM
KCEOC SUBSTANCE ABUSE/YOUTH PROGRAM. Address: 1611 First Street. Phone Number: *336-5310.* FAX Number: *336-5303.* Contact Person: Robert Cubit. Target Group..
http://www bakersfield. org/ydc/secondary/kceoc. html - size 2K - 15-Oct-96 - English

9. Kmart: HOTNEWS/Kmart Kids Race Against Drugs Race Results
Kmart Kids Race Against Drugs. And the winner is... On Saturday, January 18, Jamie Barreiro of Port St. Lucie, FL, Joshua Brown of Willingboro, NJ and...
http://wwwkmart.coiri/hotnews/hotnews.stm - size 7K - 21-May-97 - English

10. Connecticut Kidslink - Substance Abusing Mothers and Their Children
Inter-agency Committee on Substance Abusing Mothers and Their Children in Connecticut: A Summary of Problems and Solutions. Report Summary by Andy Dodge,..
http://statlab.stat.yale. edu/cityroom/kidslink2/welffire/texts/9603-03. html - size 9K - 7-Nov-96 -English

11. OMH-RC Database Record: Drug Abuse Among Minority Youth: Methodological Issues Office of Minority Health Resource Center Database Record. When available, information on where these materials may be obtained has been listed below...
http://wwwomhrc.gov/mhr2/docs/95D2315.htm - size 3K - 1-May-97 - English

12. Browne for President - Release - teenage smoking
NEWS FROM THE BROWNE FOR PRESIDENT CAMPAIGN. FOR IMMEDIATE RELEASE August 23, 1996. Clinton's new "War On Teenage Smoking" is moral grandstanding, charges.
http://www harrybrowne96. org/release-teenage-smoking html - size 4K - 24-Aug-96 - English

13. Teacher Talk, 3(3), Alcohol and Adolescents
Alcohol and Adolescents: Prevention, Intervention, Treatment, Aftercare Volume 3, Issue 3 A Publication Just for Secondary Teachers. 1996 Indiana...
http://education. indiana. edu/cas/tt/v3i3/v3i3toctext. html - size 2K - 6-Jun-96 - English

14. White House Conference on Youth Drug Use
White House Conference on Youth Drug Use. (from March/April 1996 Marijuana Policy Report) In a further attempt to defuse criticism of being soft on drugs,.
http://www.mpp.org/youfhconfhtml - size 2K - 21-May-97 - English

15. Anti-Smoking Software Installed at Bronx High School of Science
Anti-Smoking Software Installed at Bronx High School of Science. March 6, 1997: The Alumni Association of the prestigious Bronx High School of Science has.
http://www.smokefreekids.com/relO2.htm - size 2K - 21-May-97 - English

16. Optum: Live Event! Talking to Kids about Alcohol and Drugs
Optum: What is happening This Month at Optum? Check here and find out.

Query 1.4. Drug component more specific

(teenage* OR teen OR teens OR youth OR adolescent* OR kid* OR "high school")
AND (cocaine OR crack OR heroin)

2 documents match your query

1. Teenage "Huffing" - Worse Than Cocaine
Teenage “Huffing” - Worse Than Cocaine. May 22, 1996. MEEUWSEN: Imagine substances experts call deadlier than heroin or cocaine. Imagine that...
http://www.cbn.org/news/stories/huffinghtml - size 7K - 29-Oct-96 - English

2. Teen is arrested with a kil of crack cocaine
Teen is arrested with a kilo of crack cocaine. STROUDSBURG, Pa. (AP) - A 14-year-old New York City girl was busted during a bus trip through here last...
http..//www. recordernews. com/1 996/0703/natnews/teenare/teenare. html - size 2K - 25-May-97 English

Queries: Homonyms and polysemes

Query 2. wordnet (homonym: 6 meanings)

Query 3. classification (polyseme)

Query 3.1. classification AND security

Query 2. wordnet (homonym: 6 meanings)

3. **WordNet: A Lexical Database for English**
Lexical Resources for Human Language Technology. Princeton University. DARPA/ITO
http://www.ito.darpa.mil/Summaries95/B370--Princeton.html - size 12K - 12-Sep-96 -

4. **VDI - Racal WordNet Networks**
Racal Recorders. The WordNet Uses The TCP/IP "internet" Protocol, Allowing Easy Network Utilization. Search And Playback. Recorded Messages Via The Lan....
http://www.fishnet.net/~ecs/racal3.htm - size 539 bytes - 11-Oct-96 - English

6. **WordNet lexical database**
http://www.grafnetix.com/thesaurus/QueryExpansionIntro/node1.html - size 6K -

8. **WORDNET, the new generation of digital communications recorders**
Digital communications logging recorder.
http://www.abds.net/dss/wordnet.htm - size 2K - 30-Jan-97 - English

13. **WordNet's Christian Links**
Christian Web Sites. Below is your passport to a wider Christian on-line community. Some contain links to many other Christian sites
http://www.wordnet.co.uk/links.html - size 3K - 23-May-97 - English

18. **The Wordnet Story**
Wordnet Productions. Jesus, the Divine Word, casts his net, the Good News, to all through mass media. Wordnet is a Catholic television ministry dedicated..
http://www.rlagroup.com/wordnet/wrdntstr.htm - size 2K - 6-Feb-97 - English

30. **Tesi di Laurea WordNet**
Linguaggio Naturale. Proposta per Tesi di Laurea: WordNet. WordNet e` una base di conoscenza lessicale per l'inglese, disponibilile gratuitamente su..
http://ecate.itc.it:1024/cirave/wordnet.html - size 2K - 30-Sep-96 - Italian

48. **WORDNET Language Translation Service**
WORDNET is a team of language experts specializing in foreign language translation, typesetting and printing. In recent years, we have helped a number of..
http://www.wordnet.com/ - size 4K - 20-Jun-97

52. **Consortium of the EuroWordNet project**
All Rights reserved by Computer Centrum Letteren University of Amsterdam. Coordinator: builder of...
http://www.let.uva.nl/~ewn/consortium-ewn.html - size 3K - 22-Apr-97 - English

Query 3. classification (polyseme)

Examples from AltaVista search

1. **GNWT Administrative Records Classification System BUILDINGS AND PROPERTIES**

BUILDINGS AND PROPERTIES - DAMAGES 2063. Records relating to damages incurred by government buildings, facilities and structures. It includes...
http://pingo.gov.nt.ca/Records/sections/2000/1995blg9.htm - size 4K - 17-Oct-96 - English

2. **LC Classification: U - Military Science**
U - Military Science. U. 1-900. Military Science (General). 21-22.3. War, Philosophy, Military Sociology. 27-43. History of Military Science. 164-167.5....
http://www.library.yorku.ca/lc/u.html - size 6K - 13-Nov-96 - English

7. **Table Tennis Classification Procedures**
International Paralympic Committee. Sports Science | Medical | Sports | Secretariat | General. Table Tennis Classification Procedures. A. Purposes. 1) To..
http://info.lboro.ac.uk/research/paad/ipc/table-tennis/class-proc.html - size 7K - 2-Jul-96 -

8. **MPW Public Highways (Road Classification)**
ROAD CLASSIFICATION. Law No. 13 of the year 1980 (UU 13/1980) concerning roads distinguishes the category of road into public and special roads. The...
http://www.pu.go.id/publik/binama~1/html/eng/classifi.htm - size 3K - 22-May-96 -

9. **Hurricane and Tropical Storm Classification**
http://www.hiwaay.net/cwbol/scale.html - size 3K - 7-Jul-95 - English

17. **DEPARTMENT OF ENERGY FUNDAMENTAL CLASSIFICATION POLICY REVIEW**
Secretary Hazel O'Leary has emphasized the importance of improved public accountability
http://www.osti.gov/html/osti/opennet/fcprsum.html - size 10K - 11-Feb-97 - English

29. **Subject guide to the classification**
Subject guide to the Library of Congress classification. For subjects not listed here please consult the printed, red-bound Subject Index in the entrance..
http://potter.cc.keele.ac.uk/depts/li/lctable.htm - size 7K - 21-May-97 - English

30. **BRYOPHYTES: Hornwort Classification**
Phylum ANTHOCEROTOPHYTA. DENDROCEROS. Gametophyte plant with horn-like sporophyte. copyright ©1996 Southern Illinois...
http://www.science.siu.edu/bryophytes/anthocerotophyta.html - size 940 bytes - 5-Apr-97 -

31. **Policy & Planning Support - Staff Level Classification**
Staff Classification & Level. All staff are assigned a classification on employment. This data element indicates the classification...
http://wwwplan.murdoch.edu.au/stats/descript/clssfctn.html-ssi - size 4K - 21-May-97

34. **Classification Reform Approval**
March 5, 1996. FOR IMMEDIATE RELEASE. Release No. 14. POSTAL SERVICE APPROVES CLASSIFICATION REFORM RECOMMENDATIONS
http://www.usps.gov/news/press/96/96014new.htm - size 4K - 12-Apr-97 - English

57. **Universal Decimal Classification Index 5414**
NATURAL SCIENCES. MATHEMATICS. 54 CHEMISTRY. MINERALOGY. 541 GENERAL, THEORETICAL, AND PHYSICAL CHEMISTRY. 5414 CHEMICAL.
http://www.chem.ualberta.ca/~plambeck/udc/u5414.htm - size 827 bytes - 9-May-97 -

61. **Draft Public Guidelines to Department of Energy Classification of Information**
http://www.osti.gov/html/osti/opennet/document/guidline/pubgf.html - size 17K -

71. **The GNU C Library - Classification of Characters**
This section explains the library functions for classifying characters.
http://www.ia.pw.edu.pl/Pl-iso/tex-info/libc/libc_55.html - size 7K - 6-Apr-94 - English

80. **Dewey Decimal Classification System**
Dewey Decimal Classification System. Overview. 000 Generalities 100 Philosophy & psychology 200 Religion 300 Social sciences 400 Language 500 Natural...
http://www.tnrdlib.bc.ca/dewey.html - size 38K - 7-Aug-96 - English

88. **Extended Computing Reviews Classification Scheme**
Extended Computing Reviews Classification Scheme. Computing Reviews Classification System. Copyright 1994, by the Association for Computing Machinery,...
http://www.dpmms.cam.ac.uk/MR/CRclass.html - size 37K - 1-Sep-95 - English

89. **627.440 - Classification of costs.**
627.440 - Classification of costs. Standard Number: 627.440. Standard Title: Classification of costs. SubPart Number: D. SubPart Title: Administrative...
http://www.doleta.gov/regs/cfr/20cfr/toc_Part600-699/0627.0440.htm - size 12K -

90. **Pirelli Cumbria Rally 1996 Final Classification**
Pirelli Cumbria Rally 1996. Final Classification. POSITION OVERALL CLASS NUMBER CREW CLASS TOTAL 11 201 Richard Tuthill/Nick Kennedy Vauxhall Nova 1300...
http://www.idiscover.co.uk/tcs21/1996/pirelli/c_class.html - size 2K - 5-May-96 - English

117. **Classification of Students**

Classification of Students. Students at Bemidji State University are classified as regular, special, or auditor. Regular: A regular student is one who is..
http://bsuweb.bemidji.msus.edu/~catalog/catalog94_96/classify_stud.html - size 1K - 15-Mar-95 - English

Query 3. classification

Examples from Lycos search

2) **Classification of Signatures**
http://www.seas.gwu.edu/faculty/pbock/SignatureCla [99%]

5) **Supervised Classification**
Neural Network Classification of Multispectral Imagery Supervised Classificati .
http://www.ece.arizona.edu/~paola/SupervisedClass. [99%]

139) **RESIDENCE CLASSIFICATION**
Residence Classification Residence Classification Nonresident students seeking to become California residents for tuition/fee purposes must petition t.
http://www.reg.uci.edu/REGISTRAR/SOC/rc.html [99%]

152) **PRODUCT CLASSIFICATION**
EPA may classify a pesticide product for restricted use if its characteristics warrant special handling. Restricted use pestici.
http://hammock.ifas.ufl.edu/txt/fairs/26668 [99%]

426) **Dewey Decimal Classification Home Page**
DDC 21 and Dewey for Windows now available! OCLC Forest Press is pleased to announce the publication of DDC 21, the latest edition of the Dewey Decima.
http://www.oclc.org/fp/ [99%]

429) **Dewey Decimal Classification Web Site**
The Dewey Decimal Classification: Numbers You Can Count On catalog is now available. Use the online form to have.
http://www.oclc.org/oclc/fp/fptxthm.htm [99%]

634) **Library of Congress Classification System Introduction**
Introduction to the LC Classification System Some say Information is Power. Others say Information is the door to Knowledge. Libraries hold the key to.
http://snoopy.tblc.lib.fl.us/laudem/Introduction.h [99%]

Query 3.1. classification and security

Examples from AltaVista search

Restricts results but also misses a lot.

1. **EXSYS: Specific Applications: Security Classification**
   Nuclear Weapons Security Classification. US Dept. of Energy. Nuclear...
   http://www.exsysinfo.com/Appnotes/nuclear.html - size 7K - 22-May-97 - English

2. **SLATE Application Note --Security Classification and Automatic Page Marking wi**
   Introduction. If your document contains classified information, you can identify the classification by.
   http://www.slate.tdtech.com/app_notes/secclass-html.html - size 6K - 22-Feb-96 - English

3. **Computer Security Classification**
   The Classification. alert Advisories on various security vulnerabilities. dict Dictionaries and word lists. doc Security related documents.
   access_control.
   http://www.cs.purdue.edu/coast/archive/Classification.html - size 8K - 17-Mar-95 - English

4. 355 Security Classification Control (R)
   Top] -- MARC Field Guides Table of Contents -- 300 - Physical Description Fields. 355 Security Classification Control (R)Contains
   specifics pertaining to..
   http://infoshare1.princeton.edu/katmandu/marc/355.html - size 3K - 20-Jan-97 - English

5. **Security and Classification**
   By John Pike (johnpike@clark.net) The classification system is designed primarily to protect the confidentiality of certain...
   http://www.tscm.com/classification.html - size 17K - 28-Dec-96 - English
   http://www.awpi.com/IntelWeb/US/misc/classification.html - size 16K - 15-May-96 -

6. **National Security Classification Cost Estimates**
   A report to Congress from the Information Security Oversight Office
   http://vwww.clark.net/fas/sgp/isoo/costs97.html - size 9K - 10-May-97 - English

What is a thesaurus? A first look

A **dictionary** is a listing of words and phrases giving information such as spelling, morphology and part of speech, senses, definitions, usage, origin, and equivalents in other languages (bi- or multilingual dictionary).

A **thesaurus** is a structure that
manages the complexities of terminology and
provides conceptual relationships,
ideally through an embedded classification/ontology.

A thesaurus may specify descriptors authorized for indexing and searching. These descriptors form a **controlled vocabulary** (**authority list**, **index language**).

A **monolingual thesaurus** has terms from one language, a **multilingual thesaurus** from two or more languages.

A **classification** is a structure that organizes concepts into a hierarchy, possibly in a scheme of facets.
The term **ontology** is often used for a
shallow classification of basic categories or a
classification used in linguistics, data element definition, or knowledge management or (increasingly) for
any classification.

AOD navigation page here

EF route of administration

EF2 — by scope of drug action
EF2.2 . topical and local administration
EF2.2.2 . . topical administration
EF2.2.4 . . local drug administration
EF2.4 . systemic administration

EF4 — by method or body site
EF4.2 . enteral administration
EF4.2.2 . . oral enteral administration
EF4.2.4 . . rectal enteral administration
EF4.4 . mucosal administration
EF4.4.2 . . transdermal administration
EF4.4.4 . . inhalation, smoking, sniffing
EF4.4.4.2 . . . smoking
EF4.4.4.2.2 . . . . smoking w/out inhalation
EF4.4.4.2.4 . . . . smoking with inhalation
EF4.4.4.4 . . . nasal administration
EF4.4.4.6 . . . pulmonary administration
EF4.4.6 . . oral mucosal administration
EF4.4.6.2 . . . buccal administration
EF4.4.6.4 . . . sublingual administration
EF4.4.8 . . rectal mucosal administration
EF4.6 . parenteral administration
EF4.6.2 . . intravenous injection
EF4.6.2.2 . . . intravenous infusion
EF4.6.4 . . intra-arterial injection
EF4.6.6 . . intraperitoneal administration
EF4.6.8 . . intracutaneous injection
EF4.6.10 . . administration through skin implant
EF4.6.12 . . subcutaneous injection
EF4.6.14 . . intramuscular injection
EF4.6.16 . . CNS injection
EF4.6.16.2 . . . intrathecal injection
EF4.8 . skin administration
(The full entry shows Narrower Term cross-references to the more specific methods involving the skin: EF4.4.2, EF4.6.8, EF4.6.10, and EF4.6.12)
EF4.10 . oral administration
(NT to EF4.2.2, EF4.4.4.2, and EF4.4.6)
EF4.10 . rectal administration
(NT to EF4.2.4 and EF4.4.8)

EF6 drug administration by self vs. others
EF6.2 . self administration of drugs
EF6.4 . drug administration by others

Excerpt from a thesaurus hierarchy

EF route of administration

SN The way in which a substance reaches its site of action in the body. The substance may be administered for therapeutic or psychoactive effects - possibly as part of a human or animal experiment - by a third party or by the subjects themselves, or the subject may be exposed to the substance through the environment or in utero.

The major distinction between routes of administration is not the site where a substance is introduced or applied to the body, or even the way it is introduced or applied, but whether it takes effect merely in the local area where it is applied or whether it reaches its destination through systemic circulation. A further criterion is whether the drug reaches systemic circulation directly or whether it first passes through the liver, where it may be metabolized or excreted (first-pass effect in enteral administration). (Note: Drugs administered into the systemic circulation by any route, excluding intra-arterial injection, are subject to possible first-pass elimination in the lung prior to distribution to the rest of the body.)

Whether administration of a drug results in local or systemic action depends not only on the site and method of administration but also on the properties of the drug; sometimes the drug has both local and systemic action. This is particularly true for application to a mucous membrane, which may be intended for a local action but also may have - sometimes unwanted - systemic action. Furthermore, a drug may be absorbed at several sites (e.g., the mouth and the lung, the rectum and the intestine) in various proportions. To account at least partially for the very complex phenomena of the absorption of drugs into the body, the following classification uses two dimensions, or facets: By scope of drug action and by method or body site of administration. To index route of administration completely, use at least one descriptor from each facet.

ST *medication route*
ST *method of delivery of drugs or food*
ST *mode of substance administration*
ST *route of drug application*
ST *route of drug entry*
ST *route of exposure*
BT +EE12 pharmacokinetics
RT +AA2 AOD use
RT +BS AOD substance by route of administration
RT EE12.2e drug absorption
RT +EE14.4.8 drug effect by location
RT +HR drug therapy
RT MD2.2.2.2 drug paraphernalia

EF2 route of administration by scope of drug action

SN Use one of these descriptors in combination with a descriptor from **+EF4 route of administration by method or body site**.

EF2.2 . topical and local administration
SN The application of a substance to a localized area, chiefly for local effects at this site.
NT HU4.2 local anesthesia
RT GH10.2 chemical injury

EF2.2.2 . . topical administration
SN The application of a substance on the surface of the skin or on a mucous membrane (incl. the gastrointestinal membrane) so that the substance will take effect on the surface or on a localized layer under the surface.
For example, for the administration of a decongestant spray, use **EF2.2.2 topical administration** combined with **EF4.4.4.4 nasal administration**.
ST *topical application*

EF2.2.4 . . local drug administration
SN The introduction of a substance into a localized area of the skin or other tissue, as through injection.
NT EF4.6.4 intra-arterial injection
NT EF4.6.8 intracutaneous injection
NT +EF4.6.16 CNS injection

EF2.4 . systemic administration
SN The introduction of a substance into systemic circulation so that it is carried to the site of effect.
NT +EF4.6.2e intravenous injection
NT EF4.6.10 administration through skin implant
NT HU4.4 general anesthesia
RT +GH10.4 chemical poisoning

Examples of full thesaurus entries

Multilingual thesaurus problems

simian	**Affe**
monkey	*niederer Affe*
ape	Menschenaffe
timepiece	**Uhr**
clock	*Wanduhr, Standuhr, Turmuhr*
wall clock	Wanduhr
standing clock	Standuhr
tower clock	Turmuhr
watch	*Taschenuhr, Armbanduhr*
pocket watch	Taschenuhr
wrist watch	Armbanduhr
alarm clock	Wecker
blanket, rug, carpet	**Teppich**
blanket	Betteppich
rug, carpet	Bodenteppich
rug (or carpet)	*loser Bodenteppich*
long, narrow rug	Läufer
(wall-to-wall) carpet	Teppichfußboden
hanging rug	Wandteppich

Italics denotes terms created to express a concept not lexicalized in English or German, respectively.

Note that most English-German dictionaries would have you believe that the German equivalent for "monkey" is "Affe", but that equivalence holds only in some contexts.

Another difficulty arises when two terms mean almost the same thing but differ slightly in meaning or connotation, such as *alcoholism* in English and *alcoholisme* in French, or *vegetable* in English (which includes potatoes) and *Gemüse* in German, which does not. If the difference is big enough, one needs to introduce two separate concepts under a broader term; otherwise a scope note needs to clearly instruct indexers in all languages how the term is to be used so that the indexing stays, as far as possible, free from cultural bias or reflects multiple biases by assigning several descriptors.

Examples of classifications and thesauri

Alcohol and Other Drug Thesaurus (AOD Thesaurus)
(US Nat. Inst. of Alcohol Abuse and Alcoholism)
http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

Medical Subject Headings (MeSH) and Unified Medical Language System (UMLS)
(US National Library of Medicine)
www.nlm.nih.gov/mesh/meshhome.html, www.nlm.nih.gov/mesh/MBrowser.html,
www.nlm.nih.gov/research/umls/umlsmain.html, http://umlsinfo.nlm.nih.gov

Art and Architecture Thesaurus (AAT)
(Getty Foundation)
http://www.getty.edu/research/tools/vocabulary/aat/index.html

Dewey Decimal Classification
(US Library of Congress and OCLC/Forest Press)
http://www.oclc.org/dewey/about/ddc_21_summaries.htm

WordNet (Princeton University, George Miller)
www.cogsci.princeton.edu/~wn/,
www.notredame.ac.jp/cgi-bin/wn (Not reachable on July 6, 2002)

CYC Ontology (CYC Corporation)
http://www.cyc.com/cyc-2-1/cover.html, http://www.cyc.com/cyc-2-1/toc.html

Example pages form part 2 of the tutorial materials. They will be examined briefly but are intended primarily for further study.

More thesaurus examples

A few sample pages included

Yahoo **The Yahoo classification**. Web pages www.yahoo.com

Bloom **Taxonomy of educational objectives** 1956 (1 copy in the cataloging laboratory) (LB17.B55.1956), a summary at
http://www.unesco.org/webworld/ramp/html/r8810e/r8810e0e.htm
http://websites.ntl.com/~james.atherton/learning/bloomtax.htm,
http://sweep.riv.csu.edu.au/td/bloom.html,
http://faculty.washington.edu/~krumme/guides/bloom.html

SOC Standard Occupational Classification 2000
Bureau of Labor Statistics (BLS) + other agencies
http://stats.bls.gov/soc/soc_home.htm
The SOC is augmented by the **Occupational Information Network (O*NET)**, a database with additional occupational titles, definitions, and features of occupations.
http://www.doleta.gov/programs/onet

CSDGM **Content Standard for Digital Geospatial Metadata** 1998
Federal Geographic Data Committee (FGDC)
http://www.fgdc.gov/metadata/contstan.html

ERIC **Education Resources Information Center Thesaurus**. 13th ed. Bibliographic retr. http://searcheric.org/

Additional examples illustrating different functions

HS **Harmonized Commodity Description and Coding System**. World Customs Organization, Brussels. Info: http://pacific.commerce.ubc.ca/trade/HS.html

NAICS **North American Industrial Classification System**
"common industry definitions for Canada, Mexico, and the US. Developed in cooperation with the US Economic Classification Policy Committee, Statistics Canada, and Mexico's Instituto Nacional de Estadistica, Geografia e Informatica to better compare economic and financial statistics and ensure that such statistics keep pace with the changing economy. NAICS will replace the countries' separate classification systems (in the US: Standard Industrial Classification, SIC) with one uniform system for classifying industries."
Info: www.census.gov/epcd/www/naics.html, www.naics.com

ICD-10 **The International Statistical Classification of Diseases and Related Health Problems, tenth revision**. Produced by the World Health Organization. Published in many languages. Info: www.who.int/whosis/icd10/index.html, www.cdc.gov/nchs/about/major/dvs/icd10des.htm

CPT **Physicians' Current Procedural Terminology. CPT 2003**. American Medical Association. November 2002
(Info: http://www.ama-assn.org/ama/pub/category/3113.html, listing of codes https://webstore.ama-assn.org/index.jhtml)
Health Care Finance Administration (HCFA) Common Procedure Coding System (HCPCS) for Medicare reimbursement for hospital outpatient services. It has three levels - CPT (level 1), HCPCS or National (level 2), and Local (level 3).
In its data collection the Agency for Health Care Policy and Research (AHCPR) uses data standards that are based on those employed by the Census Bureau, the American Hospital Association, the Health Resources and Services Administration (Area Resource File), the National Center for Health Statistics, and codes for clinical diagnosis and procedures such as ICD-10 and CPT 1998. These standards facilitate data analysis and use by ensuring comparability, quality and interoperability. Further, uniform health care data advance medical and health care services research, the efficiency of the private sector health care delivery system, and quality improvement measurement.

- Further type of classification: **biological taxonomies**. Used in biology, agriculture, food science, and medicine. Several rivaling schemes for major areas (kingdoms) and many publications on specific areas.
  http://www.itis.usda.gov/
  http://www.ucmp.berkeley.edu/help/taxaform.html

- **Metadata schemas (such as CSDGM), data element dictionaries, object hierarchies in object-oriented programming**

Functions of a thesaurus / classification / ontological knowledge base

in the context of digital libraries

Support learning and assimilating information.

Assist researchers and practitioners with problem clarification.

Support information retrieval.

- Provide knowledge-based support for end-user searching.
- Support meaningful information display.
- Provide a tool for indexing.
- Facilitate the combination of multiple databases or unified access to multiple databases.

Support document processing after retrieval.

Support learning and assimilating information

Support learning about any topic **by providing** the learner with a **coherent, age-appropriate conceptual framework**.

Learning as information retrieval. Conceptual framework for asking the right questions.

Assist readers in understanding text.

Assist researchers and practitioners with problem clarification —

provide the conceptual basis for the design of good research and implementation and for good query formulation. Includes help with

exploring the conceptual context of a research or practical problem — a study, policy, plan, or implementation project

and with

structuring the problem.

Examples of specific functions:

Present the issues in a field or application area in a coherent framework.

Assist in problem-solving: Assist in the exploration of the dimensions of a problem and aspects to be considered in its solution; provide a classification of approaches to solving a specific problem.

Provide classification and **consistent definition of variables for research / of evaluation criteria for practical problems**, thus enhancing the comparability of research and evaluation results and making research more cumulative.

Support information retrieval

Provide **knowledge-based support for end-user searching**. Support

searching in multiple natural languages;

free-text searching;

searching multiple databases using different index languages.

Elicitation of user needs through a series of menus based on search tree, or through **guidance in the conceptual analysis of a search topic** (questions based on a facet structure, presentation of a segment of the concept hierarchy for each applicable facet).

Browsing the classification structure to identify useful concepts for a search at the level of specificity desired. Browsing a collection, as in a subject directory.

Mapping from the user's query terms to descriptors used in a database **or to the multiple natural language expressions** to be used for free-text searching.

Inclusive (hierarchically expanded) **searching**.

Enhanced ranking algorithms based on concept and term relationships.

Searching multiple databases by mapping the users query terms to the descriptors used in each of the databases, or mapping the descriptors from one database to another databases (switching); common search language.

Support information retrieval, continued

Support information display, especially presentation of search results:

Meaningful arrangement of units (document records, paragraphs, property data on a given substance assembled from several databases), including knowledge-based clustering of records retrieved.

This supports **exploration of large retrieved sets** and, by extension, **exploration of the content of an entire collection** or subcollection.

Meaningful arrangement of information within a record (for example meaningful ordering of descriptors assigned).

Support information retrieval, continued

Provide a tool for indexing.

Vocabulary control.

User-centered (request-oriented, problem-oriented) **indexing**.

Indexing **several databases** in a field with a **common index language** and sharing the results of indexing to reduce overall indexing effort.

Mapping indexing descriptors from one system to another.

Support information retrieval, continued

Facilitate the combination of multiple databases or unified access to multiple databases through

mapping the users query terms to the descriptors used in each of the databases;

mapping the query descriptors from one database to another (switching);

providing a **common search language** from which to map to multiple databases;

providing a **common index language** for a number of databases in a field;

mapping indexing descriptors from one database to another.

Support document processing after retrieval

For example

Highlight descriptors responsible for retrieval, using different colors for different facets.

Highlight terms belonging to a given category, for example, **personal names**, again using different colors for different categories.

Prepare document summaries, possibly in a different language , taking into account the query topic.

Translate full documents.

Extract facts from text. Compile and arrange facts extracted from several texts.

The underlying function of a knowledge base on concepts and terminology

Map out a concept space, relate concepts to terms, and provide definitions, thus providing orientation and serving as a reference tool.

Provide a **semantic road map and common language** for an individual field and, perhaps more importantly, map the relationships among fields.

Clarify concepts by putting them in the context of a classification / typology and to provide a system of definitions.

Relate concepts and terms across disciplines, languages, and cultures.

Thesaurus/ontology functions

Reference list

Functions of a thesaurus / classification / ontological knowledge base
Overview

Provide a **semantic road map** to individual fields and the relationships among fields.
Map out a concept space, relate concepts to terms, and provide definitions, thus providing orientation and serving as a reference tool.

Improve communication generally. Support learning and assimilating information.

Support learning through conceptual frameworks. Conceptual framework to help the learner ask the right questions.

Support the development of instructional materials through conceptual frameworks.

Assist readers in understanding text by giving the meaning of terms.

Assist writers in producing understandable text by suggesting good terms.

Support foreign language learning.

Provide the conceptual basis for the design of good research and implementation.

Assist researchers and practitioners with problem clarification.

Consistent data collection, **compilation of statistics** (related to information analysis)

Provide classification for action. Classification for social and political purposes

a classification of diseases for diagnosis,

of medical procedures for insurance billing,

of commodities for customs.

Support information retrieval and analysis. Organizing and keeping track of goods and services for commerce (esp. ecommerce) and inventory

Provide a tool for searching, particularly knowledge-based support for end-user searching, including hierarchically expanded searching.

Provide a tool for indexing.

Facilitate the combination of or unified access to multiple databases

Support document processing after retrieval.

Support meaningful, well-structured display of information.

Ontology for data element definition. Data element dictionary.

Conceptual basis for knowledge-based systems.

Do all this across multiple languages

Mono-, bi-, or multilingual dictionary for human use.
Dictionary/knowledge base for automated language processing

The underlying function of a knowledge base on concepts and terminology:

Provide a semantic road map to individual fields and the relationships among and across fields.

Map out a concept space, relate concepts to terms, and provide definitions, thus providing orientation and serving as a reference tool.

Provide a **semantic road map and common language** for an individual field and, perhaps more importantly, map the relationships among fields.

Clarify concepts by putting them in the context of a classification / typology and to provide a system of definitions.

Relate concepts and terms across disciplines, languages, and cultures.

Many specific functions build on this foundation.

Improve communication generally.
Support learning and assimilating information

Support learning about any topic **by providing** the learner/reader with a **coherent, age-appropriate conceptual framework**. Conceptual frameworks help the learner ask the right questions; **learning as information retrieval**.

Support the development of instructional materials by providing a conceptual framework to the instructional developer / writer and by suggesting didactically useful arrangements of topics.

Assist readers in understanding text; help them ascertain the proper meaning of a term and placing it in context.

Assist writers in producing understandable text by helping them to conceptualize the topic and suggesting from a semantic field the term that best conveys the intended meaning and connotation.

Support foreign language learning

Provide the conceptual basis for the design of good research and implementation.

Assist researchers and practitioners with problem clarification

Includes help with

exploring the conceptual context of a research or practical problem — a study, policy, plan, or implementation project

and with

structuring the problem and providing a conceptual framework for asking the right questions and devising good query formulations for retrieval.

Examples of specific functions:

Present the issues in a field or application area in a coherent framework.

Assist in problem-solving: Assist in the exploration of the dimensions of a problem and aspects to be considered in its solution; provide a classification of approaches to solving a specific problem (for example, a classification of approaches to drug abuse prevention as a help in designing drug abuse prevention projects).

Provide classification and **consistent definition of variables for research / of evaluation criteria for practical problems**, thus enhancing the comparability of research and evaluation results and making research more cumulative.

Support the compilation and use of statistics

This is a very important function. The Census Bureau, the Bureau of Labor Statistics, and other statistical agencies are heavily involved in developing classifications and defining concepts.

Support data collection

> The concepts in a classification used for statistics not only make the collected data retrievable, they define the very nature of the data.

Support data aggregation

> For example, get the value of all *electronic goods* imported into the US in the year 2000, or the tonnage of *green leafy vegetables* produced in a given year in the US.

Support retrieval of specific numbers (also part of information retrieval)

Support data tabulation and analysis (Need to have proper variables available)

Provide classification for action

This list addresses the functions of formal classifications. In a broader perspective, classification is the basis for much of everyday action, where we put people, things, and events in certain categories and, based on these categories, predict the behavior of persons and things and the course and effects of events, determine our attitudes towards them, and plan action accordingly.

For example,

a classification of diseases for diagnosis,

a classification of medical procedures for insurance billing,

a classification of medical outcomes to assist with treatment evaluation,

a classification of commodities for customs,

a classification of educational objectives for instructional development,

a classification of occupations for matching job applicants with job openings and for pay scale;

a classification of skills for employee task assignments.

a classification of crimes for determining sentences

a classification of types of expenses for tax purposes

Classification for social and political purposes. Socially charged classification

For example

Establishing that a profession has its own knowledge base, thereby enhancing the recognition of the profession (for example, the Nursing Intervention Classification)

Establishing a persons condition or behavior as normal, or as a disease, or as a moral failing or otherwise deviant. Different groups may want the same condition or behavior classified in different ways to further their agenda

Examples:

Should homosexuality be classified as a disease?

Is alcoholism or other drug abuse a disease or a moral failing?

Is mental illness a disease on a par with physical illness, and thus covered by health insurance the same way?

Is some levy to be classified as a *tax* or as a *user fee*

Support information retrieval 1:

A tool for searching, particularly knowledge-based support for end-user searching. Support

searching in any kind of database — bibliographic, full-text and hypermedia, directory, numeric, etc.;

searching in any kind of medium — printed indexes, CD-ROM systems, online systems, and the Internet;

searching in multiple natural languages independent of the language used in each database;

free-text searching;

searching multiple databases using different index languages.

Elicitation of user needs through a series of menus based on a search tree, or through **guidance in the conceptual analysis of a search topic** (questions based on a facet structure, presentation of a segment of the concept hierarchy for each applicable facet).

Browsing the classification structure to identify useful concepts for a search at the level of specificity desired. (The user may not have command of the vocabulary needed.) Browsing a collection (as on the shelves or in a subject directory)

Mapping from the user's query terms to descriptors used in a database **or to the multiple natural language expressions** to be used for free-text searching.

Inclusive (hierarchically expanded) **searching**.

Enhanced ranking algorithms that use concept and term relationships.

Searching multiple databases by mapping the users query terms to the descriptors used in each of the databases, or mapping the descriptors from one database to another databases (switching); common search language.

Support information retrieval 2: Provide a tool for indexing.

Vocabulary control.

User-centered (request-oriented, problem-oriented) **indexing**.

Indexing **several databases** in a field with a **common index language** and sharing the results of indexing to reduce overall indexing effort.

Mapping indexing descriptors from one system to another.

Support information retrieval 3:

Facilitate the combination of multiple databases or unified access to multiple databases through

mapping the users query terms to the descriptors used in each of the databases;

mapping the query descriptors from one database to another (switching);

providing a **common search language** from which to map to multiple databases;

providing a **common index language** for a number of databases in a field;

mapping indexing descriptors from one database to another.

Support information retrieval 4: Document processing after retrieval

Sample functions that require knowledge-based support:

Meaningful arrangement of search results (see next box)

Highlight descriptors responsible for retrieval, using colors to show facets.

Highlight terms belonging to a given category, for example, **personal names**, again using different colors for different categories.

Prepare document summaries, possibly in a different language , taking into account the query topic.

Translate full documents.

Extract substantive data from text. Compile and arrange data extracted from several texts.

Support meaningful, well-structured display of information

Meaningful arrangement of units (document records, paragraphs, property data on a given substance assembled from several databases), including knowledge-based clustering of records retrieved. This includes meaningful structure for **Web sites** and **subject directories**

This supports **exploration of large retrieved sets** and, by extension, **exploration of the content of an entire collection** or subcollection.

Meaningful arrangement of information within a unit (for example meaningful ordering of descriptors within a bibliographic record).

Organizing and keeping track of goods and services for commerce (esp. ecommerce) and inventory

The functions detailed for information retrieval apply to this special case

Organize a store, an inventory, an online merchandise catalog, a yellow page directory so items can be found

Display the inventory in a meaningful arrangement so users can find tings (as in a store)

Keep track of inventory

These functions apply both to business-to-consumer and to business-to-business commerce. Classification by function or purpose is especially important here.

Ontology for data element definition.

Data element dictionary.

Consider data processing systems in a multinational corporation

Conceptual basis for knowledge-based systems.

Do all this across multiple languages

Mono-, bi-, or multilingual dictionary for human use.

Printed or machine-readable, such as dictionary on CD-ROM or a thesaurus used in conjunction with a word processor

Dictionary/knowledge base for automated language processing

Machine translation and natural language understanding (data extraction, automatic abstracting/indexing). (It should be noted that parsing natural language requires not only morphological information and information about the possible syntactic roles of a term but also a great deal of semantic information.)

Spell check dictionary

Knowledge base for grammar checking.

Functions of an ontological knowledge base in software development

Assist in the design and implementation of the **user interface, esp. choice of terms and icons**.

Terms and icons must be chosen with the sometimes conflicting goals of communicating to the intended user group and of adhering to standards.

Assist in the organization and formulation of **help messages and of documentation** and third-party software books.

Serve as the **lexicon for machine translation** of interfaces and software-related documents

Assist the user in understanding interfaces and documentation, esp. in a foreign language.

Support retrieval of software for the end user or for **software reuse**.

Data element definition and standardization and organization of CASE tool databases.

All this functionality must be provided in **multiple languages** (for example, **software localization** for end users, **CASE tool databases for multinational development teams**)

End of reference list of thesaurus/ontology functions

User-centered indexing / request-oriented indexing

Construct a **classification/ontology**
(embedded in a thesaurus)
based on actual and anticipated **user queries and interests**.

Thus provide a **conceptual framework**
that organizes user interests and
communicates them to indexers.

Index materials from users' perspective:
Add need-based retrieval clues
beyond those available in the document.
Increase probability that a retrieval clue
corresponding to a query topic is available.

Index language as checklist.
Indexing = judging relevance against user concepts.
Relevance rather than aboutness

Implementation:
Knowledgeable indexers
Expert system using syntactic & semantic analysis
& inference.

User-centered indexing / request-oriented indexing.

Sample concepts included in the index language due to user interest

Systemic administration

Intergenerational social mobility

Biochemical basis of behavior

Longitudinal study

User-centered / request-oriented indexing. Sample documents with descriptors

Document

The drug was injected into the aorta

User concept: *Systemic administration*

Document:

The percentage of children of blue-collar workers going to college

User concept: *Intergenerational social mobility*

Document:

CSF studies on alcoholism and related behaviors
User concept: *Biochemical basis of behavior*

User concept: *longitudinal study*
(Longitudinal not mentioned in the document; determined through careful examination of the methods section.)

Design of a classification scheme for fiction based on analysis of actual user-librarian communication

Annelise Mark Pejtersen

Dimensions	*Freq.*	*Sub-classes*	*Freq.*	*Examples*
1. subject matter	38	a. action and course of events (plot)	10	a. mystery novel, book with action
		b. psychological development / description	5	b. love story, book with psychology
		c. social relations	19	c. family chronicle, not with politics
2. frame	25	a. time	13	a. historical novel, books from 16th and 17th centuries
		b. place (geographical, social environment, professions)	12	b. travelogue, books from the countryside, books about working people
3. author's intention /attitude	37	a. emotional experience	34	a. humoristic, suspense, amusing
		b. cognition / information	3	b. philosophical, political, not too many problems
4. accessibility	34	a. readability	16	a. easy, not complicated, not heavy
		b. physical characteristics	18	b. typography, modern/old, series, size, volume
Other formulations	87	a. author's name / title	25	
		b. author's name / title as example	22	b. something like Emily Bronte
		c. good book	32	
		d. diverse	8	

From: **Theory and application of information research**. London: Mansell; 1980. ISBN 0-7201-1513-2. p. 149

Sample user concepts for indexing images

Good scientific illustration

Useful for fundraising brochure

Appealing to children

Cover page quality

User-centered /request-oriented / problem-oriented indexing

As summarized in the overheads, user-centered indexing involves analyzing actual and anticipated user queries and interests and constructing a framework, a hierarchically structured controlled vocabulary, that includes the concepts of interest to the users and thus communicates these interests to the indexers or an expert system that can infer user-relevant concepts from text. The indexers then become the "eyes and ears" of the users and index materials from the users' perspective. The indexer uses the structured list of user-relevant concepts as a checklist, applying her understanding of a document (or other object) to judge its relevance to any of these concepts. This process ensures that users will find the documents that they themselves would judge relevant upon examination.

Request-oriented indexing contrasts with document-oriented indexing, where the indexer simply expresses what the document is about or where simply the terms in the text are used. But, as the examples show, a document can be **relevant** for a concept without being **about** the concept: a document titled *The percentage of children of blue-collar workers going to college* is not necessarily about *intergenerational social mobility*, but a researcher interested in that topic would surely like to find it, so it is relevant.

Request-oriented indexing is essential for good performance in fiction retrieval and even more so in image retrieval. Image retrieval profit from descriptors that capture imponderables, such as the mood of an image or from descriptors indicating possible uses of an image (such as

This perspective on indexing has implications for cross-language retrieval: The conceptual framework must be communicated in every participating language to allow a meeting of minds to take place, regardless of the languages of the user and the indexer. This is particularly salient in the context of indexing : One needs to make sure that, as far as possible, the term used by the indexer in one language communicates the same mood as the term given to the user in another language for searching.

Web-based thesaurus display and incorporation into search functions

Vignette on thesaurus use in searching a digital library. The director of a drug-free community coalition is faced with developing a prevention project and the funding for it. Signing on to the AOD Digital Library, she begins by **browsing the prevention section of the thesaurus hierarchy** to get a structured overview of various prevention approaches. From the **thesaurus scope notes**, some of these approaches seem particularly applicable to her community, so she follows the links to more in-depth explanations. She returns to the thesaurus and follows a link from *prevention through education* to a funding program announcement. She opens the guidelines for submitting proposals to this program and copies a proposal template into her private space (shown in another window) and fills in some text and copies some text (which is transferred with the proper source). From the program announcement, she follows a link to projects funded previously and further to project reports and evaluations. She comes across the **unfamiliar term *triangulation*** and clicks on it to see the **thesaurus entry, which gives an explanation and the hierarchical context**. In another document she highlights the phrase *prevention program evaluation* to initiate a search in the system and one external database. She copies three references with abstracts to her private space. (Later she will return to these, select one for detailed reading, and add more notes and quotes to her emerging proposal.) Returning to the program announcement, she follows a link to relevant research, selects some articles to read, and adds more material to her outline. One of the papers compares the effectiveness of several prevention curricula. She follows a link to the curriculum that came out on top and from there finds further reviews. She also finds some discussion of resources required. She needs some more data — namely, demographics of her community and funding sources for the required local match — so she initiates searches in two external databases, incorporating the results into her proposal. Now she completes the first draft, including the text itself and annotations that explain why a piece is included or why certain language is used. Before submitting the proposal, she emails two board members and a city staff member for comments, giving them access to her private space. The three people read the draft and add their annotations, including suggested wording. The director now revises the draft, requests the final document in PDF format, links to the agency's submission system, and sends off her proposal.

The next page shows a proposed digital library structure built around a thesaurus and the pages following show two steps in a search using the thesaurus.

DL structure diagram and two search steps see file dlthestut2.pdf

search step 1

Search step 2

Web-based thesaurus display

Requirements

Browsing a hierarchy at different levels of detail

Hyperlinks for following relationships

Searching for compounds containing any combination of elemental concepts

Searching for a word or phrase (full complement of Boolean and adjacency operators).
Search in the combination of the descriptor field and the synonymous term field

For a controlled vocabulary search:
Insert descriptor or descriptor + narrower terms into search form

For a free text search:
Insert descriptor + synonyms or
descriptor + synonyms
+ narrower terms + their synonyms
into search form

The following pages have examples of a proposed interface that is very simple but functional. (Fancy graphics are often more a hindrance than a help.)

Thesaurus interface pages are in file dlthestut2.pdf

Fig. 2a from DL proposal

Fig 2b from DL proposal

Searching with elemental concepts

Example 1. AOD Thesaurus

Search for:

central nervous system AND disorder

Result:

GH6.10.2 brain injury
GX4 CNS disorder

Search for:

central nervous system

Result:

EF4.6.16 CNS injection
EW8 CNS function
GH6.10.2 brain injury
GX4 CNS disorder
XV4.4.4 CNS sensory pathway
XZ central nervous system

Searching with elemental concepts

Example 2. LC Classification

Search for:

buildings, architecture AND acoustics

Result:

NA2800 Architectural acoustics
TH1725 Soundproof construction

Search for:

vehicles AND acoustics

Result:

TL681.S6 Airplanes. Soundproofing
VM367.S8 Submarines. Soundproofing

Searching with elemental concepts. Ex. 3 DDC

Search for:
Payment in exchange for some consideration

Result:
general concepts containing this component with examples of more specific Dewey classes (many, but not all, in 330 Economics)

Wage
331.21 Labor economics / Compensation
658.32 General management / Personnel management / Wage and salary administration

Price/Cost
338.52 Production economics / Prices
339.42 Macroeconomics / Cost of living (Prices)
354.5285 Public administration / Admin. of agriculture / Agricultural price supports

Interest
332.82 Financial economics / Interest
336.2426 Public finance / Income taxes / Interest income

Rent
336.11 Public finance / Non-tax revenue / Revenues fr. rents
346.043 44 Law / Private l. / Property / ... / Rent and rent control

Fees
025.11 Operation of libraries / Finance (incl. user fees)
371.206 Schools / Financial management (incl. tuition)
378.106 Higher ed. / Financ. management (incl. tuition)
384.555 Communication / Wireless / TV / Pay television

Example for a word search in an online thesaurus

Search for

commercial AND organization

finds the following record

corporation
ST business organization
ST commercial enterprise
ST company

Search for the German words

Hirn AND Entzuendung

finds the following record

meningitis
ger Hirn-haut-entzuendung
fre menengite

Thesaurus structure

Concept-term relationships

Conceptual structure

- **Semantic analysis and facets**
- **Hierarchy**

Concept-term relationships (Terminological structure)

Controlling synonyms

Term	*Preferred synonym*
Teenager	Adolescent
Teen	Adolescent
Youth (young person)	Adolescent
Pubescent	Adolescent
Black	African American
Afro-American	African American
Alcoholism	Alcohol dependence
Inheritance	Heredity
Ultrasonic cardiography	Echocardiography

Soergel, p. 215, enlarged

Disambiguating homonyms

administration 1 (management)

administration 2 (drugs)

Läufer 1 (Sportler) English: runner (athlete)

Läufer 2 (Teppich) English: long, narrow rug

Läufer 3 (Schach) English: bishop (chess)

discharge 1 (From hospital or program)
German: Entlassung

discharge 2 (From organization or employment)
Preferred synonym: Dismissal
German: Entlassung

discharge 3 (Medical symptom)
German: Absonderung, Ausfluss

discharge 4 (into a river)
German: Ausfluss

discharge 5 (Electrical)
German: Entladung (which also means unloading)

Importance of terminological structure

The terminological structure is equally important
in controlled vocabulary systems and
in free-text searching.

In free-text searching,

synonym expansion of query terms is important for recall

homonym indicators can trigger a question to the user on the intended meaning of the query term.

Conceptual structure

A well-developed conceptual structure

sine qua non for user-centered indexing

very useful for free-text retrieval as well.

The two principles of conceptual structure

facet analysis

hierarchy

Facets.
Semantic factoring or feature analysis

Analyzing a concept into its defining components (elemental concepts or features).

Concept frame with facet slots

liver cirrhosis

Pathologic process:	inflammation
Body system:	liver
Cause:	not specified
Substance/organism:	not specified

alcoholic liver cirrhosis

Pathologic process:	inflammation
Body system:	liver
Cause:	chemically induced
Substance/organism:	alcohol

hepatitis A

Pathologic process:	inflammation
Body system:	liver
Cause:	infection
Substance/organism:	hepatitis A virus

Facet principles

A facet groups concepts that fall under the same aspect or feature in the definition of more complex concepts; it groups all concepts that can be answers to a given question. In frame terminology: The facets listed above are slots in a disease frame; a facet groups all concepts that can serve as fillers in one slot.

Using elemental concepts as building blocks for constructing compound concepts drastically reduces the number of concepts in the thesaurus and thus leads to conceptual economy. It also facilitates the search for general concepts, such as searching for the concept *dependence*, which occurs in the context of medicine, psychology, and social relations.

Facets can be defined at high or low levels in the hierarchy, as illustrated in the next overhead.

Top-level facets

organism
body part
chemical substances by function
chemical substances by structure

Low-level facets

route of administration
- route of administration by scope of drug action (local/topical or systemic)
- route of administration by body part
- route of administration by method of application (injection, rubbing on, etc.)

liver
- liver tissue (hepatocyte, Kupffer cell, etc.)
- liver part (hepatic lobule, portal lobule, etc.)

Hierarchy

groups at high risk of drug use

- suicidal or physically or mentally disabled
- persons from unstable or low-cohesion families
- children of alcoholic or other drug-abusing parents
  - SN Adult or still under age
- children of single teenage mothers
- persons subjected to abuse or neglect (now or past)
  - persons subjected to abuse/neglect by parents
    - latchkey children
  - persons subjected to abuse/neglect by spouse
- single teenage mothers
- school dropouts or those at risk of dropping out
- unemployed or in danger of being unemployed
- economically disadvantaged
- homeless
  - runaway youth
- gateway drug users
- persons engaged in violent or delinquent acts

See also examples given previously in
What is a thesaurus

Uses of facet analysis and hierarchy

Help to organize the concept space and establish concept relationships.

- Discover concepts, esp. general concepts spanning several disciplines

Assist the user in analyzing and clarifying a search problem:

- elicit the facets involved
- present hierarchical structure within each facet

Facilitate the search for general concepts, such as

- inflammation or
- dependence (which occurs in the context of medicine, psychology, and social relations)

Hierarchic query term expansion

These functions are useful in both

- controlled vocabulary and
- free-text searching.

Concept discovery through facet analysis and hierarchy building

Through facet analysis and hierarchy building,
one often discovers concepts
that are needed in searching or
that enhance the logic of the concept hierarchy.
Need to create terms for these concepts.

Consider

train station, *bus station*, *harbor*, *airport*

Common semantic component: ***traffic station***

gin, whiskey, cherry brandy, tequila, etc.

common semantic component: ***distinct distilled spirits***
(counterpart of the already lexicalized
neutral distilled spirits)

transactional analysis, *dream analysis*, *insight therapy*, *Gestalt therapy*, *reality therapy*, *cognitive therapy*

Umbrella concept for structuring the hierarchy and for retrieval: ***analytic psychotherapy***
(methods that seek to assist patients in a personality reconstruction through insight into their inner selves)

Payment in exchange for some consideration (see above)

Searching interaction:

Facets for eliciting user needs

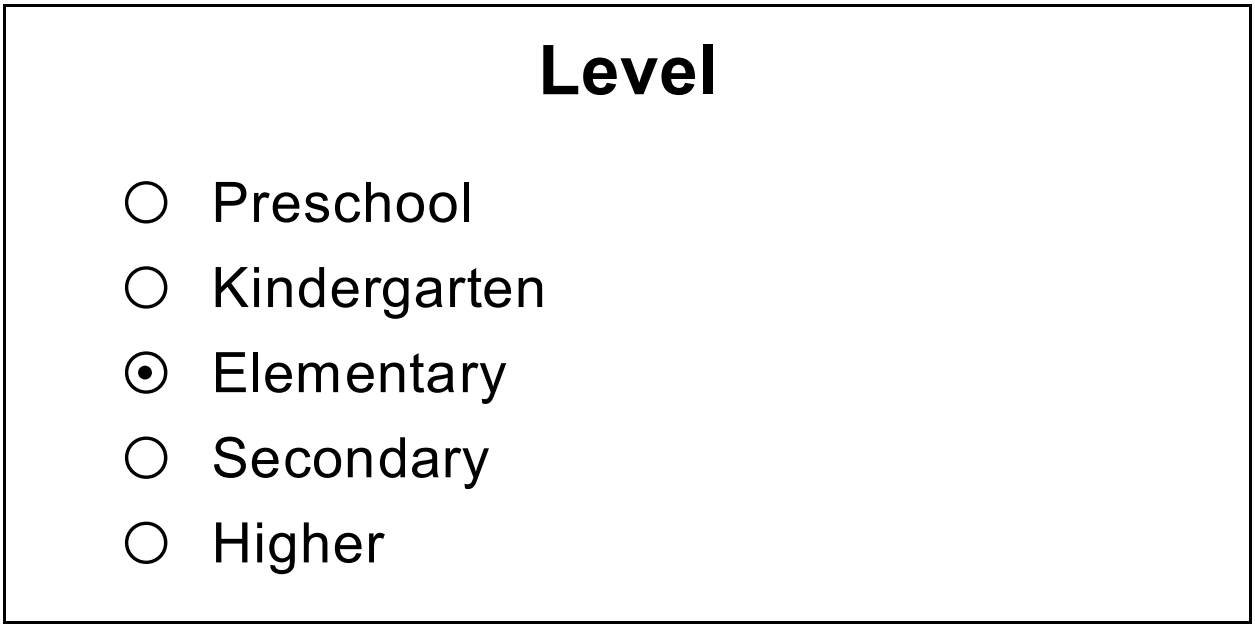
User enters subject field of search.
System displays list of facets (limiting aspects).
User indicates first aspect for limiting the search

Subject field of search: **Education**

Indicate limiting aspects to be used:

- ◉ Level
- ○ Ethnic origin of students
- ○ Giftedness/handicap of students
- ○ Subject
- ○ Country
- ○ Public/private

User selects level descriptor



Searching interaction:

Facets for eliciting user needs

System displays query formulated so far.
User indicates ***Ethnic origin***
as a limiting aspect

Subject field of search

Education

Indicate limiting aspects to be used

- ⦿ Level **Elementary**
- ⦿ Ethnic origin of students
- ○ Giftedness/handicap of students
- ○ Subject
- ○ Country
- ○ Public/private

User selects *Ethnic origin* descriptor

Ethnic origin

- ○ Latin American / Spanish American
- ○ Mexican American
- ○ Puerto Rican
- ◉ African American
- ○ Asian American
- ○ Chinese American
- ○ Japanese American

. . .

- ○ Gypsy

Searching interaction:

Facets for eliciting user needs

System displays query formulated so far.
User indicates *Subject* as the next limiting aspect

Subject field of search

Education

Indicate limiting aspects to be used

⦿	Level	**Elementary**
⦿	Ethnic origin of students	**African American**
○	Giftedness/handicap of students	
⦿	Subject	
○	Country	
○	Public/private	

Searching interaction:

Facets for eliciting user needs

After a few more interactions, the system displays the completed query formulation

Subject field of search

Education

Indicate limiting aspects to be used

⊙ Level	**Elementary**
⊙ Ethnic origin of students	**African American**
○ Giftedness/handicap of students	
⊙ Subject	**Reading**
⊙ Country	**U.S.**
○ Public/private	

Thesaurus-supported Web search engines

These sites are still experimental; they come and go

Using synonym expansion

www.simpli.com/ (was working 2001, did not respond July 6, 2002)

Using homonym disambiguation

www.hotbot.com/ (used to do this at one time)

www.oingo. com (will change name to appliedsemantics.com)
Finds Open Directory categories (Full Web search with homonym disambiguation is hard)

Using a large enriched thesaurus/ontology

www.seruba.com/ (defunct)

Implementing thesaurus functions in retrieval systems with emphasis on cross-language retrieval

Important principle: Let the system do the work. Full exploitation of thesaurus power cannot rely on users learning how to use a thesaurus but rather requires a system that gives behind-the-scene knowledge-based assistance with the thesaurus serving as the knowledge base.

Controlled vocabulary

With a controlled vocabulary there is a defined set of concepts used as descriptors in indexing and searching. The user can browse the thesaurus hierarchies to identify search concepts; or the user can start from a term or phrase and consult the thesaurus to find the proper descriptor(s) or let the system do the mapping behind the scene. In either case, the user need not worry about the various ways each search concept is expressed in natural language. In cross-language retrieval this simply means that the user should be able to use a term in his own language to find documents (or whatever the retrieval objects are) indexed by the corresponding descriptor (concept identifier). The user can consult a multilingual thesaurus that includes for each concept corresponding terms from several languages and has an index for each language. Or the mapping from a user term in any covered language to the corresponding descriptor could be automatic. As an example, consider a library catalog using the Library of Congress Subject Headings, for which French and Spanish translations are available. In the VTLS automated library system, each subject heading is identified by a number that is used in the document records. The authority file includes for each subject heading the preferred term and any synonyms; this information can be included in multiple languages. From any user term in English, French, or Spanish the system can map to the corresponding subject heading number through a free-text search on authority records to find any subject heading for which either the preferred term or any synonym contains the user's query word or phrase in any language.

Whenever the mapping from user terms to descriptors is done "behind the scenes", transparent to the user, the system should ask the user for clarification whenever the query word or phrase has multiple meanings and cannot be disambiguated automatically. Beyond that, showing the user the descriptor(s) the system came up with in their hierarchical context might improve the accuracy of the query formulation and thus retrieval. The success of this type of interaction depends on the quality of the hierarchy and the interface.

If voice input is available, one might even include the spoken form of terms in the thesaurus to enable voice input of query terms which would then be mapped to the appropriate descriptors.

A controlled vocabulary system must promote correct use of descriptors in indexing. Hierarchy and scope notes assist the indexer in understanding the meaning of a descriptor. Request-oriented indexing assures that important descriptors are not overlooked.. In cross-language retrieval the thesaurus version in each language must make sure that the indexer in that language fully understands the meaning of a descriptor that originated from another language; otherwise, the indexing of such a descriptor will not be consistent across the database.

Automated indexing with a controlled vocabulary, particularly if it is to take a request-oriented slant, can be accomplished with a knowledge base that (1) allows recognition of important words and phrases (in spoken form for speech indexing) and allows for homonym disambiguation and (2) gives mapping rules that lead from the (possibly weighted) set of words and phrases identified for a document to a set of descriptors that should be assigned.

Such mapping rules can take many forms. In their simplest form, they specify a direct mapping from text words or phrases to the appropriate descriptors for each word or phrase (and possibly even word or phrase combinations). To increase accuracy, the mapping can be made dependent on context (Hlava 97). A more complex mapping relies on association strengths between terms (words and phrases) and descriptors. Broadly speaking, the association strength between term T and descriptor D could be seen as the predictive probability that the document containing term T should be indexed with descriptor D. Such association strengths can be computed from a training set of indexed documents. This is the approach often taken in automated text categorization, where often, but not always, the goal is to index each document by only one descriptor (assign it to one of a set of non-overlapping categories). An advanced version of this approach is the use of "topic signatures", profiles consisting of a set of terms with weights; a document is assigned the topic if its terms match the topic signature (Lin 1997). In effect, a topic signature is a query which identifies documents relevant to the topic.

As the foregoing discussion illustrates, the knowledge base needed to support automated indexing is more complex than a thesaurus for manual indexing. It must include more terms and term variants so that the words and phrases important for indexing can be recognized in the text, and it must include information needed for the disambiguation of homonyms (which often requires determining the part of speech of a text word).

For indexing and searching, a controlled-vocabulary cross-language retrieval system can be seen as a set of monolingual systems, each of which maps the terms from its language to a common system of concepts used in indexing and searching. For manual indexing and query formulation, this is accomplished through a multilingual thesaurus, which may in fact consist of multiple monolingual thesauri linked through common descriptor identifiers (such as Dewey Decimal class numbers). Automated indexing in cross-language text retrieval with texts in multiple languages means mapping from each language to the common conceptual structure represented in the controlled vocabulary. The knowledge base component dealing with identification of words and phrases for automated indexing can be developed independently for each language. Mapping rules that are entirely term-based can also be developed independently for each language. However, some mapping rules, for example rules based on context or topic profiles, may include conceptual elements that could be shared across languages.

There are a number of controlled-vocabulary cross-language retrieval systems based on manual indexing in use in bilingual or multilingual areas such as Switzerland, Belgium, Canada, and areas of the US with large Spanish-speaking populations; in international organizations, such as the European Community; and in international collaborative systems, such as AGRIS. These systems are based on the Universal Decimal Classification, which has been translated into many languages (library of the ETH, Zurich); on the Library of Congress Subject Headings (translated into French); on EUROVOC, an EC thesaurus in 9 languages; and AGROVOC, a thesaurus in

three languages created by translation from its original English-only version. There are a large number of thesauri that either have been developed as multilingual thesauri or have been translated into several languages.

Free-text searching

High-recall (even moderate-recall) free-text searching requires query-term expansion as discussed above. Cross-language free-text searching, finding texts in one language that are relevant for a query formulated in another language without relying on controlled vocabulary indexing, is an extension of this principle: Each query term must be mapped to a set of search terms in the language of the texts, possibly attaching weights expressing the degree to which occurrence of a search term in a text would contribute to the relevance of the text to the query term. To assist with this task, a thesaurus must include the mapping information. If the thesaurus includes fine-grained definitions that deal with subtle differences of meaning, distance between such definitions can be used to derive term weights.

A major difficulty of this mapping is that a homonym used in the query gives rise to multiple translations, each corresponding to one of its meanings. The target terms may in turn be homonyms in their language and thus retrieve many irrelevant documents unless text terms are disambiguated. (This problem exists in synonym expansion in one language as well but is exacerbated in cross-language text retrieval.) When the mapping goes to a term that has multiple meanings, the specific meaning should be identified, possibly in interaction with the user. For best retrieval results the terms in the texts should also be disambiguated so that only documents that include the term in the right sense score

The issue of homonymy in retrieval is not as straightforward as it may seem at first glance (Sanderson 1994). First of all, quite a bit of disambiguation may occur "naturally", in that a given term may assume only one of its meanings in the specific domain of the collection and therefore in the queries. Second, in a multi-component query, a document that includes a homonymous term from the first query component in a meaning other than that intended in the query is unlikely to also include a term from another query component; thus excluding irrelevant documents may not require disambiguation in either the query or the texts. On the other hand, with single-concept query to a general collection (such as the World Wide Web), disambiguation can be expected to have a beneficial effect on retrieval performance. Failing that, a system might be able to suggest to the user an additional query component that would separate out the documents that include the query term but in a different meaning. Note that information extraction is much more dependent on homonym disambiguation.

In any event, for best support of free-text retrieval a thesaurus should flag homonyms, give their senses, and include rules for disambiguation.

The greater difficulty of free-text cross-language retrieval stems in no small measure from the fact that one must work with actual usage, while in controlled-vocabulary retrieval one can, to some extent, dictate usage.

Thesauri for knowledge-based search support

Whether searching is by controlled vocabulary or by free text, it is often helpful to the user to browse a well-structured and well-displayed hierarchy of concepts, preferably with the option of including definitions. A more sophisticated system may guide a user through a facet analysis of her topic. These aids provided by the system enable the user to form a better idea of her need and to locate the most suitable descriptors or free-text search terms. The guidance through facets and their hierarchical display must be available in the language of the user. These suggestions are based on the assumption that browsing a hierarchy is natural to most users and that users will appreciate the structure provided. This assumption rests on the belief that people try to make sense of the world and that guided facet analysis and browsing well-structured hierarchies help them do so. There is anecdotal evidence to support this assumption, but it needs to be investigated by building prototype systems and studying users' success (see, for example, Pollitt 1996).

This is one example of using a thesaurus as a knowledge base to make searching more successful. The assistance provided does not require that the user be an expert in classification and thesauri. This is even more true for "behind-the-scenes" assistance. There is no need to teach users about following a cross-reference from a synonym to a descriptor if the system searches for the descriptor automatically. There is no need to tell the user to look under narrower terms also if the system can do a hierarchically expanded search. There is no need to tell the user about strategies of broadening the search if the system, in response to a user input that not enough was found, can suggest further descriptors to be searched based on cross-references in the thesaurus. Sophisticated retrieval software can make the use of thesauri in retrieval independent of the user’s knowledge and thereby can get much more mileage out of the investment in thesauri.

Key issues in multilingual thesauri

Conceptual systems in different languages differ

What concepts are lexicalized differs from language to language

Translation of an English thesaurus into French does not make a French thesaurus

Develop common conceptual structure integrating perspectives from multiple languages. Harmonize concepts where possible, keep concepts where necessary, invent a term if a concept is not lexicalized in a language

Problems of structure: simplified versus real

Simplified

English term 1		French term 1
English term 2	Concept	French term 2
English term 3		French term 3

Real

English term 1		French term 1
English term 2	Concept	French term 2
English term 3		French term 3

Special issues in multilingual thesauri

A multilingual thesaurus for indexing and searching with a controlled vocabulary can be seen as a set of monolingual thesauri that all map to a common system of concepts. With a controlled vocabulary, indexing is concept-based; cross-language retrieval is simply a matter of providing designations for these concepts in multiple languages so that queries can be written in multiple languages. However, as the example given above illustrates, conceptual systems represented in the vocabulary of different languages do not completely coincide.

The crux of the matter, then, is which concepts to include. Ideally, the thesaurus should include all concepts needed in searching by any user in any of the source languages. Language differences often also imply cultural and conceptual differences, more so in some fields than in others. We need to create a classification that includes all concepts suggested by any of the languages. At a minimum this includes all relevant concepts lexicalized in at least one of the source languages. Also, different languages often suggest different ways of classifying a domain; the system needs to be hospitable to all of these. The problem that has bedeviled many developers of multilingual thesauri is that a concept lexicalized in one language may not be lexicalized in another and that the terms that do exist often vary slightly in meaning, possibly giving rise to different relationships. Starting from the misguided notion that a thesaurus should include only concepts for which there is a term in the language and that term meanings cannot be adjusted for purposes of the thesaurus, they had difficulty making the system of concepts the same for all languages. But, as we have seen, even in a monolingual thesaurus the lexicographer often discovers concepts needed in searching or to enhance the logic of the concept hierarchy and then needs to create terms for these concepts. In multilingual thesauri this necessity arises more often, particularly when different languages differ in the hierarchical levels at which they lexicalize concepts.

The principle proposed here is to establish a common conceptual system, which may require an arduous. and expensive, process of negotiation, and then arrange for the terms in all languages to fit, giving proper definitions, of course. In contrast, many multilingual thesauri have been produced by translating an established monolingual thesaurus, thus accepting the conceptual system of one language and limiting the number of synonyms (if any) in the other languages. EuroWordNet is based on a more comprehensive, but still limited approach: Rather than developing a conceptual structure based on an analysis of the vocabulary in all participating languages, EuroWordNet accepts the conceptual system of the English language WordNet . On the other hand, EuroWordNet does not simply translate WordNet but develops synonym sets independently in each language and then links them to the concepts (synsets) established in WordNet.

So far we have described a multilingual thesaurus with a common conceptual system, however constructed, where the terms of each language are linked to a concept but not to each other. Relationships between terms from different languages are established through their relationships to concepts. This simple model will do for most information retrieval. But unless the concepts are exceedingly fine-grained and include in their definition affective components of meaning and usage considerations, this model is too simplistic for natural language processing, especially

translation. There one needs direct relationships between terms to enable the proper word choice in translation.

The problems discussed here and illustrated in the example above have major implications for cross-language free-text searching: Each query term should be mapped from the source language to its multiple equivalents in the target language; each of these equivalents may have other meanings in the target language, presenting potential problems for retrieval. The query term may not have a precise equivalent in the target language; one may need to map to broader or narrower terms, distorting the meaning of the original query.

Evaluation of Thesauri

Introductory example: Yahoo classification

Reference and General Interest	Subjects
	Science Animals, Astronomy, Engineering ...
Reference Libraries, Dictionaries, Quotations ...	**Health** Medicine, Diseases, Drugs, Fitness ...
	Social Science Archaeology, Economics, Languages ...
Computers & Internet Internet, WWW, Software, Games ...	**Society & Culture** People, Environment, Religion ...
	Government Elections, Military, Law, Taxes ...
	Business & Economy B2B, Finance, Shopping, Jobs ...
News & Media Full Coverage, Newspapers, TV...	**Education** College and University, K-12 ...
Entertainment Movies, Music, Humor, Cool Links ...	**Arts & Humanities** Literature, Photography ...
Recreation & Sports Sports, Travel, Autos, Outdoors...	
Regional Countries, Regions, US States ...	

Yahoo Classification. Home. Meaningful arrangement

Yahoo classification. Education. Meaningful arrangement.

Home >
Education

Categories

Browse by Region (170)
By Culture or Group (398)
By Subject (11)

Information sources
Bibliographies (4)
Web Directories (47)
News and Media (83)
Chats and Forums (40)
Conferences (52)
Journals (36)
Statistics (6)

Education by level
Early Childhood Education (90)
K-12 (53910)
Higher Education (16638)
Adult and Continuing Education (325)

Special students and subjects
Special Education (168)
Disabilities@
Literacy (12)
Bilingual (24)
Career and Vocational (236)
Correctional@

Educational methods
Theory and Methods (659)
Teaching (63)
Instructional Technology (334)
Distance Learning (476)
Standards and Testing (63)
Academic Competitions (79)
Graduation (53)

Political and economic aspects
Policy (52)
Reform (70)
Equity (27)
Financial Aid (395)
Employment (143)

Organizational aspects
Government Agencies (77)
Organizations (3008)
Companies@
Programs (322)

Evaluation of Knowledge Organization Systems (KOS)

Characteristics for describing and evaluating KOS

(classifications/ontologies/taxonomies/index languages/thesauri/glossaries/dictionaries)

(For some items, a section number from Soergel, Organizing information (starting with a digit), and/or Soergel, Indexing languages and thesauri (starting with a capital) is given.)

1. **Overall description and evaluation**

1.1 **Purpose**, for example

Providing "conceptual infrastructure"

Mapping out the conceptual structure and providing a common language for a field

Providing classification/typology and concept definitions. Clarifying concepts by putting them into context. Thus providing orientation and serving as a reference tool for individual researchers and practitioners and thereby

Assisting with the exploration of the conceptual context of a research problem and in structuring the problem, thereby providing the conceptual basis for the design of good research, for the consistent definition of variables, and thus the cumulation of research results.

Providing the conceptual basis for the exploration of the various aspects of a program in program planning, in the identification of approaches and strategies, and in the development of evaluation criteria

Information storage and retrieval (IR)

One information system

Several IR systems, switching language. Support the coordination or combination of several databases in the same area to facilitate access to multiple databases.

Assisting readers in understanding text

Assisting writers with conceptualizing a topic and with finding the proper term

Translation

Language learning

In each case specify the intended audience

If purpose is IR specify

Information system(s) in which the vocabulary is to be used

Use of the vocabulary

Vocabulary control in indexing and searching (controlled vocabulary)

Vocabulary control only for searching. Assist with clarifying a search topic and assembling all applicable concepts and terms, whether searching with a controlled vocabulary of free-text.

IR technique(s) (such as: printed index, computer search system). Support of inclusive (hierarchically expanded) searching

Automated vs. manual indexing or query formulation. Approach to indexing to be supported: Request-oriented vs. entity-oriented

Techniques for eliciting user needs (e.g., menu based on search tree; questions based on facet structure)

1.2 **Relationship to other KOS**, especially standard schemes

1.3 **Summary evaluation** of the vocabulary's adequacy for the stated purpose based on the more detailed analysis as outlined below.

2. **Coverage of concepts and terms. Sources, quality of usage analysis**.

2.1 Concepts: scope, breadth of coverage (See also 2.3.1)

2.2 Concepts: specificity, depth of coverage

Completeness of coverage at each level of specificity considering all concepts (descriptors and other preferred terms) and descriptors alone (F0.4.3)

Specificity must be adapted to the purpose. Assistance in the choice of terms or the comprehension of text requires many nuances. An IR system for propositions requires high specificity . A bibliographic IR systems may require only low specificity.

2.3 Sources from which concepts and terms are included (natural languages, classifications/thesauri, etc.).
Relationship to other vocabularies, especially standard schemes.

For each source:

2.3.1 Completeness of coverage; all vs. selected concepts; all vs. selected terms for each concept (this includes coverage of synonyms)

2.3.2 Quality of analysis of actual term usage in the source.

2.3.3 Recency

Specifically: Completeness of coverage of the terminology from a given language (English, French, German, Chinese, etc.; the language is the source)

2.4 Augmentation of sources through concepts created in concept analysis (15, C3)

2.4.1 Are all necessary facets included?

2.4.2 Formation of new concepts arising from semantic factoring and other methods of concept analysis. Specifically: Are the concepts applicable across disciplines? Are the concepts applicable across different societies and cultures? (See also 3.4)

2.5 Choice of terms

2.5.1 Form of terms - consistency, adherence to common usage.

2.5.2 Appropriateness of selection of preferred terms from among synonyms.

2.5.3 Choice of terms to designate descriptors (F0.4.2) Closeness to user terminology.

2.6 Nature of notation (if none, state that) (D4)

3. **Conceptual analysis and conceptual structure. Terminological analysis**

3.1 Quality of conceptual structure (14, C1)

Types and degree of differentiation of conceptual relationships included:

3.1.1 Expression of concepts through elemental concepts (closely related to definition)

3.1.2 Hierarchical relationships (polyhierarchy)

3.1.3 Associative relationships

Completeness of conceptual relationships included.

3.2 Quality of definitions, explications, scope notes (correctness, detail, clarity). (C3)

3.3 Completeness of terminological relationships.

Does the thesaurus contain terms that are synonymous or quasi-synonymous without indicating the relationship?

4. **Use of precombination in the index language** (cuts across 2 and 3) (14, 15, C2)

4.1 To what degree are descriptors precombined?

4.2 To what extent are precombined descriptors enumerated and/or given in the alphabetical index? Built by the indexer? Updating characteristics.

Are precombined descriptors designated by an independent symbol or a string of symbols? Citation order free or fixed? To what extent do the components of a precombined descriptor determine its place in the arrangement? (Relates also to 5)

5. **Access and display. Format of presentation of the vocabulary**

Consider for each format access/retrieval by concepts versus access/retrieval by terms

Access can be provided through arrangement in a printed document or through a computerized search system.

5.1 Format of printed document

5.1.1 Overall format (D1)

Thesaurus parts and information given in each, connections between them.

Is the overall format clear and helpful for finding the appropriate concepts and terms or notations in indexing and query formulation?

5.1.2 Display of conceptual relationships
- through arrangement (15.5.2, C2, D3)
- through cross-references (D3.1.1,1)
- through descriptor-find index (15.5.1, D3.6)

How well does the display reflect the conceptual analysis (e.g., sequence of concepts on some hierarchical level) (D3.1.2)

5.1.3 Display of terminological relationships. Format of alphabetical index (C5)

5.2 Access through computer systems. Retrieval of concepts and terms. Navigation. Format of on-line displays

5.2.1 Overall format. Available windows and their relationships

5.2.2 Display of conceptual relationships, esp. hierarchy. Localized hierarchical chains vs. entire hierarchy. Overviews and total hierarchy. Expandable levels vs. expanded or expand-all option. Graphical displays, concept maps. Are cross-references active hyperlinks? Is there an online descriptor-find index.

5.2.3 Access by words and phrases. Is the thesaurus database searchable? How does the search work. What is searched? Just the term itself, synonyms, scope notes, all cross-references (not good!)?

5.3 Format of machine-readable form (if any). What standard is followed

5.4 Detail of keeping records of the origin of information included in the vocabulary.

6. **Updating**

Outline for the analysis of subject access vocabularies. DDC

1. Purpose

1.1 **Information system** or type of information system in which to be used

Bibliographic information system. Intended for public and school libraries.

1.2 **Intended for** controlled vocabulary indexing ☉ or query term expansion ☐ (Ch. 12, Introduction)

1.3 **Type of file and search mechanism** for which originally designed

Shelving ☉ Card catalog ☐ Online system ☐ (Now promoted for Web subject directories)

2. Coverage and designation of concepts. Coverage and format of terms

2.1 **Concepts: Scope**, breadth of coverage. Recency of concepts

Universal — covers all of knowledge. But focus on Western culture, esp. US.

2.2 **Concepts: Specificity**, depth of coverage. (Section16.2.2). Coverage at each level of specificity.

Medium specificity. Would need closer analysis by subject area. Geography table quite specific.

2.3 Are all needed **facets** included? Concepts formed in semantic factoring and facet analysis? (S.a. 3.1)

Some general concepts included in the general tables and the in-schedule tables. Many others not included by themselves but only as components in one or more precombined descriptors. Completeness of explicit and implied facets? Answer would require extensive analysis.

2.4 **Terms**: Completeness of coverage (completeness of lead-in vocabulary). Recency of terms

Some lead-in synonyms included in the alphabetical index. How complete? Would need extensive analysis!

2.5 **Form of terms**: Consistency, adherence to common usage. *Terms seem appropriate. Many classes cannot be expressed by a simple term but need a phrase devised by the editor.*

2.6 **Nature of notation** (if none, state that). (Section 15.5.2) *Decimal, highly expressive (with some exceptions).*

3. Terminological and conceptual analysis and conceptual structure.

3.1 **Quality of conceptual structure** (14): Facet analysis. Types and degree of differentiation of conceptual relationships included. For each type indicate the completeness of inclusion. (Fill in 3.1.1 - 3.1.3)

3.1.1 Expression of concepts through elemental concepts (closely related to definition)

For enumerated compound concepts: Sometimes done implicitly in the relative index. For precombined descriptors constructed according to DDC rules: Done by the indexer.

3.1.2 Hierarchical relationships (polyhierarchy) (Shown by arrangement or Broader Term / Narrower Term X-ref)

Monohierarchical. A few additional BT/NT through cross-references. Many hierarchical relationships implied by the relative index (Example: The classes shown under Blind).

3.1.3 Associative relationships. (Implied by physical proximity in the arrangement or explicit Related Term X-ref)
Some explicit cross-references

3.2 **Quality of definitions**, explications, scope notes (correctness, detail, clarity).

Many notes throughout the schedules and in the Manual. Mostly usage notes explaining the difference between classes or instructions on how to form new precombined descriptors. A few definitions

3.3 Completeness of terminological relationships: Does the vocabulary contain terms that are synonymous or quasi-synonymous without indicating the relationship? *Not a problem in a classification like DDC.*

4. **Use of precombination in the index language** (concerns both 2 and 3) (14, 15, esp. 15.4)

4.1 To what degree are descriptors precombined?

DDC can be used with a medium to high degree of precombination, depending on how many new precombined descriptors the indexer builds.

4.2 To what extent are precombined descriptors enumerated and/or given in the alphabetical index?

Medium degree of enumeration in the schedules, some addl. precombined descriptors in the index.

To what extent can the indexer build additional precombined descriptors?

To a large extent. Libraries differ in their use of this option.

Are precombined descriptors designated by an independent symbol or a string of symbols? Combination order free or fixed? To what extent do the components of a precombined descriptor determine its place in the arrangement? (Relates also to 5) (Section 15.5.2)

Enumerated precombined descriptors have their own independent symbol (which sometimes is constructed using notation components from tables). Combination order is fixed. The components completely determine the place of a precombined descriptor built by the indexer.

5. **Access and display. Format of presentation of the vocabulary**

Consider for each format access/retrieval by concepts versus access/retrieval by terms.

Access can be provided through arrangement in a printed document or through a computer search system.

5.1 **Format of printed document** (Fill in 5.1.1- 5.1.3)

5.1.1 Overall format: Thesaurus parts and information given in each, connections between them. Is the overall format clear and helpful for finding the appropriate concepts and terms or notations in indexing and query formulation?
Introduction (v.1), Tables (v.1), Schedules (v.2+3), Relative Index (v.4), Manual (v.5)
Need to go back forth between schedules and manual, otherwise reasonably helpful.

5.1.2 Display of conceptual relationships (Broader Term, Narrower Term, Related Term)

- through linear arrangement or graphical display (Section15.5.2)

In the tables and schedules.

- through cross-references (Section 14.1)

In the tables and schedules.

- through descriptor-find index (Section 15.5.1)

The relative index combines the functions of an alphabetical index and a descriptor find index of sorts.

How well does the display reflect the conceptual analysis, e.g., sequence of concepts on the same hierarchical level (sequence of the children of a concept, that is, the concepts one level further down).
Usually the sequence of classes makes good sense.

5.1.3 Display of terminological relationships (Synonymous Term)

Terminological relationships are displayed only in the relative index, which gives the lead-in term and points to the appropriate class number.

5.2 **Access through computer system**. Navigation. Format of on-line displays

This would be an analysis of Dewey for Windows. Not required here.

Some points on procedure for evaluating a thesaurus

What went into the construction of a thesaurus

Check sources used.

Check procedures used in thesaurus development.

Examine the thesaurus structure and content

Use knowledge of thesaurus structure for analysis of structure and internal consistency.

Check against other thesauri and against encyclopedias, dictionaries, or other authoritative sources.

In this examination, collect data on all the criteria in parallel by looking through the thesaurus, probing for examples and following leads as they arise. Interact with the thesaurus. Keep notes according to the outline of criteria. (Much like anthropological field work, where the investigator observes as events occur, keeping the variables of interest in mind but is always open to aspects not though of beforehand. At some point, the notes are indexed and sorted by the variables of interest.)

Check how the thesaurus works

Try the thesaurus on search request and on documents; indexing and retrieval experiments (see F6). Online interaction with the thesaurus.

Can be done by the evaluator - for example, to shed light on completeness - or by real indexers and real searchers. In the latter case, knowledge of the subjects must be considered. Indexers may require training. Indexing experiments where several indexers index the same documents can be very useful; disagreements may point out problems in the thesaurus.

Thesaurus development, with emphasis on multilingual thesauri

Building a thesaurus, especially a multilingual thesaurus, takes a lot of effort. Some term relationships can be derived by statistical analysis of term occurrence in corpora, but this will not result in the kind of well-structured conceptual system described above. Developing such a structure requires intellectual effort.

A common method for thesaurus construction in a single language is to work bottom-up: One collects a list of terms (words and phrases), preferably from search requests, but also from documents, free-term indexing, and other thesauri. These terms are then sorted into increasingly fine-grained groups, until a group contains only synonyms or terms that, for purposes of the thesaurus, can be considered synonyms. In this process at least some homonyms will be detected; they must be disambiguated into several senses, each expressed by its own (possibly newly coined) term having one meaning and being grouped accordingly. A group of synonyms can be considered to represent a concept; usually a preferred term to designate the concept is selected, but some other concept identifier can be used. A first rough hierarchy of concepts emerges from this process.

Now perform conceptual analysis, especially facet analysis at various levels, resulting in a well-structured faceted hierarchy. Next write definitions (scope notes) (often results in rethinking the hierarchy) and introduce relationships between concepts that complement the hierarchy.

The development of a multilingual thesaurus is, naturally, an even more complex undertaking; the basic approaches are summarized in the table below. The ideal way to develop a multilingual thesaurus is to start from a pool of terms in all covered languages and carry out the process without regard to the language of the terms. This will bring together terms from different languages that have the same meaning into one group. This process gives all languages an equal chance to contribute concepts and concept relationships. It also forces a careful analysis of the meaning of each term in each language to determine the degree of equivalence, making it possible to develop the fine-grained structure of definitions that has the potential of providing powerful support to free-text cross-language retrieval.

Of course, this process requires a lexicographer knowledgeable in the subject matter of the thesaurus and fluent in all covered languages, not a very practical requirement. A more practical variation that still maintains the spirit of this approach is to start with two languages and develop the conceptual structure — a bi-lingual lexicographer is needed in any event. Definitions should be written in both languages. One would then work on a pool of terms in a third language and fit it into the structure, creating new concepts as necessary. This is not at all the same as translating the thesaurus into the third language. This requires a lexicographer fluent in one of the starting languages and the third language. Add other languages the same way.

The result of such a process is a conceptual system that brings the conceptual structures embedded in the different languages under one roof, so to speak.

The most common approach to the construction of a multilingual thesaurus is to translate an existing monolingual thesaurus into one or more languages. But this approach is problematic: The original language and its vocabulary determine the conceptual structure, and one merely looks for equivalent terms in the second language with-out covering its terminological richness .

In some multilingual thesauri, only one term in the target languages is provided, making the thesaurus unsuitable for query term expansion in free-text searching.

In between is an approach in which one starts with a monolingual thesaurus as the center and fits terms from one or more other languages into the structure of this central thesaurus without changing the concepts or the hierarchy. EuroWordNet (Gillaranz 1997) takes an improved variation of this approach, working with the English WordNet as its central thesaurus. In EuroWordNet, separate and independent word nets are constructed in each language in parallel efforts, each identifying synonym sets in that language (A synset can be considered a concept). The synsets of each language are then mapped indepedently to WordNet synsets; no changes are made to WordNet. In addition to identity, this mapping allows for hyponym and hypernym relationships, thus indicating that the concept identified in the language being worked on is not included in WordNet, but giving at least the hierarchical location. EuroWordNet also uses a very weak variation of approach 5: The participants developed a "top ontology", which presumably reflects and integrates perspectives from their individual cultures. In addition to being mapped to WordNet, the individual language synsets are also mapped to this top ontology.

Building a multilingual thesaurus

Requirements: Must cover all concepts of interest to the users in the various languages, at a minimum all domain concepts lexicalized in any of the participating languages.
Must accommodate hierarchical structures suggested by different languages.

Approaches (by increasing complexity and quality)

(1) Start from monolingual thesaurus and translate. This approach does not capture concepts lexicalized only in another language and is biased to the conceptual structure underlying the starting language. May not produce all synonyms in the second language.

(2) Start from a monolingual thesaurus as the center. Collect terms from other languages and establish correspondences of these terms to the central thesaurus. Suffers from similar bias toward the starting language as (1), but may cover more synonyms in the other languages.

(3) Work with a central thesaurus as in (2), but after collecting terms from a second language first group them into synsets, that is, derive concepts each of which is represented by a set of terms, and then map each concept to the corresponding concept in the central thesaurus or indicate that the concept is new and give the nearest broader or narrower concept in the central thesaurus. Note that the central thesaurus remains unchanged.

(4) As (2), but add concepts not in the starting thesaurus. This mitigates bias, but the central thesaurus now becomes a moving target.

(5) Start from a pool of terms from all participating languages and organize them into a conceptual framework, establishing term correspondence in the process. This approach results in a true "conceptual interlingua" not biased to any one language, but offering a home to multiple conceptual perspectives. This approach requires most effort.

.

Thesaurus development example

Audience/Demographic Characteristics

Raw term list

Terms collected from lists used in three NCADI databases, from the NCADI request form, and from *Breaking New Ground for Youth At Risk*, duplicates eliminated, in alphabetical order

A/D prevention professional
A/D treatment professionals
Administrator/Manager
Adults (25-59 years)
African Americans/Black
Asians and Pacific Islanders (Chinese, Japanese, Vietnamese, etc.)
Attorney
Biomedical researchers
Blacks
Caucasians
Children subjected to abuse and neglect
Children and youth who are economically disadvantaged
Children (pre-adolescents)
Children of alcoholic or other drug-abusing parents
Clergy
College students
Community organization leader
Community service groups
Correction officer
Criminal/juvenile justice
Disabled
EAP Practitioners
Educator/teacher/trainer (specify grade[s])
Elderly (60 + years)
Elementary youth (5-12)
Employees
Employer
General public
General public, personal concern
General public, concern for family/friend
Grantee
Handicapped/Disabled
Health care providers (physicians, nurses, Pas, NAs, pharmacists)
Health care professional
High-risk families
High-risk youth
High-risk families/youth (including COAs and ACOAs)
Hispanics/Latinos
Homeless or runaway youth
Homosexuals (males and females)
IV drug users
Jr. High Youth
Judge
Latchkey children
Legislator
Librarian/Information Specialist
Media representatives
Mental health professional
Native Americans (American Indians and Eskimos)
Other
Parents (specify age of child)
Parole/Probation officer
Patients
Police officer
Policy makers/Administrators
Preschool (age 4 and under)
Psychosocial researcher
Recreation/Sports Personnel
Reporter/writer
Researcher
School dropouts or those at risk of dropping out
School Administrator
Scientists and researchers
Single teenage mothers and their children
Social service professional
Sr. High Youth (16-18)
Student
Unemployed youth or those in danger of being unemployed
Unknown/anonymous
Women
Young Adults (18-25 years) (19-25 years)
Youth who use gateway drugs
Youth (adolescents)
Youth who are suicidal or physically or mentally disabled
Youth who are engaged in violent or delinquent acts

Terms collected arranged in broad groupings

Age

- Preschool (age 4 and under)
- Elementary youth (5-12)
  - ST Children (pre-adolescents)
- Youth (adolescents)
  - Jr. High Youth
  - Sr. High Youth (16-18)
- Young Adults (18-25 years) (19-25 years)
  - College students
- Adults (25-59 years)
- Elderly (60 + years)
- Student?

Gender

- Women
- Men

Sexual preference

- Homosexuals (males and females)

Racial/ethnic group

- African Americans/Black
- Asians and Pacific Islanders
- Caucasians
- Blacks
- Hispanics/Latinos
- Native Americans (Amer. Indians, Eskimos)

group by ability/handicap

- Disabled
- Handicapped/Disabled

Groups at high risk of drug use

- Children subjected to abuse and neglect
- Children and youth economically disadvantaged
- Children of alcoholic or other drug-abusing parents
- High-risk families
- High-risk families/youth (including COAs and ACOAs)
- High-risk youth
- Homeless or runaway youth
- Latchkey children
- School dropouts or those at risk of dropping out
- Single teenage mothers and their children
- Unemployed youth or those in danger of being unempl.
- Youth who use gateway drugs
- Suicidal or physically or mentally disabled youth
- Youth who are engaged in violent or delinquent acts

IV drug users

By profession or position

- A/D prevention professional
- A/D treatment professionals
- Administrator/Manager
- Attorney
- Clergy
- Community organization leader
- Community service groups
- Correction officer
- Criminal/juvenile justice
- EAP Practitioners
- Educator/teacher/trainer (specify grade[s])
- Health care providers (physicians, nurses, Pas, NAS, pharm.)
- Health care professional
- Judge
- Legislator
- Librarian/Information Specialist
- Media representatives
- Mental health professional
- Parole/Probation officer
- Police officer
- Policy makers/Administrators
- Recreation/Sports Personnel
- Reporter/writer
- Researcher
  - Biomedical researcher
  - Psychosocial researcher
- School Administrator
- Scientists and researchers
- Social service professional

By employer/employee relationship

- Employees
- Employer

Other groupings

- Patients
- Parents (specify age of child)
- General public
  - General public, concern for family/friend
  - General public, personal Concern
- Grantee
- Other
- Unknown/anonymous

One area conceptually refined

Groups at high risk of drug use

Suicidal or physically or mentally disabled
Persons from unstable or low-cohesion families
Children of alcoholic or other drug-abusing parents
SN Grown up or still under age
Children of single teen-age mothers
Persons subjected to abuse or neglect
SN Now or in the past
Persons subjected to abuse and neglect by parents
Latchkey children
Persons subjected to abuse and neglect by their spouse
Single teenage mothers
School dropouts or those at risk of dropping out
Unemployed or in danger of being unemployed
Economically disadvantaged
Homeless
Runaway youth
Gateway drug users
Persons engaged in violent or delinquent acts

The concept *Youth at risk of drug use* or any of its subordinate concepts (as specified by group and age range) can be produced by combination with *Adolescent*

An observation on defining groups by combination: For any group defined by personal/demographic characteristics, there are several derivative groups, e.g.:

Parents of members of the group
Children of members of the group
Spouses of members of the group
Teachers of members of the group

The thesaurus needs to make provision for forming such combinations.

Thesauri and ontologies in digital libraries

Tutorial

Part 2

Design, evaluation, and development

Part 2. Outline

Introduction and overview

Scope:
"Thesaurus" is used as shorthand for Knowledge Organization Systems (KOS)
Includes Thesauri, classifications, ontologies, taxonomies, concept maps, dictionaries, etc.

Main objective:

Participants should be able to crystalize the conceptual structure of a domain

Outline

The process of thesaurus construction

Developing the conceptual structure

The structure and processing of thesaurus data

The process of thesaurus construction

The overall process of thesaurus construction

Sources of concepts, terms, relationships, definitions
Methods of data collection

Merging data from many sources

The overall process of thesaurus construction

Diagram from DS 1974 copied in here. Need orig

Sources of concepts, terms, relationships, definitions

Reuse knowledge in existing Knowledge Organization Systems.
Much intellectual capital was invested in their development

But: Adapt content and structure to user requirements and background.

Most important source: search requests and other statements of user requirements.

Types of sources

(1) **Prearranged sources** (terms are already arranged according to some principle)

(2) **Open-ended sources** (terms are not ordered or terms must be inferred or derived)

Find machine-readable sources

Internal and external sources

Sources of concepts, terms, relationships, definitions

(1) **Prearranged sources** (terms are already arranged according to some principle)

(1.1) Descriptor lists, classification schemes, thesauri (universal classification schemes, such as LCC or UDC, and special classification schemes).

(1.2) Nomenclatures of single disciplines, esp. if approved by an international body.

(1.3) Treatises on the terminology of a subject field

(1.4) Encyclopedias, lexica, dictionaries, glossaries (universal or discipline-oriented; mono-, bi-, or multilingual).

(1.5) The tables of contents and indexes of conference proceedings, textbooks, handbooks, and course syllabi.

(1.6) Indexes of journals, abstracting journals, other publications, databases.

(1.7) Term-association lists produced by subjects in term association studies.

(1.8) Output from automatic classification programs based on term co-occurrence data or citations.

Sources of concepts, terms, relationships, definitions

(2) **Open-ended sources** (terms are not ordered or terms must be inferred or derived)

(2.1) Lists of search requests and interest profiles and other statements of user requirements obtained from search logs and user studies (individual interviews, focus groups).

(2.1a) Mooers' method: Focus group, present documents, ask "Why would this be of interest?"

(2.2) Descriptions of R&D projects and other activities to be supported.

(2.3) Free indexing of a sample of documents, each by several experts (to get synonyms).

(2.4) Titles, abstracts, full text, reviews of books, journal articles, conference papers, Web sites, internal documents, etc.

(2.5) For more information on individual terms: Web searches

Methods of data collection

For prearranged sources

If machine-readable, include all information, can always delete later

If not machine-readable and highly relevant, scan or have transcribed

Otherwise go through and select

May need to reformat for input to thesaurus software; use Perl scripts or word processor macros

Methods of data collection

For open-ended sources

Extract terms and **phrases** automatically, using a large general phrase dictionary, syntactic analysis, or a system such as http://www.nzdl.org/Kea/

Possibly use frequency data for further selection.

Extract term relationship automatically (often a feature of text mining programs).

Extract terms manually, being on the look-out for term relationships that can be inferred from text.

Merging data from many sources

Merge terms

Need to consolidate term variants

Use broad-scope sources to get more information on terms collected

Assemble synonym sets / concepts

Use ST relationships from many sources
Source 1: elderly ST aged person,
Source 2: aged person ST senior citizen

Merge relationships

Need to consider that often the same conceptual relationship is expressed in different terms

“Afterburn” collection from specialized sources to fill gaps

Developing the conceptual structure

Facet analysis 1: Education

More facet examples:
- **Yahoo Education (from Part 1)**
- **Job titles**

Facet exercise (in pairs)

Principles for meaningful arrangement

Rules for selection of concepts as descriptors. Rules for selection of terms

Facet analysis

Education (starting with classes from DDC)

Conceptual analysis and synthesis
in three steps:

Step 1. **Semantic factor compound concepts, make a list of elemental concepts.**

Step 2. **Arrange elemental concepts into facets.**
Arrange each facet in a well-structured hierarchy.

Step 3. **If needed, fit compound concepts into the framework of the hierarchy.**

Concept list for conceptual analysis and synthesis

(from Dewey Decimal Classification)

Note: A broader class is given in ( ), if necessary to specify the meaning of a term.

372.19	Curriculums of elementary schools
372.35043	Science in the elementary school curriculum
372.414	Methods of instruction for reading in elementary schools
372.72043	Arithmetic in the elementary school curriculum
373.19	Curriculums in secondary schools
373.243	Military schools (Secondary Education)
376	Education of women
376.63	Secondary education of women
378.19	Curriculum of colleges and universities
378.33	Fellowships (Higher Education)
371.911	Blind and partially sighted students
371.912	Deaf and hard-of-hearing students
371.95	Curriculums for gifted students

More facet examples

Job titles. Can you spot the facets?

Lawyer

Paralegal

Law office receptionist

Librarian

Library assistant

Library clerk

Physician

Physician's assistant

Doctor's office clerk

Ophthalmologist (eye doctor)

Ophthalmologic technician

Surgeon

Facet exercise (in pairs?)

Yahoo Health

Arrange the terms in front of you into meaningful groupings.

Use the blank strips to write a heading for each group.

Time: 30 minutes (leaving 10 minutes for discussion)

Principles for meaningful arrangement

Sequence and two-dimensional graphical arrangements (concept maps) can convey important information about concept relationships.

Collocate closely related concepts.

Often a principle of arrangement intrinsic to the subject matter suggests itself. The following examples and guidelines are intended to sharpen “informed intuition”.

Meaningful arrangement

Example 1

<*size: photograph formats*>	**size: photograph formats**
double whole plate	sixteenth plate
half plate	ninth plate
mammoth plate	sixth plate
ninth plate	quarter plate
quarter plate	half plate
sixteenth plate	whole plate
sixth plate	double whole plate
whole plate	mammoth plate
Art and Architecture Thesaurus	**Suggested meaningful sequence**

Alphabetical vs. meaningful sequence on same hierarchical level

Meaningful arrangement

Example 2. **Body systems. Fuller version**

XF body system or organ
XG . musculoskeletal system
XH . skin system
XJ . cardiovascular system
XK . respiratory system
XL . mouth, larynx, vocal organ
XM . digestive system
XN . urogenital system
XP . . urinary system
XQ . . reproductive system
XR . blood, immune system
XS . . blood
XT . . immune system
XU . endocrine system
XV . sensory system
XW . nervous system
XX . . nervous system structures and components
XY . . peripheral nervous system
XZ . . central nervous system

Meaningful arrangement

Example 3. Art genres

Trying to find a meaningful arrangement for a list of concepts often reveals a facet structure.

See the example in the tutorial notebook.

Graphical arrangement: Concept maps

See the examples in the tutorial notebook.

<art genres>

academic art
amateur art
apocalyptic art
art brut
children's art
commercial art
community art
SN Includes art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference to widely established standards. For art intended to beautify or enrich public places, use **public art**.
computer art
court art
crafts
cybernetic art
didactic art
dissident art
ethnic art
fantastic art
figurative art
folk art
funerary art
naive art
nonrepresentational art
primitive art
public art
SN Use for art whose purpose is to beautify and enrich public places. For art undertaken in conjunction with particular communities, usually to produce an effect or inspire response specifically within those communities, use **community art**.
rock art
cave art
serial art
sofa art
street art

a. **Original alphabetical sequence**

art genres

art genres by content or other intrinsic characteristics

figurative art
fantastic art
apocalyptic art
nonrepresentational art
cybernetic art
serial art
crafts

art genres by standard

academic art
folk art
dissident art

art genres by type of artist or origin

amateur art
naive art
art brut
children's art
computer art
ethnic art
primitive art

art genres by audience, purpose, or display context

sofa art
court art
public art
SN Art whose purpose is to beautify and enrich public places.
community art
SN Public art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference to widely established standards.
street art
rock art
cave art [prehistoric, esp. paleolithic]
didactic art
commercial art
funerary art

b. **Suggested meaningful sequence**

Figure 3. **Example from the Art and Architecture Thesaurus**

Concept map PHD

Concept map instr design

Meaningful arrangement

Guidelines

"Natural" principles

(1) Chronological – e.g., historical events.

(2) Evolutionary – arrange entities in the order they evolved, e.g., biological species, ideas.

(3) Sequence of steps – e.g., production processes, research methods, sequence of logical steps

(4) Increasing extension

(5) Geographical – spatial proximity.

More conceptual principles

(6) Increasing complexity (integrative levels)

(7a) From abstract to concrete or vice versa

(7b) From general to specific

(7c) From universal to local

(8) Canonical – an order given by an authority, e.g., books of a holy scripture

(9) Consistency of comparable subdivisions that appear in two or more different places

(8) Importance for indexing and query formulation

Rules for selection of preferred terms from a group of synonyms

Include in the thesaurus any term that falls in scope.
A large lead-in vocabulary is good!

Then select preferred terms.

The preferred term should

- be the best to reflect the meaning of the concept;
- be recognized in the user community;
- be unambiguous;
- be simple and short in spelling.

These criteria may conflict

Frequency data and occurrence in authentic sources can help in the selection.

Rules for selection of concepts as descriptors

The following criteria are helpful:

- Usefulness for searching and other functions;
- Are there alternative solutions:
  use a combination of descriptors,
  use a broader descriptor,
  consolidate with another concept to form a broader concept;
- Logical structure: is the concept needed as a heading?
- Frequency in indexing.

The structure and processing of thesaurus data

Interoperability of thesauri/ontologies. Crosswalks

The structure of a thesaurus/ontology database

The many forms of Knowledge Organization Systems (KOS) and their standards

Thesaurus software and its evaluation

Interoperability of thesauri/ontologies.

Crosswalks

Primary question:

- take a query formulated in vocabulary A,
- map the descriptors to vocabulary B,
- how good is the search in B as compared to using a query formulated in vocabulary B directly?

The answer determines searching compatibility.

Searching compatibility is directional, complex, and depends on the individual descriptors used.

Vocabulary A	**Vocabulary B**
Aircraft	Aircraft Airplane Helicopter
Military aircraft	Aircraft AND Military
Pest control Pesticides	Pest control (no narrower terms)

Insert index language page here

Insert compat figures here

The structure of a thesaurus database

Thesaurus data are relational.
Relational database is the most natural structure.

Many types of relationships – structure should not be restrictive. (See sample list in notebook.)

Examples of Synonymous-Term-type relationships

ST	Synonymous Term
ET	Equivalent Term
SP	Spelling Variant
AB	Abbreviation
FT	Full Term

Structure should allow for a **relationship to be the object of another relationship**
(for example, a scope note explaining the relationship)

Relationship strength

Appendix 2. Relationship types presently recognized by TermMaster

Note: This list is extensible by simply updating a table in the program and recompiling

Symbol	Meaning	Reciprocal	Group	Reference to
FN	Full form Note (If full form of term > 61 char)	-	SN	Text
SN	**Scope Note**	-	SN	Text
QN	Qualifier Note	-	SN	Text
HN	History Note	-	SN	Text
IN	**Internal Note** Expands on the external scope note, esp. reasons for term inclusion, term placement, and other decisions.	-	SN	Text
AN	**Action Note** Notes on actions to be taken on the term, such as look up definition, ask Ms. X, etc.		SN	Text
SQ	Source (for additional subset record)	-	SN	
SR	**Detailed source**	-	SN	Text
SI	Synonym, Internal	SI	ST	Term
SH	From non-hyphenated to hyphenated	SG	ST	Term
SG	From hyphenated to non-hyphenated	SH	ST	Term
SP	**Spelling variant**	SP	ST	Term
SB	Spelling British	SA	ST	Term
SA	Spelling American	SB	ST	Term
AB	**Abbreviation**	FT	ST	Term
FT	**Full Term**	AB	ST	Term
ST	**Synonymous Term**	ST	ST	Term

ET	**Equivalent Term**	ET	ST	Term
TR	Translation	TR	ST	Term
NA	Narrower of Facet	FA	NT	Term
NX	Narrower term of a broad category used in preliminary sorting	BX	NT	Term
NF	Narrower term - compound containing factor	BF	NT	Term
NM	Narrower Term - compound containing Modifier	BM	NT	Term
NC	Reciprocal of BC	BC	NT	Term
NT	**Narrower Term**	BT	NT	Term
NG	Narrower term - Generic	BG	NT	Term
NTT	Narrower term - Token	BTT	NT	Term
NPT	Narrower term - Partitive	BP	NT	Term
FA	Facet	NA	BT	Term
BX	Broader term for preliminary. sorting	NX	BT	Term
BF	Broader term - Factor	NF	BT	Term
BM	Broader term - Modifier	NM	BT	Term
BC	Broader term that might have NT to be used in combination	NC	BT	Term
BT	**Broader Term**	NT	BT	Term
BG	Broader term - Generic	NG	BT	Term
BTT	Broader term - Type of token	NTT	BT	Term
BPT	Broader term - Partitive	NPT	BT	Term
RC	Related term for combination (pop-up menu showing terms to use)	RD	RT	Term
RD	Inverse of RC	RC	RT	Term

RG	One-directional related term	RH	RT	Term
RH	Inverse of RG	RG	RT	Term
RN	Related term in scope note, generated by the program	RO	RT	Term
RO	Inverse of RN	RN	RT	Term
RT	**Related Term**	RT	RT	Term
EX	Excludes	EF	EX	Term
EF	Excluded From	EX	EX	Term
UN	Unspecified relationship	UN	RT	Term
HT	Homonymous Term	HB	HT	Term
HF	Homonym From	HT	HT	Term
ME	Meaning Equivalent	MF	HT	Term
MF	Meaning equivalent From	ME	HT	Term
BW	Broader Word	NW	RT	Term
NW	Narrower Word	BW	RT	Term
AF	Affects	AY	AF	Term
AY	Affected by	AF	AF	Term
PC	Precursor	PB	AF	Term
PB	Produced by	PC	AF	Term
RW	reacts with	RW	AF	Term
IB		IB	ID	
//	From a relationship to a term. Internal symbol TH	/ TI		Relation
/<	Inverse of // Internal symbol TI	// TH		Relation

The structure of a thesaurus database

Three levels

Level 1: Link term variants to terms

AST FT aspartate aminotransferase

GOT FT glutamate oxaloacetate transmaninase

(FT Full Term)

Level 2: Link terms to concepts

aspartate aminotransferase

ST glutamate oxaloacetate transmaninase

Level 3: Relate concepts to concepts

aspartate aminotransferase

BT aminotransferases

Levels 1 and 2 are often confounded.

The structure of a thesaurus database

Two models

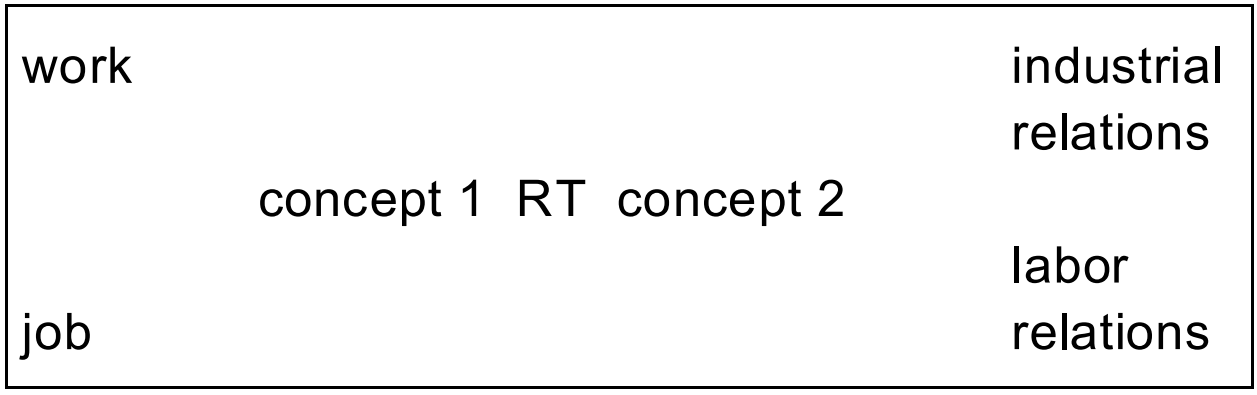
Concept-based model

Terms are mapped to concepts. This mapping expresses Synonymous Term relationships.

Concept relationships are expressed using concept identifiers.

Elegant, but in a multi-thesaurus database requires universal commitment to the term-concept mapping.

UMLS uses this model



The structure of a thesaurus database

Term-based model

All relationships are expressed as relationships between terms.

A concept relationship may be expressed in many ways, using different synonyms for each concept

Requires extensive processing to discover all concept relationships starting from a given concept.

job ST work industrial relations ST labor relations	job RT industrial relations work RT industrial relations job RT labor relations work RT labor relations

The many forms of Knowledge Organization Systems (KOS) and their standards

The purpose of standards

1 Input of thesaurus data into programs /
Transfer of thesaurus data from one program into another

1.1 Format for original input files (but XML difficult for that, use a more user-friendly format, such as TermMaster input formats)

1.2 Transfer from one thesaurus development program to another

1.3 Transfer from a thesaurus development program to an information system that uses a thesaurus for authority control, query expansion (synonym and /or hierarchic), display/browse/search, or other purposes

1.4 Transfer from a thesaurus development program to a thesaurus display / browse / search program

2 Querying thesauri and viewing results (for example, using Z39.50)

2.1 By people

2.2 By systems to use data from external thesauri for query term expansion etc.

3 Identifying specific terms/concepts in specific thesauri

This requires rules for URIs that uniquely identify specific term/concept records in specific thesauri. Probably requires some sort of name resolution service (such a thesaurus registry)

3.1 Links from one thesaurus to another

3.2 Indexing terms/concepts in the metadata for an object, or any other reference to a term/concept in a text/object

Standards that give a general format, leaving the user to develop specifics (e.g., relationship types) vs.

Standards that give specifics

The many forms of Knowledge Organization Systems (KOS) and their standards

Dictionaries

ISO 12200:1999, Computer applications in terminology--Machine Readable Terminology Interchange Format (MARTIF)--Negotiated Interchange

ISO 12620:1999, Computer applications in terminology--Data Categories.

Thesauri

ISO 2788-1986(E) / ANSI/NISO Z39.19-1993(R1998) (www.niso.org)

ZThes (using Z39.50, strictly ANSI Z39.19)

http://lcweb.loc.gov/z3950/agency/profiles/zthes-04.html)

Browser at http://muffin.indexdata.dk/zthes/tbrowse.zap

Vocabulary Markup Language (VocML) (under discussion at NKOS)

See also http://ceres.ca.gov/thesaurus/

ISO 5964-1985(E) (multilingual)

USMARC format for authority data

(http://lcweb.loc.gov/marc/authority/ecadhome.html)

Topic maps (reference works, encyclopedias) (http://www.topicmaps.org/about.html)

ISO/IEC 13250:2000 Topic Maps

XML Topic Maps (XTM) 1.0 (http://www.topicmaps.org/xtm/1.0/)

Concept maps

Classification schemes

USMARC format for classification data

http://lcweb.loc.gov/marc/classification/eccdhome.html

Ontologies

Knowledge Interchange Format (KIF) NCITS.T2/98-004

(http://meta2.stanford.edu/kif/dpans.html)

Ontology Markup Language (OML) /

Conceptual Knowledge Markup Language (CKML)

(http://www.ontologos.org/OML/CKML-Grammar.html)

Ontology Interface Layer (OIL) (http://www.ontoknowledge.org/oil/)

Generic standards for knowledge structures, entity-relationship models

Resource Description Framework (RDF) (http://www.w3.org/RDF/)

Open Information Model (OIM) (http://www.mdcinfo.com/OIM/) (Seems to be no longer active)

XTM might also fit here

Appendix B. The Zthes Abstract Model in XML
(from http://www.loc.gov/z3950/agency/profiles/zthes-04.html)

Appendix B.1. The Zthes DTD for XML

This DTD was supplied by Thomas Place. It is put forward not as a ``good'' XML representation of thesaurus information (whatever that might be construed to mean) but as a pragmatically valuable alternative encoding of the Zthes abstract record. Real Zthes data sets have been exchanged in the form of XML documents conforming to this DTD.

```
<!-- Zthes DTD
    Based on Z39.50 Profile for Thesaurus Navigation, version 0.1 (20 Feb 1999)
    Version of DTD: 25 Feb 1999 -->

<!-- #PCDATA: parseable character data = text
    occurrence indicators (default: required, not repeatable):
    ?: zero or one occurrence (optional)
    *: zero or more occurrences (optional, repeatable)
    +: one or more occurrences (required, repeatable)
    |: choice, one or the other, but not both
 -->

<!ENTITY % term "termId, termName, termQualifier?,  termType?, termLanguage?">
<!ENTITY % admin "termCreatedDate?, termCreatedBy?, termModifiedDate?,
termModifiedBy?">

<!ELEMENT Zthes (%term;, termNote?,  %admin;,relation*)>

<!ELEMENT relation (relationType, sourceDb?, %term;)>

<!ELEMENT termId              (#PCDATA)>
<!ELEMENT termName            (#PCDATA)>
<!ELEMENT termQualifier       (#PCDATA)>
<!ELEMENT termType            (#PCDATA)>
<!ELEMENT termLanguage        (#PCDATA)>
<!ELEMENT termNote            (#PCDATA)>
<!ELEMENT termCreatedDate     (#PCDATA)>
<!ELEMENT termCreatedBy       (#PCDATA)>
<!ELEMENT termModifiedDate    (#PCDATA)>
<!ELEMENT termModifiedBy      (#PCDATA)>
<!ELEMENT relationType        (#PCDATA)>
<!ELEMENT sourceDb            (#PCDATA)>
```

(This appendix should include a crosswalk with any pre-existing thesaurus DTDs if appropriate.)

Appendix B.2. Sample Zthes-in-XML Document

This document was supplied by Thomas Place.

```
<?XML version="1.0" ?>
<!DOCTYPE Zthes SYSTEM "zthes.dtd">
<Zthes>
    <termId>102067</termId>
    <termName>video art</termName>
    <termType>PT</termType>
    <termNote>
        Use for works of art that employ video technology, especially videotapes. For the study
        and practice of the art of producing such works, use "video."
    </termNote>
    <relation>
        <relationType>UF</relationType>
        <termId>102067/001</termId>
        <termName>art, video</termName>
        <termType>ND</termType>
    </relation>
    <relation>
        <relationType>BT</relationType>
        <termId>185191</termId>
        <termName>[time-based works]</termName>
        <termType>NL</termType>
    </relation>
    <relation>
        <relationType>RT</relationType>
        <termId>54153</termId>
        <termName>video</termName>
        <termType>PT</termType>
    </relation>
    <relation>
        <relationType>RT</relationType>
        <termId>253827</termId>
        <termName>video artists</termName>
        <termType>PT</termType>
    </relation>
</Zthes>
```

Dagobert Soergel ds52@umail.umd.edu http://www.clis.umd.edu/faculty/soergel/

Elements of an XML thesaurus data specification

This schema is parsimonious yet allows the recording of many types of data. It gives enough information to derive a full XML specification.

This spec assumes that data from each source are grouped, so that source attribution is not needed for each element; otherwise the structure would be much more complex. This works for a communications format but not for an internal database format.

The term itself is indicated in a relationship of type TERM. This allows for terms in multiple languages for the same concept and simplifies the schema since elements in *term* would be the same as in *relationship target*.

Addition of the *scope* element was inspired by the Topic Map Standard (see http://www.topicmaps.org/xtm/1.0/)

The scheme needs a method for indicating a relationship set defined elsewhere and used within the source or for defining a relationship set for the source.

Default is minOccurs="1" maxOccurs="1"

Source (minOccurs="0" maxOccurs="unbounded")
- Pointer to or definition of relationship set used
- Unit: Concept or term or group of terms (minOccurs="0" maxOccurs="unbounded")
  - Unique identifier
  - Hierarchy position (minOccurs="0" maxOccurs="unbounded")
    - Hierarchical level
    - Class number / notation
    - Scope for which this concept/term holds (minOccurs="0" maxOccurs="unbounded")
  - Relationship (minOccurs="0" maxOccurs="unbounded")
    - Relationship type
    - Relationship target
      - /* See below for structure. */
    - Relationship strength (minOccurs="0" maxOccurs="1")
    - Audience level /* Of this relationship */ (minOccurs="0" maxOccurs="unbounded")
    - Perspective /* Of this relationship */ (minOccurs="0" maxOccurs="unbounded")
    - Scope for which this relationship holds (minOccurs="0" maxOccurs="unbounded")
    - Relationship, added information (minOccurs="0" maxOccurs="unbounded")
      - /* This could be a scope note explaining the relationship, an image illustrating the relationship, another term, etc. */
      - Type of added information /* Relationship types might be reused here. */
      - Relationship target
      - Audience level /* Of this piece of info. */ (minOccurs="0" maxOcc="unbounded")
      - Perspective /* Of this piece of information */ (minOccurs="0" maxOcc="unbound")

Where relationship target has this structure (unifying term, text, images, multimedia document)

Relationship target
- Type
  - /* Includes types of terms (descriptor, other preferred term, non-preferred term and types of texts and other documents, may be an elaborate hierarchy. */
- Target value (a term or a document)
  - Term
    - Term variant (minOccurs="0" maxOccurs="unbounded")
      - Type of variant
        - /* Such as Preferred Spelling, other SPelling, ABbreviation, Full Term. */
      - Term form (complete term or Stem plus suffix)
        - Complete term
        - Stem plus suffix
          - Stem
          - Suffix
  - Document
- Language (zero to many, exactly one for terms)
- Audience level /* Of this relationship target */ (minOccurs="0" maxOccurs="unbounded")
- Perspective /* Of this relationship target */ (minOccurs="0" maxOccurs="unbounded")
- Scope for which this/term holds (minOccurs="0" maxOccurs="unbounded")

Thesaurus software and its evaluation

Different types of software

- Thesaurus management software specifically
- Concept mapping software
- Ontology editors
- Description-logic- based software

Thesaurus software selection criteria

General criteria for evaluation of software

Customizable?

Special functions of thesaurus management

A General system parameters

Multiple thesauri? Multiple languages?
Relationship types supported

B Input and editing (batch and online)

Preserve arrangement?

C Output in various formats

Nicely formatted hierarchical displays, concept maps. Web

Map detailed internal relationship types to less detailed external

D Processing of data

Check or create reciprocal relationships.
Create notations

Dagobert Soergel ds52@umail.umd.edu http://www.clis.umd.edu/faculty/soergel April 2002

Requirements for Thesaurus Management Software.

Criteria for Evaluation

Outline

General criteria for description and evaluation of software

Special functions of thesaurus management

- A General system parameters
- B Input and editing (of input data files and online)
- C Output in various formats
- D Processing of data

General criteria for description and evaluation of software

Only a few points that are especially important in connection with thesaurus software are dealt with here.

Database management system used. Is it easy to produce tailor-made output performance.

Efficiency of storage

Version control. Does the program keep track of all changes

User interface

- Menus versus commands. Use of function keys, etc.
- Use of windows
  - Window positions fixed in program
  - Window positions on the screen can be specified by user
- Navigation possibilities (see editing)
- Program asks for verification before actually recording a change in the thesaurus database.
- Consistency of the user interface
- Help

Case sensitivity. Are upper and lower case treated the same or different in sorting and retrieval? If the same, is this true for all characters or are there exceptions (for example, in Index 4.1 sorting is different for upper and lower case umlauts).

Note: Case is often important to distinguish words, e.g. turkey and Turkey. If case insensitive, need turkey (bird), Turkey (country)

User influence on how the program works.

. The user can influence the program behavior through data input without changing the program itself.

. . The program reads parameter from a file (possible from the line (s) at the beginning of an input file), that can be modified by the user.

. . Program uses external files that can be changed by the user.

. . Program accepts specifications written by the user (e.g. specification of a record structure through giving data fields) (example: database system).

. . The user can change menus, error messages, help messages, etc.

. The program itself can be modified according to user wishes

. . Program change through the user himself or herself (source code available)

. . Program modification only through the producer

. . Effort needed for changing the program (this depends on the modularity of the program and the programming technique used. Example: in the program language C constants such as the maximum length of a term or the character used to mark a line as bold can be defined in a header file. To change these constants one needs only to change the header file and then compile the program anew, which could be done by a properly instructed non-programmer.)

Special functions of thesaurus management

Note.

For all parameters and functions of the program being evaluated the question arises how much the user can influence it. This criterion is always applicable and is explicitly mentioned only in special cases. For example, one should know whether the user can define term types, relationship types, etc. One should keep in mind, however, that many such values have a semantics which must be operated on by the program. For example, if the program has the ability to construct an overall hierarchal structure by binary NT relations, the parts of the program doing this function must use all NT-type relations, and only those. If the user defines a new relation that is a special case of NT then this can become complicated.

Whenever there are user choices, the system should provide default values so that the user who has no special requirements can use those defaults without further ado and need not concern herself with the choice of parameters and the methods for changing the parameters.

A General system parameters

Types of vocabularies supported

Remark: The following types of vocabularies overlap considerably

. Vocabularies used primarily for information retrieval
. . Classifications and thesauri
. . . Thesauri without a well-structured classification
. . . Well-structured classification
. . . Concept map
. . Topic map (relationally rich thesaurus)
. . Indexes for books or journals
. . Record filing scheme
. . Data dictionary (in systems analysis and software development)
. Nomenclatures and taxonomy (chemistry, biology, etc.)
. Dictionaries or lexica, general or special
. . Mono- or multi-lingual dictionaries
. . . Mono-lingual dictionaries
. . . Multi-lingual dictionaries
. . Glossaries
. . Lexica
. . Picture dictionary

Thesaurus database as a whole

. Number of thesauri in a thesaurus database
. . One thesaurus per database
. . . One of several thesauri being worked on can be specified when calling the program (but each thesaurus is stored in its own database)
. . Multiple thesauri integrated in one database
. . . Number of thesauri that can be included
. . . Only thesauri which are subsets of one unified thesaurus (micro-thesauri within one large thesaurus), or really different thesauri?
. . . All thesauri on an equal footing or one main thesaurus with connections to terms of other thesauri
. . . Are there relationships between terms from different thesauri? How are these relationships determined?
. . . . Derived from the structure of the database

. . . . Through reference to a "switching language"

. . . . Through direct bilateral relationships between pairs of thesauri

. . Marking subsets in a single thesaurus (notations are the same across subsets)

. Is there a starting database of terms and concepts that can be processed by the program?

. Languages that can be processed: number of languages and list of languages. (This is relevant for functions that depend on the language such as normalization of plural forms to singular, decomposition of terms that include several roots - multi-word terms in English, composite words in German, spell checking, or use of a stop word list.)

. . All languages on a equal footing

. . One main language

. Subjects that the program can work on: number and list (This is relevant for spell checking and possibly for operations that use certain structural properties of the terms in a special subject.)

. Maximum number of terms

. Stop word list

. . For data input (for example for the decomposition of terms that contain multiple roots, in English these are usually multi-word terms.)

. . Additional stop word list for KWIC or KWOC Index

. . Can the stop word list be changed by the user?

. Does the program support hierarchical arrangement?

. . Maximum number of hierarchical levels

. . Does the program preserve sequencing on same level of the hierarchy (see below)

Other characteristics of the system as a whole

. Code lists for various types of data (term types, relationship types, languages, etc. that are used for checking input and/or for presentation of menus. Can the user change these lists?)

Data that can be given for each term and for relations between terms

Note: This list is just a small subset of all the data that might be needed by varied applications.

. Maximum term length (Recommended at least fifty, especially if there are many multi-word (or multi-root) terms and long names. Also important for input of source term lists that have long terms.)

. . Maximum defined by the system

. . Maximum can be defined by the user (within system limits) (This is needed if a thesaurus is produced for an ISAR system that has its own maximum term length.)

. . . Is it possible to define a separate term length for each of multiple thesauri integrated in a thesaurus database

. Treatment of homonyms. How are the separate meanings of homonyms identified.

- Language of the term. Maximal length of the language indication. Does the thesaurus use a standard list of language symbols (In a multilingual thesaurus databases indication of the language is necessary for the unique identification of a term.)
- Sort form (if different from display form)
- Part of speech for a term
- The gender of a term
- Other syntactic or morphological data
- Language level (day-to-day language, discipline specific language, outdated, etc.)
- Indication of whether this term may participate in relationships to other terms
- Term types (See attached list for examples)
  - Term types predefined in the system: number and list
  - User definable term types: how many
  - Can a separate list of term types be defined for each thesaurus included in an integrated thesaurus database?
- Perspective, a value that can be used for selecting terms into lists (Index 4.1)
- Marker, another value that can be used to select terms into lists (Index 4.1)
- Notation
  - Coarse notation (for example, for identifying broad subject groupings or facets)
  - Detailed notation (can at the same time fulfill the functions of a coarse notation)
  - External notation
  - Internal notation (for example, a notation expressing the hierarchical structure to be used by a retrieval program for inclusive searching)
  - For each kind of notation: maximum length (can the maximum length be specified by the user?)
  - Can the user specify whether a descriptor can have several or only one notation (MeSH, for example, has for each descriptor as many notations as the descriptor has places in the parley hierarchy.)
  - How much influence does the user have on the form of the notation
  - Support for the generation of notations
- The sequence of the terms on the same level of a hierarchy can be stored (This can be implemented through notation)
- Relationship types (See attachment for examples.) (At a minimum, thesaurus software should support the relationship types specified in thesaurus standards.)
  - Relationship types predefined in the system: number and list
  - Relationship types that can be defined by the user: number (But see note at the beginning.)

. . . Can the user define/change the rules used by the system in processing relationship types?

. . Possibility of specifying many detailed relationship types in the database but map these to a few general relationship types in the user version

. . Can the relationship type names for the user version be freely chosen

. Rules for relationship types (Rules serve for consistency checking but can also introduce unnecessary restrictions.) Examples for rules:

. . Synonym relationship always from descriptor to nondescriptor

. . Abbreviation relationship always from descriptor to nondescriptor

. Data about relations

. . Strength of connection

. . Aspect used in establishing the relation. For hierarchical relationships: The characteristic of subdivision (However, it is preferable to create an own heading for each characteristic of subdivision to group all the narrower terms that correspond to that characteristic.)

. . Qualification through context (that is, the relation is valid only for a certain context, or in any case the connection strength is dependent on the context. Put differently, the relation is itself an object related to another object, such as a term.)

. . Scope note for a relation. Explains why the relation was introduced.

. . In what output formats should the relation appear (This does not refer to the relationship type, but to the specific relation between two terms)

. Maximum number of relations of a given relationship type that can be given for a term (This may differ from one relationship type to another.) (For example, some systems allow only one BT; this is not good, since mono-hierarchy is too restrictive.)

. Maximum number of relations for a term altogether

. Is it possible to establish two relations of different types for an ordered pair of terms (for example, NT as well as RT or ST as well as RT)? (There are cases where it makes sense to have two relations coexisting.)

. . In a single thesaurus

. . In the integrated thesaurus database

. Scope note and other text information

. . How many types of text information (for example, is it possible to have internal notes)

. . How many notes of each type of term

. . Maximum text length

. . Can descriptors inside a scope note be marked and treated specially?

. For terms, notations and/or relations

. . Status value (Such as *included in present edition*, *kept for later decision*, *deleted*. The *deleted* status is important so that decisions on the same term must not be made again

when, for example, this term appears in a newly processed source; it is also needed in order to reconstruct the state of the thesaurus at the time of indexing a given document.)

. . Source indication

. . . Maximum length of the source indication

. . . Maximum number of sources of a term or relation

. . Date indications (Dates for various events such as inclusion in the thesaurus database, inclusion in a given thesaurus, approval by an editor, deletion from the thesaurus, etc.)

. . Frequency of use (in a system that indexes with weights: Frequency of use with weight 2, frequency of use with weight 1 or 2) (Keep in mind that one always must specify the frequency with a time span.)

. . Indication of the editor/lexicographer and reviser

. . Editing history (edited when and by whom, revised and approved when and by whom)

. Data on the sources as such (Does the program allow for a directory of sources?)

. Other kinds of data provided for in the program

. Can the user define additional types of data? In what limits? (Since many data about a term can be given through relationships the possibility of defining additional relationship types is important.)

. Data Structure

B Data input and editing

Data input

. Batch input

. . Batch input of other thesauri

. . Batch input of thesaurus files that have been created with a word processor or otherwise. (In many cases this is the most efficient method of imputing data. This method also allows editors to work independently from the program wherever there is a computer.)

. . . Format(s) of such input files

. . Command structure that allows for scheduling the input of several files in sequence (this is important because the input of a file may take a long time. With such a command one can input several files over night without intervention.)

. Online data input (see also online editing)

. . Input of individual terms and data about them

. . . Input of term and data about the term in one step

. . . . Online form for all data about a term. Details about this form (for example, are there fixed fields for relationship types or is the relationship type given through an explicit name, scope note as one continued text for a number of lines, scrolling if not all information fits on one screen)

. . . . Script: The system prompts for the various data for a term in a fixed sequence. Is the content and the sequence of these prompts defined by the system or definable by the user?

. . . . How does the system treat cross-terms that have not yet been entered as main terms

. . . Input of terms and relations in separate steps

. . . Are all data about a term shown on the screen once input is completed? Can they be modified at that point?

. . Input of whole lists, especially hierarchies, that have been composed on the screen under thesaurus program control (Functionally this is very similar to batch input of thesaurus files as discussed.)

Editing

. For the selection of the terms to be edited in an editing session and for the format of display of the data to be edited see the criteria under Output

. General functions in editing (Some of these are also important for input.)

. . Effort for different types of changes

. . Consistency check for changes made (see consistency check under D)

. . Is the user asked to verify the change?

. . Can changes be made with "hierarchical force"? (E.g., deleting a broad term and all its narrower terms.)

. . Does the system give a message if the user enters a term or relation that was considered earlier and either rejected for inclusion or deleted after it was once included?

. Types of changes. For each type: How much effort

. . Changes for terms

. . . Adding a term

. . . . Specifying of the position of the new term in the hierarchical sequence (the input of a BT relation alone is not sufficient if one wants to maintain a meaningful sequence of terms on the same level)

. . . Deleting a Term

. . . . Are all relations deleted as well (or at least not output any more? Possibility differentiated by output formats for editing and output formats for the user version)

. . . . Is there a consistency check after a term was deleted? Especially the effects on the hierarchy need to be checked. It is problematic to delete a descriptor that has narrower descriptors that ought to be kept. Some systems do not allow deletion of a term that is linked to other terms through relations; the editor must first delete these relations.

. . . Adding a term that was deleted earlier

. . . . Are the relations that were in the system while the term was still there also added automatically?

. . . Change in term type (especially from descriptor to nondescriptor and vice-versa)

. . . Replace one term through another

. . Notation changes

. . . Are other affected notations automatically changed accordingly (important especially when a term is added at a given position)

. . Changes in relationships

. . . Adding a relationship

. . . Deleting a relationship

. . Global changes (for example, add EN to all terms in the thesaurus database if one wants to change from an English-only thesaurus database to a multilingual database)

. Batch Editing

. . File of editing commands

. . The program produces a file for editing (as part of its output functions). This file can be edited and re-input (All data in the file for editing where given a temporary *deleted* status. For any data not contained in the edited file, that *deleted* status becomes permanent.)

. . . Format of the file for editing (for example, Generic Word Processor format or a format that can be used by an outline processor)

. . . See also online editing and output regarding the criteria for selection of terms and the display format

. Online editing (most of the functions given here apply also to online input)

. . History functions

. . . Is navigation history kept? Can the user retrace steps?

. . . Complete transaction log for error recovery?

. . Manipulation of lists of terms that must undergo editing

. . . Editing lists can be stored and recalled

. . . . During one session

. . . . From one session to the next

. . . . Editing lists named by the user or by the system (For example, in Index 4.1 an editing list, as given in a window, is identified by the coordinate of the left upper window corner.)

. . . Navigation in the editing list

. . . . Screen by screen

. . . . Scrolling

. . . Deleting elements from an editing list

. . Navigation in a batch of forms

. . Switching between editing lists and editing batch of forms

. . Editing data for an individual term

. . . Editing data about an individual term in a list

. . . . Which data are displayed (see C)

. . . . Which data can be edited (These editing changes can be changes to the database or they can be changes that influence further editing, such as marking a term as processed or moving a term to another list.)

. . . . Can new terms be input while working on an editing list?

. . . . Does system display available options (for example, when working on BT relationships, the system might display a list of the terms that would be legal and the user would select; see consistency checks)

. . . Editing data for a term on an online form (most systems would always allow input of new terms in this context by having the user request an empty form)

. . . . Screen format and editing options (for example, is it possible to do full screen editing as in a word processor using the general keys like arrows and delete, can text be copied from one place to another, from one form to another, can scope notes be edited as continuous text, mouse support.)

. . . . Function for exchanging descriptor with one of the synonyms

. . . Jumping to a cross-referenced term, editing it, and returning to the term previously worked on (possibly do this multiple steps)

. Editing entire structure, especially a section of a hierarchy, without detailed data for each term. This is functionally equivalent to editing and re-input of an editing file in hierarchical format as discussed above, but may be more convenient.

. . Functions offered for editing (for the editing of hierarchies the functions of outline processing are especially useful)

Reports on inconsistencies (For example, relationships to a nonexisting term) in a form that facilitates the input.

. Batch

. Online

Reports on changes, especially if there is a procedure for the edition and final approval.

C Output

Note: Output can be for human use, either printed or online, by thesaurus users or for editing, or for use by another system. Furthermore, many of the functions/criteria discussed here apply also to the selection of a group of terms for online editing. This includes the selection and sequencing of terms to be edited online, the data displayed on the screen, and the extent to which the user can control these parameters.

General criteria for all output functions

(One and the same thesaurus management program can have different values for different output formats.)

. Domain of the output

. . An individual thesaurus (either the only thesaurus in the database or an individual thesaurus from an integrated database)

. . Terms that appear in multiple thesauri

. . . User can specify a list of thesauri

. . . Concordance

. . . Comparison print: a printout that shows how the terms occurring in one or more source thesauri are dealt with in a target thesaurus, highlighting especially terms missing from the target thesaurus

. Selection of terms from the domain (Many of these criteria are important especially for editing.)

. . Scope in a hierarchy (identified by beginning and ending notation or all terms under a broad term)

. . Selection by relationship to another term or object

. . Selection by facet

. . Selection by hierarchical level

. . Scope in alphabetical sequence (identified by beginning and ending term)

. . Selection by status

. . Selection by markers or perspective

. . Selection by absence from a given thesaurus. (This is important for editing: If a new source is added to the thesaurus database, check terms absent from the thesaurus being worked on to see whether they should be included.)

. . Selecting terms that are not yet revised and approved

. . Select terms not included in the last printed or otherwise published version

. . Selection by language

. . Selection by string pattern contained (free text searching). How powerful are the possibilities for defining patterns (wild cards for characters, for strings, etc., phrase searching vs. just word searching, etc.)

. . Selection by internal term number (record number)

. . Selection by specific notation

. . Selection by a boolean combination of the criteria

. . Selection of a small list by marking terms in a big list

. Sequencing of the selected terms for presentation (this is important to achieve a meaningful sequence for editing)

. . Hierarchical sequence

. . . Stored hierarchical sequence (usually implemented through notation)

. . . . If the domain includes several thesauri: Can the editor select one thesaurus as a guide that will determine the hierarchical sequence?

. . . Hierarchical sequence generated by the program based on hierarchical relationships (This usually implies alphabetical sequence of the children under the same parent.)

. . Alphabetical sequence

. Determining the entry point for the list

. Method for calling up a list (This may be different for the different selection criteria. For example: Index 4.1 the editor working on a term can position the cursor on the facet field and call up a list corresponding to the value; when the facet field for the term being worked on has the value "Person" then the list called up includes all terms from the facet "Person".)

. Content and format of the output (for screen forms and for lists) (For each criterion: how much control does the user have?)

. . Data for each term

. . . Suppressing relationships that are shown through arrangement (especially suppression of hierarchical relationships that are shown through the sequence and indication of the hierarchical level)

. . Differentiation of relationships types

. . Symbols for relationship types

. . Sequence of data and relationships for one term

. . Sequence of the cross terms within the same relationship type

. . Are cross terms shown with their notation

. . Are cross terms that have narrower terms identified (for example, by a plus before or after the notation or before or after the term) (This is important because the searcher or indexer should check to see whether one of the narrower terms is more suitable than the cross term.)

. . . If yes, is this indication fixed by the system or selectable by the user? Is the symbol chosen (in the example plus) user selectable?

. . Orientation aids for the user (such as giving the first and last term on the page in an alphabetical list or the first and list notation on a page in a hierarchical list)

. Number of languages presented in the output format

. . Monolingual thesaurus

. . Multilingual thesaurus

. . . Parallel arrangement with a column for each language

. Specification of the output format

. . Only predefined formats (The evaluation of a thesaurus management program should include detailed descriptions and sample pages of these redefined formats.)

. . Specification of the output format through the user

. . . Specification online. Can the resulting specification be stored and recalled under a name?

. . . Specification through a special specification file that can be produced with a word processor

. . . How complex is this specification (this must be seen in relation to the number of formatting options offered)

. . . How compact is the specification

. . . Does the program come with predefined formats or specification files which the user can simply use as is or modified, which would be less work than creating these files from scratch. (Include in the evaluation detailed description and sample pages of these redefined specifications.)

. Possibility to order several outputs at the same time (e.g., for overnight processing)

Printed thesaurus for public use

Note: Many of the format specifications listed here apply also to online displays, particularly Web displays.

. Printing methods supported: especially laser printer support (for example, through output of a file in the format of a word processing or desk top publishing program), Photo Type Setting Support. File with general markup language

. . Formatting into pages, especially considering proportional fonts and different font sizes

. . Formatting into columns

Note: Formatting into pages or columns important for producing orientation aids for the user

. Can the output file be edited before printing?

. Output formats

. . Hierarchal lists of terms

. . . Sequence of the hierarchy, see above

. . . Specificity of the hierarchical list

. . . . Hierarchical outline

. . . . Hierarchical list of all terms

. . . Degree of detail of the hierarchy

. . . . Quick hierarchical list

. . . . Annotated hierarchical list

. . . Method for showing the hierarchical level

. . . . Showing the hierarchical level through indention

. . . . . Indentions with a special symbol (for example, a dot) for each level

. . . . . With additional explicit indication of the hierarchical level

. . . . . Indention, type size, and normal/bold as a function of the hierarchical level

. . . . . Maintaining the hierarchical context through repeating the hierarchical change at the beginning of each (left that is even)

. . . . Hierarchy without indention with explicit indication of the hierarchical level, especially for two or more column printouts

. . Graphical representation of conceptual relationships (concept maps, topic maps)

. . Alphabetical lists of terms

. . Alphabetical index

. . . KWOC index

. . . . KWOC index in which the access words are normalized to singular form

Online search for navigation in the thesaurus using the Web or the program itself (also important for editing)

. Web files

. . Generation of hyperlinks and anchors for jumping from an outline to a quick hierarchy to an annotated hierarchy and for following relationships

. . Explorer-type expandable hierarchy

. . Control over partitioning the thesaurus to get Web files of reasonable size

. . Capability for showing coordinated windows on the Web

Files for communicating thesaurus data to retrieval systems (such as DIALOG or BRS) or to other thesaurus management programs

. Files compliant with a given standard, for example ZThes

. Files that ca be input into a database system for searching the thesaurus. If the database is Web-enabled, this can be combined with thesaurus Web files.

Change reports

. Report of changes since a given date

. Report of changes since the last printed or otherwise published edition

Statistical reports (Number of descriptors and entry terms, number of descriptors in each major class, number of descriptors on each hierarchical level, number of each type of relationship)

D Processing of data (consistency checks, etc.) through the program

In general: how much support does the program offer in the processing and generation of data (for example, constructing a hierarchy from BT/NT relationships, generation of notations)? The other way around: To what extent is the program limited to managing the data input by the user.

Checking input data for formal correctness (in batch input or during online editing)

. Checking the term length for main terms and cross terms

. Checking the relationship symbols, term type symbols, language symbols, etc.

. Checking for illegal terms in a hierarchy (A jump by more than one level down is illegal.)

. Checking completeness (for example, checking whether a notation is given for a term when one is required)

. Spell check

Consistency checks (during batch input and online editing)

. General characteristics of consistency checks

. . Is the check mandatory or user selectable

. . Force of the check (maybe be different for different kinds of check)

. . . There is no way to input inconsistent data

. . . Merely a warning to the editor

. Consistency checks for terms

. . Check for form of term

. . . Check whether the term agrees with the rules of form established for the thesaurus (for example, preference for singular, preference for nouns over adjectives or verbs)

. . . Singular/plural check (whichever is preferred in the thesaurus) or conversion

. . . Spell check

. . . Capitalization check or conversion (Some terms must always be capitalized; can this be enforced, for example by having these terms in the thesaurus database?)

. . Duplication check for terms

. . . Does duplication check consider singular and plural as the same? (In an integrated database both can appear.)

. . . Does the duplication check consider variant spellings the same?

. . . Does the duplication allow the editor to take care of homonyms

. . . Can the program handle identical strings of characters that denote separate terms in different languages?

. Consistency check for relationships

. . Duplication checks for relationships

. . Check for reciprocal relationships and creation of reciprocal relationships where needed

. . Check for several relationships between the same ordered pair of terms (If this is not allowed, it should be checked, but only within an individual thesaurus.)

. . Check for a relation of a term to itself

. . Check for terms that are not preferred terms used as cross term in a concept relationships. Alternatively, replacing a term through the appropriate preferred term when producing output

. . Check for ST-type relationship from descriptor to descriptor

. . Check the consistency of hierarchical relationships

. . . Check for hierarchical relationships that jump a level, for example, A NT B, B NT C, A NT C

. . . Check for hierarchy cycles, for example, A NT B, B NT C, C NT A (Such cycles could throw the program for a loop in the generation of a complete hierarchical structure from hierarchical relationships.)

. . Check for incomplete relationships, for example, semantic factoring with only one semantic factor

. . . Check for terms that do not participate in any relationship (orphan terms)

. . More complex checks of the semantic consistency of a relationship (example for a rule: hierarchical relationships are allowed only between terms belonging to the same category (we do not say here whether this rule is good or bad). ST-type relationship only in some language if TRanslation relationship is used from one language to another (using ST-type relationships regardless of the languages involved might actually be better. Formal ontologies defines properties of concepts such that only concepts that agree in these properties can be hierarchically related. For example, for a concept that represents a class of objects, each instance has identity, but for a concept that designates an amount this is not the case.

. Check whether input data conform to the field or relationship rules (For example, some relationship may be valid only to places, so the cross term must be a place name.)

Support in the editing of terms

. Normalization of terms to singular (while storing the original form)

Generation of notations

. The editor can input some or all notations, but where notations are missing the program generates them through hierarchical extension.

. Format of the notations generated

Support for the processing of relationships, for example

. Support for the generation of relationships, for example

. . Extracting single words from a multi-word term and presenting them as candidates for semantic factors

. . Presenting candidates for semantic factors based on hierarchical inheritance from the broader terms

- Generating hierarchical relationships from an input list in hierarchical format
- Conversely, generation of a hierarchical sequence from binary hierarchical Relationships
- In an integrated thesaurus database: use of synonym relationships in detecting the identity of conceptual relationships, for example Thesaurus 1: A BT B; Thesaurus 2: A BT C; any thesaurus: B ST C; conclusion: the two hierarchical relationships are the same conceptual relationships.

Thesauri and ontologies in digital libraries

Tutorial

Resources

Resources

A brief bibliography and a few examples of directories of thesauri and dictionaries on the Web. The tutorial Web site has more resources

Web site: http://www.clis.umd.edu/faculty/soergel/dlthestut

Printouts from the following Web sites were included in the paper tutorial notebook:

www.darmstadt.gmd.de/~lutes/thesoecd.html Web Thesaurus Compendium (representative list with descriptions)

www.onelook.com OneLook Dictionaries. The Faster Finder

www.yourdictionary.com/

www.strategic-road.com/pratique/dicofr.htm Strategic Road Dictionaries

www.emich.edu/~linguist/dictionaries.html

www.mikesart.net/giantglossarycom Terminology - Search

www.asel.udel.edu/natlang/nlp/lrd.html The Language Representation Database Project

http://nkos.slis.kent.edu

Networked Knowledge Organization Systems (NKOS). Has a workshop at JCDL

www.ukoln.ac.uk/metadata/desire/classification/

The role of classification schemes in Internet resource description and discovery

www.verity.com/products/k2developer/index.html

www.excalib.com/products/rw/rwarchitecture.shtml Excalibur RetrievalWare™

Thesaurus software Web sites

http://www.willpower.demon.co.uk/thessoft.htm

http://sky.fit.qut.edu.au/~middletm//cont_voc.html

http://www.fbi.fh-koeln.de/fachbereich/labor/Bir/thesauri_new/indexen.htm

http://bak-information.ub.tu-berlin.de/software/term.html (covers a wider range of software, annotations in German)

Concept mapping resources

at http://158.132.100.221/CMWkshp_folder/CM.ResFolder.html

(Educational Development Resource Centre, Hong Kong Polytechnic University)

Many links to concept mapping programs and other resources

Ontology editor example: Ontolingua editor, accessible through http://WWW-KSL-SVC.stanford.edu:5915/doc/network-services.html

On **description logic** see, for example http://potato.cs.man.ac.uk/seanb/publications.php

The URLs given on the standards page are also useful more generally

Search terms for a Web search for thesauri etc.

(ontolog* OR classification* OR Klassifikation* OR taxonom* OR thesaur* OR dictionar* OR dictionnaire OR Woerterbuch OR glossar* OR glossaire OR "word list" OR lexicon OR lexique OR Lexik* OR terminolog* OR vocabulaire OR vocabulary OR "knowledge organization" OR "knowledge structure" OR "authority list")

Possibly add OR concept OR mot-clé OR keyword OR "subject heading" OR definition

It is best to require these terms in the title. Otherwise there will be a lot of irrelevant material retrieved, especially by the term *classification*.

Short bibliography

Website: http://www.clis.umd.edu/faculty/soergel/dlthestut

Basic information retrieval and classification concepts

Soergel, Dagobert, 1985

Organizing Information. Principles of data base and retrieval systems.

Orlando: Academic Press, 1985. 450 p.

Vickery, Bryan C.

Faceted classification.

London: Aslib, 1970.

Thesaurus textbooks

Soergel, Dagobert

Construction and maintenance of indexing languages and thesauri

New York: Wiley, 1974. 632 p.

Lancaster, F. Wilfrid

Vocabulary control for information retrieval. 1.ed.

Washington, D.C.: Information Resources Press, 1986. 233 p.

(2. ed. not as good)

Aitchison, Jean; Gilchrist, Alan; Bawden, David

Thesaurus Construction and Use : A Practical Manual. 4. ed.

London: Fitzroy Dearborn, 2000. 230 p.

Also watch for the Proceedings of the ASIS SIG/CR Classification Research Workshop, published as **guidelines for the establishment and development of monolingual thesauri** by Information Today

Standards (use with caution) (see the section on Standars in Part 2 of the tutorial)

National Information Standards Organization

Guidelines for the construction, format, andmanagement of monolingual thesauri.

Bethesda, MD: NISO Press; 1993. ANSI/NISO Z39.19-1993. Borrows heavily from

International Organization for Standardization.

Documentation--guidelines for the establishment and development of monolingual thesauri. 2. ed. Geneva: International Organization for Standardization; 1986.

International Standard ISO 2788-1986(E).

International Organization for Standardization.

Documentation--guidelines for the establishment and development of multilingual thesauri. Geneva: International Organization for Standardization; 1985.

International Standard ISO 5964-1985(E).

Machine-Readable Dictionaries and Computational Linguistics Research

Walker, Don, ed. 1995; Zampolli, A., ed.; Calzolari, N., ed.. **Automating the Lexicon: Research and Practice in a Multilingual Environment**. Oxford University Press, 1995.

Cole, Ronald A., editror-in-chief 1996. **Survey of the State of the Art in Human Language Technology.** With Chapter 12 Language Resources and Section 12.4 Lexicons.

http://www.cse.ogi.edu/CSLU/HLTsurvey/HLTsurvey.html

Hutcheson, H.M. (1995) **Preparation of multilingual vocabularies**. *Standardizing and Harmonizing Terminology: Theory and Practice.* Philadelphia, PA: American Society for Testing and Materials. (1995): 102-114.

Other relevant publications by the tutorial instructor

A universal source thesaurus as a classification generator.

J. Amer. Soc. for Info. Sci. 1972.9; 23(5): 229-305.

Indexing and retrieval performance: The logical evidence.

J. Amer. Soc. for Info. Sci. 1994.9; 45(8): 589-599. (Invited paper)

Reprinted in From classification to "knowledge organization": Dorking revisited or "Past is prelude" / Edited by Alan Gilchrist. - The Hague: FID, 1997. - xiv, 186 p. - (FID pub. no. 714; FID Occasional paper 14). - ISBN 92 66 00 714 5

The Art and Architecture Thesaurus (AAT): A critical appraisal.

Visual Resources. 1995; 10(4): 369-400.

Software support for thesaurus construction and display.

Proceedings of the 5th ASIS SIG/CR Classification Research Workshop. Held at the 57th ASIS Annual Meeting, Oct. 16-20, 1994, Alexandria, VA.

Silver Spring, MD: American Society for Information Science. Special Interest Group / Classification Research. 1994.10; 5: 157-184. (Advances in Classification Research. v. 5)

Data structure and software support for integrated thesauri.

Paper presented at the Research Seminar on Compatibility and Integration of Order Systems, Warsaw, Poland, September 13-15, 1995.

Published in Compatibility and Integration of Order Systems: Research Seminar. Proceedings of the TIP/ISKO Meeting. Issued by International Society for Knowledge Organization; Polish Library Association; Soc. for Professional Information. Warsaw: Wydaw. SBP; 1996. p. 47-57.

(Included in the notebook).

SemWeb. Proposal for an open, multifunctional, multilingual system for integrated access to knowledge base about concepts and terminology.

Proceedings of the Fourth International ISKO Conference, 15-18 July 1996, Washington, DC. Frankfurt/Main: Indeks Verlag; 1996. (Advances in Knowledge Organization, v. 5). p. 165 - 173

Multilingual thesauri in cross-language retrieval. Paper persented at the AAAI-97 Spring Symposium Series. Cross-Language Text and Speech Retrieval. Stanford, CA March 24-26, 1997. Published in the Symposium Technical Report

Bibliographies of thesauri

Gerstenkorn, A. 1985 ed.; Rolland, M. Th. ed.; et al.

Thesaurus guide. Analytical directory of selected vocabularies for information retrieval.

Amsterdam: Elsevier; 1985. 748p.

Basis for the Echo database of thesauri

International Classification and Indexing Bibliography. Vol. I: Classification Systems and Thesauri 1950-1982. ICIB 1. 160 pages, DIN A4, DM 48.80; ISBN 3-88672-300-3; FID-Publ.610. Frankfurt/M: Indeks Verlag; 1982. This comprehensive bibliography of all universal and special classification systems and thesauri which could be found in the literature as well as in libraries, listing some 2300 titles from the time 1950-1982,

Includes bibliography of editions in multiple languages of

Universal Decimal Classification (UDC) Library of Congress Classification (LCC)

Dewey Decimal Classification (DDC) Library of Congress Subject Headings (LCSH)

Chan, Lois Mai; Pollard, Richard.

Thesauri used in online databases: an analytical guide.

New York: Greenwood; 1988. 268 p.

Brewer, Annie M.ed. **Dictionaries, Encyclopedias, & Other Word-Related Books.** 4th ed. Detroit: Gale Research; 1988. 1333 p. ISBN 0810304406

Molho, Emanuel. **The dictionary catalogue**. Second edition., 178 pages; French & European Publications, Inc., New York; 1989. (A bibliography of mono-, bi-, and multilingual dictionaries)

Examples of multilingual thesauri

Thesaurus EUROVOC: Official journal of the European communities. Office for Official Publications of the European Communities; 1995.

Viet, J. and Georges van Slype. **EUDISED Multilingual thesaurus for information processing in the field of education**, English version. 307 pages. Mouton Publishers, Berlin,, New York, Amsterdam; 1984.

EUDISED R&D Bulletin, volume 45, ISSN 0378-7192;. 127 pages.; K.G. Saur, Munich; 1993.

Food and Agriculture Organization of the United States. **AGROVOC multilingual agricultural thesaurus**. Second edition, English version; 798 pages; APIMONDIA, Rome; 1992. (Not latest)

International Atomic Energy Agency. **INIS: Thesaurus**. 887 p. and **INIS multilingual dictionary**. 314 p. IAEA, Vienna; 1993, 1983 (not latest editions).

Organization for Economic Cooperation and Development. **Multilingual dictionary of fish and fish products**. Fourth edition., 352 pages; Fishing News Books, Cambridge; 1995. LCC Ql64.7.M84.1995

Centre for Computer-Aided Egyptological Research. **Multilingual Egyptological thesaurus**. http://www.ccer.ggl.ruu.nl/thes/thsaur.html. 1995.

Verity K2 Toolkit

http://www.verity.com/products/k2toolkit/index.html

COMBINES ADVANCED SEARCH AND RETRIEVAL WITH STATE-OF-THE-ART PRECISION

The Verity K2 Toolkit combines enterprise-level performance and unlimited scalability withVerity´s state-of-the-art retrieval precision. The Verity K2 Toolkit provides complete access to Verity´s search engine so software designers can incorporate full text, metadata and concept-based Topics® searching within applications. All key Verity features are supported including relevancy ranking, highlighting, natural language query parsing, **thesaurus-based searching**, linguistic tools and advanced query navigation. The Verity K2 toolkit is fully compatible with current Verity collections, so existing hardware and software can be leveraged easily into bigger and faster applications.

What is the Verity K2 ToolKit?

The Verity K2 Toolkit is a software development tool that combines the market leading precision of the Verity Developer's Kit with the scalability and high-performance necessary to manage vast amounts of documents and users. The Verity K2 Toolkit enables organizations to build scalable fault-tolerant applications allowing thousands of users to search hundreds-of- millions of unstructured documents online, with nearly instantaneous results.

Do you need to build Topics to use K2?

Use of Topics is optional but provides users with the unique ability to share the expert queries tailored to your specific business rules that produce precise results.

Where can I get Topics if I don´t want to build them?

Verity resells Topicsets called Sageware Knowledgesets produced by Sageware Corp. These expert-created libraries include queries for over 700 industry segments in 20 industries. Libraries may be purchased that range from a single subject area to entire markets, covering company names and business intelligence terms in depth. In addition, there are a number of partners and consultants who sell Topicsets or help Verity customers to build custom Topicsets on contract. Contact Verity Consulting or your local sales representative for further information.

Examples of Thesauri

and other Knowledge Organization Structures

The paper tutorial notebook includes copies of sample pages from print thesauri, Web pages on thesauri, and of search results in various Web thesauri. The pdf file does not include these pages, but it does include the URLs of applicable Web pages.

Alcohol and Other Drug Thesaurus

Alcohol and Other Drug Thesaurus: A Guide to Concepts and Terminology in Substance Abuse and Addiction.

3rd ed. Washington, DC: U.S. Department of Health and Human Services, 2000.

Volume 1: Introduction and Overview, 387 p.

Volume 2: Annotated Hierarchy, 848 p.

Volume 3: Alphabetical Index, 406 p.

Volume 4: Annotated Alphabetical List, 896 p.

		2nd ed.	3rd. ed.
Number of:	Descriptors	10,315	11,323
	Lead-in Terms	6,675	7,783
	Total Terms	16,990	19,106
	History Notes		2,900
	Scope Notes (2.ed. incl. HN)	2,351	2,085
	Total descriptor cross-references	35,108	**39,720**

Web: http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

To order:

CSR Inc

1400 Eye St, NW, Ste 200

Washington, DC 20005

tel. (202) 842-7600

US$ 100 plus shipping

Art and Architecture Thesaurus

Art and Architecture Thesaurus. 2nd ed. Getty Art History Information Program. New York: Oxford University Press, 1994.

Volume 1 & 2: Introduction and Hierarchies, 455 p., 533 p.

Volume 3 - 5 : Aand - Zutu, 586 p., 586 p., 546 p.

US$ 375

Number of:	Descriptors	24,500
	Guide Terms	2,750
	Synonyms	20,000
	Total Terms	**47,000**
	British Variants	2,000
	Alternate Terms (singular/plural)	16,000
	Permutations	27,000
	Terms and Variants	**92,000**

Electronic Editions:

See http://www.getty.edu/research/tools/vocabulary/obtain.html

On the Web

http://www.getty.edu/research/tools/vocabulary/aat/index.html

Getty Vocabulary Program home page (copy included in this package)

http://www.getty.edu/research/institute/vocabulary/introvocabs/

Review article

Dagobert Soergel. **The Arts and Architecture Thesaurus (AAT). A critical appraisal**. Visual Resources 1995; 10(4): 369-400.

A few sample pages from an expanded version of this article are included here.

Dagobert Soergel

The Arts and Architecture Thesaurus (AAT). A critical appraisal.

4 Introduction: Thesauri in information retrieval

What is a thesaurus and what is its purpose? Describing the functions of a thesaurus in a nutshell will provide the background for a critical examination of the AAT. A thesaurus is a structured collection of concepts and terms for the purpose of improving the retrieval of information. A thesaurus should help the searcher to find good search terms, whether they be descriptors from a controlled vocabulary or the manifold terms needed for a comprehensive free-text search — all the various terms that are used in texts to express the search concept. Most thesauri establish a controlled vocabulary, a standardized terminology, in which each concept is represented by one term, a descriptor, that is used in indexing and can thus be used with confidence in searching; in such a system the thesaurus must support the indexer in identifying all descriptors that should be assigned to a document or other object in light of the questions that are likely to be asked. A good thesaurus provides, through its hierarchy augmented by associative relationships between concepts, a semantic road map for searchers and indexers and anybody else interested in an orderly grasp of a subject field.

A good thesaurus can be used for automatic search query expansion in two ways:

(1) synonym expansion, adding all the synonyms for a search term needed for free-text searching. For example,

color proofs addcolor separations

barrel vaults addcradle vaults, tunnel vaults, wagon vaults, wagonhead vaults

bluish gray addaqua gray, baby blue, blue black, blue gray, centroid color 191, light Payne's gray, pewter, powder blue, slate

(2) hierarchic expansion, adding all the narrower terms for a search term (also called inclusive searching). This is needed whether one searches with a controlled vocabulary or free-text, for example,

humanities addarts, linguistics, literature, philosophy, history, etc.

gold addelectrum, chryselephantine sculpture

barrel vaults addannular vaults, half barrel vaults, rampant barrel vaults, spiral vaults

saints add**hagiography**, **hagiographies**

B Associated concepts facet (1,018)

BM Associated concepts (1018)

D Physical attributes facet (890)

DC Attributes and properties (353)

DE Conditions and effects (46)

DG Design elements (162)

DL Color (329)

F Styles and periods facet (3,382)

FL Styles and Periods (3,382)

H Agents facet (1,093)

HG People (958)

HN Organizations (135)

K Activities facet (2,034)

KD Disciplines (318)

KG Functions (287)

KM Events (177)

KQ Physical activities (87)

KT Processes and techniques (1,165)

M Materials facet (2,869)

MT Materials (2,869)

P/V Objects facet (13,210)

PC Object groupings and systems (202)

PE Object genres (154)

PJ Components (3,066)

R Build Environment (1,943)

RD Settlements and landscapes (241)

RG Built complexes and districts (287)

RK Single built works (1,185)

RM Open spaces and site elements (230)

T Furnishings and equipment (5,592)

TC Furnishings (1,363)

TE Costume (721)

TH Tools and equipment (1,463)

TK Weapons and ammunition (256)

TN Measuring devices (315)

TQ Containers (622)

TT Sound devices (607)

TV Recreational artifacts (183)

TX Transportation vehicles (462)

V Visual and verbal communication (1,853)

VC Visual works (574)

VK Exchange media (169)

VW Information forms (1,110)

Numbers in parentheses give the number of descriptors to indicate emphasis.

Figure 1. **Top-level outline**

Facet	**Sample descriptors**
physical attributes	quarter plate, opacity, vivid red
styles and periods	Rococo
agents	painters (artists), photographers
activities and processes	gilding, gelatin silver process, color photography, carving, deterioration
materials	color film, wood
objects	chairs, negatives

Figure 4. **Facets and sample descriptors**

VC1 *<visual works>*
VC2 *<visual works by form>*
VC34 *<visual works by function>*
VC70 *<visual works by location or context>*
VC75 *<visual works by medium or technique>*
VC283 **photographs**
VC284 *<photographs by form>*
VC285 negatives
VC289 *<negatives by color>*
VC290 black-and-white negatives
VC291 color negatives
VC292 *<negatives by process>*
VC295 gelatin silver negatives
VC299 positives
VC310 photographic prints
VC312 later prints
VC315 *<photographic prints by color>*
VC316 black and white prints (photographs)
VC317 color prints (photographs)
VC318 *<photographic prints by process>*
VC322 chromogenic color print
VC346 *<photographs by form: color>*
VC347 black-and-white photographs
VC348 color photographs
VC349 *<photographs by form: format>*
VC357 slides (photographs)

VC358	black-and-white slides
VC359	color slides
VC360	*<photographs by function>*
VC363	news photographs
VC364	*<photographs by technique>*
VC365	*<photographs by picture-taking technique>*
VC366	aerial photographs
VC381	*<photographs by processing or presentation technique>*
VC390	manipulated photographs
VC391	composite photographs
VC400	*<photographs by subject type>*
VC406	studio portraits

Figure 5. **Example for minor facets and precombined descriptors**

Photography

D Physical Attributes Facet, DC Attributes and Properties

DC111 *<size: photograph formats>*
DC116 quarter plate

D Physical Attributes Facet, DE Conditions and Effects

DE38 *<conditions and effects: photography>*
DE39 oxidative-reductive deterioration

H Agents Facet, HG People

HG299 photographers

K Activities Facet, KT Processes and Techniques

KT487 *<photography and photographic processes and techniques>*
KT503 photographic processes
KT526 gelatin silver process
KT567 *<photographic techniques>*
KT570 *<picture-taking techniques>*
KT571 chronophotography
KT598 *<photographic processing and presentation techniques>*
KT602 enlarging

KT616 reduction (photography)

M Materials Facet, MT Materials

MT1416 paper
MT1463 *<paper by function>*
MT1481 photographic paper
MT2364 photographic materials
MT2367 photographic film

P/V Objects Facet, TH Tools and Equipment

TH746 photographic equipment
TH747 *<cameras and camera accessories>*
TH788 *<photographic processing equipment>*
TH794 enlargers
[no reducers]

P/V Objects Facet, VC Visual Works

VC283 photographs
VC284 *<photographs by form>*
VC285 negatives
VC292 *<negatives by process>*
VC295 gelatin silver negatives
VC364 *<photographs by technique>*
VC364 *<photographs by picture-taking technique>*
VC367 chronophotographs

Figure 6. **Facet arrangement dispersing concepts from same subject area**.
a. **Hierarchy excerpts concerning the subject Photography**

<art genres>

academic art

amateur art

apocalyptic art

art brut

children's art

commercial art

community art

SN Includes art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference to widely established standards. For art intended to beautify or enrich public places, use **public art**.

computer art

court art

crafts

cybernetic art

didactic art

dissident art

ethnic art

fantastic art

figurative art

folk art

funerary art

naive art

nonrepresentational art

primitive art

public art

SN Use for art whose purpose is to beautify and enrich public places. For art undertaken in conjunction with particular communities, usually to produce an effect or inspire response specifically within those communities, use **community art**.

rock art

cave art

serial art

art genres

art genres by content or other intrinsic characteristics

figurative art

fantastic art

apocalyptic art

nonrepresentational art

cybernetic art

serial art

crafts

art genres by standard

academic art

folk art

dissident art

art genres by type of artist or origin

amateur art

naive art

art brut

children's art

computer art

ethnic art

primitive art

art genres by audience, purpose, or display context

sofa art

court art

public art

SN Art whose purpose is to beautify and enrich public places.

community art

SN Public art undertaken in conjunction with particular communities, often socially deprived, usually with the idea of producing an effect or inspiring response specifically within those communities, with no reference

a. **Original alphabetical sequence**

b. **Suggested meaningful sequence**

Meaningful arrangement

Example from the Art and Architecture Thesaurus

Medical Subject Headings

Medical Subject Headings - Annotated Alphabetic List. 2002.

Bethesda, MD: National Library of Medicine, Nov. 2001, 1338 p.

Cost: US$ 51.95 Order No.: PB2001-964801

Medical Subject Headings - Tree Structures. 2002.

Bethesda, MD: National Library of Medicine, Nov. 2001, 908 p.

Cost: US$ 46.95 Order No.: PB2002-964901

Permuted Subject Headings. 2002.

Bethesda, MD: National Library of Medicine, Nov. 2001, 669 p.

Cost: US$ 42.95 Order No.: PB2002-965101

General MeSH info:www.nlm.nih.gov/mesh/meshhome.html
Ordering info: www.nlm.nih.gov/mesh/pubs.html
MeSH Files Available to Download: www.nlm.nih.gov/mesh/filelist.html

MeSH on the Web
www.nlm.nih.gov/mesh/MBrowser.html (more powerful)
www.ncbi.nlm.nih.gov/htbin-post/Entrez/meshbrowser (simpler)

Used in searching the bibliographic database Medline through PubMed

www.ncbi.nlm.nih.gov/PubMed/medline.html

PubMed uses MeSH and UMLS for synonym expansion and the MeSH hierarchy for hierarchic expansion

Unified Medical Language System (UMLS)

http://umlsinfo.nlm.nih.gov
www.nlm.nih.gov/pubs/factsheets/umlskss.html
www.nlm.nih.gov/pubs/factsheets/umlsmeta.html

Structure of the UMLS Metathesaurus

2000: 75 source vocabularies and 25 translations. Growth since 1998: x 1.5

Strings	Terms	Concepts
1,593,730 (1,718,083 tokens)	1,338,650	730,155

Substance Dependence Substance dependence substance dependence	Substance Dependence	Substance Dependence
Addiction, chemical addiction, chemical chemical addiction chemical addictions	Addiction, chemical	

adolescent	adolescent	adolescent
Teenager Teenagers teenager	Teenager	
teen teens	teen	
youth (young person) youths youth <1>	youth (young person)	

youth	youth

youth <2> youth (stage of life)	youth <2>	youth <2>

UMLS semantic types

https://umlsks.nlm.nih.gov/KSS/00/Specialist/Semantic_Net/semtype.list.html

Last Modified: Monday, February 07, 2000, copied May 31, 2000

Entity
- **Conceptual Entity**
  - Idea or Concept
    - Functional Concept
      - Body System
    - Temporal Concept
    - Qualitative Concept
    - Quantitative Concept
    - Spatial Concept
      - Body Location or Region
      - Body Space or Junction
      - Geographic Area
      - Molecular Sequence
        - Amino Acid Sequence
        - Carbohydrate Sequence
        - Nucleotide Sequence
  - Finding
    - Laboratory or Test Result
    - Sign or Symptom
  - Organism Attribute
    - Clinical Attribute
  - Intellectual Product
    - Classification
    - Regulation or Law
  - Language
  - Occupation or Discipline
    - Biomedical Occupation or Discipline
  - Organization
    - Health Care Related Organization
    - Professional Society
    - Self-help or Relief Organization
  - Group Attribute
  - Group
    - Age Group
    - Family Group
    - Professional or Occupational Group
    - Population Group
    - Patient or Disabled Group
- **Physical Object**
  - Anatomical Structure
    - Anatomical Abnormality
      - Acquired Abnormality
      - Congenital Abnormality
    - Embryonic Structure
    - Fully Formed Anatomical Structure
      - Body Part, Organ, or Organ Component
      - Cell
      - Cell Component
      - Tissue
      - Gene or Genome
  - Manufactured Object
    - Clinical Drug
    - Medical Device
    - Research Device
  - Organism
    - Animal
      - Invertebrate
      - Vertebrate
        - Amphibian
        - Bird
        - Fish
        - Mammal
          - Human
        - Reptile
    - Archaeon
    - Bacterium
    - Fungus
    - Plant
      - Alga
    - Virus
    - Rickettsia or Chlamydia
  - Substance
    - Body Substance
    - Chemical
      - Chemical Viewed Functionally
        - Biologically Active Substance
          - Neuroreactive Substance or Biogenic Amine
          - Hormone
          - Enzyme
          - Vitamin
          - Immunologic Factor
          - Receptor
        - Biomedical or Dental Material
        - Pharmacologic Substance

- Antibiotic
- Indicator, Reagent, or Diagnostic Aid
- Hazardous or Poisonous Substance
- Chemical Viewed Structurally
- Organic Chemical
- Amino Acid, Peptide, or Protein
- Carbohydrate
- Lipid
- Eicosanoid
- Steroid
- Nucleic Acid, Nucleoside, or Nucleotide
- Organophosphorus Compound
- Inorganic Chemical
- Element, Ion, or Isotope
- Food

Event

- Activity
- Behavior
- Social Behavior
- Individual Behavior
- Daily or Recreational Activity
- Occupational Activity
- Educational Activity
- Governmental or Regulatory Activity
- Health Care Activity
- Diagnostic Procedure
- Laboratory Procedure
- Therapeutic or Preventive
- Procedure
- Research Activity
- Molecular Biology Research
- Technique
- Machine Activity
- Phenomenon or Process
- Human-caused Phenomenon or Process
- Environmental Effect of Humans
- Injury or Poisoning
- Natural Phenomenon or Process
- Biologic Function
- Physiologic Function
- Cell Function
- Molecular Function
- Genetic Function
- Organ or Tissue Function
- Organism Function
- Mental Process
- Pathologic Function
- Cell or Molecular Dysfunction
- Disease or Syndrome
- Mental or Behavioral Dysfunction
- Neoplastic Process
- Experimental Model of Disease

UMLS semantic relations

https://umlsks.nlm.nih.gov/KSS/00/Specialist/Semantic_Net/relation.list.html

Last Modified: Monday, February 07, 2000, copied May 31, 2000

- **associated_with**
  - **physically_related_to**
    - branch_of
    - connected_to
    - consists_of
    - contains
    - ingredient_of
    - interconnects
    - part_of
    - tributary_of
  - **spatially_related_to**
    - adjacent_to
    - location_of
    - surrounds
    - traverses
  - **functionally_related_to**
    - affects
      - manages
      - treats
      - disrupts
      - complicates
      - interacts_with
      - prevents
    - brings_about
      - produces
      - causes
    - performs
      - carries_out
      - exhibits
      - practices
    - occurs_in
      - process_of
    - uses
    - manifestation_of
    - indicates
    - result_of
  - **temporally_related_to**
    - co-occurs_with
    - precedes
  - **conceptually_related_to**
    - analyzes
      - assesses_effect_of
    - conceptual_part_of
    - evaluation_of
    - degree_of
    - assesses_effect_of
    - measurement_of
    - measures
    - diagnoses
    - property_of
    - derivative_of
    - developmental_form_of
    - method_of
    - issue_in
- **isa**

Dewey Decimal Classification

`

Dewey Decimal Classification and Relative Index. 21st ed. Library of Congress. Albany, NY: OCLC Forest Press, 1996.

Volume 1: Introduction and Tables, 625 p.

Volume 2: Schedules 000 - 599, 1200 p.

Volume 3: Schedules 600 - 999, 1105 p.

Volume 4: Relative Index, 1207 p.

Cost: US$ 325, UK 220

World Wide Web:

Dewey Decimal Classification home page
www.oclc.org/oclc/fp/
www.oclc.org/dewey/products/webdewey/about.htm

Good display of top three levels
www.oclc.org/dewey/about/ddc_21_summaries.htm
www.tnrdlib.bc.ca/dewey.html
www.anthus.com/CyberDewey/CyberDewey.html

Examples of Internet Resources Classified by Dewey
www.oclc.org/dewey/worldwide/
http://link.bubl.ac.uk:80/linkbrowse
http://www.oclc.org/oclc/man/colloq/v-g

More complete listing to be on

www.clis.umd.edu/faculty/soergel/dlthestut

WordNet

WordNet Lexical Database. Version 1.7. Princeton University, Cognitive Science Laboratory, 2002.

WordNet is an online lexical database that is organized semantically rather than alphabetically.

		synsets (concepts)	word senses (terms, homonyms disambiguated
Number of:	nouns	60557	107424
(version 1.5)	verbs	11363	25761
	adjectives	16428	28749
	adverbs	3243	6201
	total	91591	168135

Web home page: www.cogsci.princeton.edu/~wn

Especially useful:

www.cogsci.princeton.edu/~wn/obtain.shtml
www.cogsci.princeton.edu/~wn/links.shtml
www.cogsci.princeton.edu/~geo/reader.html (WNet as lexical aid: click on word in text)
www.cogsci.princeton.edu/~wn/man1.7/wngloss.7WN.html
www.globalwordnet.org/
On EuroWordNet: http://www-ksl.stanford.edu/onto-std/eurowordnet.pdf

Best search: www.notredame.ac.jp/cgi-bin/wn.cgi (Not reachable on July 6, 2002)

Interesting site: http://www.beingmeta.com/brico/ (combines WordNet, Roget's 1911 Thesaurus, and the published top level of the CYC ontology)

D. Soergel **Top level hierarchy of WordNet's main categories**

Arranged building on the structure from the WordNet literature. Categories in [] added.

nouns	verbs	adjectives
thing, entity living thing, organism plant, flora animal, fauna person, human being and care non-living thing, object natural object body, corpus artifact substance food [other things or entities] group, collection	verbs of bodily function contact verbs	
process, action, event] process act, action, activity event, happening natural phenomenon	[**process verbs**] verbs of change creation verbs motion verbs [**social interaction and competition verbs**] verbs of social interaction competition verbs consumption verbs weather verbs	
[**time and place**] time place		
[**knowledge, communication, feeling**] cognition, knowledge communication feeling, emotion motive	[**knowledge, communication, feeling verbs**] cognition verbs perception verbs verbs of communication emotion or psych verbs	
[**attributes and relations**] attribute, property state, condition shape quantity, amount possession relation	[**stative and possession verbs**] stative verbs verbs of possession	descriptive adjectives color adjectives relational adjectives
		reference-modifying adjectives (e.g., *former* president)

Cyc Ontology

Cyc Ontology. Version 2.1. Cycorp, Inc.; 1997

The Cyc Ontology is a subset of the CYC system, a multi-conceptual knowledge base and inference engine. It is produced by

Cycorp, 3721 Executive Center Dr., Austin, TX 78731

Number of:	Concepts in the Cyc Ontology guide (upper ontology) "the topmost few percent of the hierarchy in the Cyc® Knowledge Base."	3000
	Concepts in the Cyc Knowledge Base	?

Web: www.cyc.com/cyc-2-1/cover.html

Especially

www.cyc.com/cyc-2-1/toc.html CYC® Ontology Guide: Table of Contents

www.cyc.com/cyc-2-1/intro-public.html Cyc® Ontology Guide: Introduction

CYC ontology top level outline (43 classes)

From http://www.cyc.com/cyc-2-1/toc.html Updated 1997-8-12, accessed 2001-4-15

Reformatted

Fundamentals	Biology
Top Level	Chemistry
Time and Dates	Physiology
Types of Predicates	General Medicine
Spatial Relations	Materials
Quantities	Waves
Mathematics	
Contexts	Devices
Groups	Construction
	Financial
"Doing"	Food
Transformations	Clothing
Changes Of State	Weather
Transfer Of Possession	Geography
Movement	Transportation
Parts of Objects	Information
Composition of Substances	Perception
	Agreements
Agents	
Organizations	Linguistic Terms
Actors	Documentation
Roles	
Professions	
Emotion	
Propositional Attitudes	
Social	

Supporting Documentation

The Syntax of CycL
The CYC® Functional Interface
Glossary of Common CYC® Terms

CYC Social Vocabulary Outline (created by DS from full file)

Some groupings, indicated by blank lines, introduced by DS (this outline would profit from better organization)

controls : <Agent> <Individual>
SocialBeing
affiliatedWith : <Agent> <Agent>
acquaintedWith : <IndividualAgent> <IndividualAgent>

Workplace

spectators : <Event> <Agent>
beneficiary : <Event> <Agent>

owns : <Agent> <SomethingExisting>
recipientOfService : <ServiceEvent> <Agent>
providerOfService : <ServiceEvent> <Agent>

socialParticipants : <SocialOccurrence> <Agent>
residesInDwelling : <Animal> <ShelterConstruction>
residesInRegion : <Animal> <GeographicalRegion>

HumanOccupationConstructResident
languageSpoken : <IntelligentAgent> <NatLanguage>
fieldsOfFormalEducation : <Person> <FieldOfStudy>
fieldsOfCompetence : <Person> <FieldOfStudy>
fieldsOfActivity : <Person> <FieldOfStudy>

representsAgentToAgent : <Agent> <Agent> <Agent>
socialClass : <Person> <SocialClass-Lifestyle>
SocialClass-Lifestyle
competingAgents : <Competition> <Agent>
eventHonors : <SocialOccurrence> <Agent>
positiveVestedInterest : <Agent> <TemporalThing>
negativeVestedInterest : <Agent> <TemporalThing>

AdultFemalePerson
HumanInfant
HumanChild
HumanAdult

SportsEvent
EntertainmentPerformance
EntertainmentEvent

spouse : <Person> <Person>
acquaintances : <Person> <Agent>
<AcquaintanceAttribute>
SimpleContactAcquaintance
AcquaintanceAttribute
friends : <Animal> <Animal>
boss : <Person> <Person>
cohabitingFamilyMembers : <Animal> <Animal>
cohabitants : <Animal> <Animal>
likesAsFriend : <SentientAnimal> <SentientAnimal>
loves : <SentientAnimal> <Agent>

maritalStatus : <Person> <MaritalStatusOfPeople>
MaritalStatusOfPeople

socialStatus : <Person> <SocialStatusAttributeType>
SocialStatusAttribute
SocialStatusAttributeType
SocialAttributeType
EducationLevelAttribute
schooling : <Person> <EducationalOrganization>
StudentStatusAttribute
educationLevel : <Person> <EducationLevelAttribute>
FieldOfStudy
ScientificFieldOfStudy
Religion

Title
CourtesyTitle
titleOfPerson-String : <Person> <CharacterString>
firstName : <Person> <HumanGivenNameString>
middleName : <Person> <HumanGivenNameString>
lastName : <Person> <HumanFamilyNameString>
ProperNameString
HumanNameString
HumanGivenNameString
HumanFamilyNameString
salutation : <Person> <CourtesyTitle>
nameOfAgent : <Agent> <ProperNameString>

ethnicity : <Person> <EthnicGroupType>
skinColor : <Person> <HumanSkinColor>
PersonalityAttribute
HumanCultureType
EthnicGroupType
Nationality

SocialOccurrence
SociabilityBasedAction
PublicEvent
SocialGathering
MeetingTakingPlace
Transaction
Party-Celebration
SocialRitual
Ritual

RudeAction
HostileSocialAction

TransferringPossession
GreetingSomeone
MeetingSomeone
VisitingSomeone

Competition
AthleticActivity
Bartering
MakingSomethingAvailable
AppropriatingSomething
ObtainingPermission

CommercialActivity
Advertising
Negotiating
BusinessRelationshipActivity

PhysicallyAttackingAnAgent
Battle
WagingWar
DisputeEvent
Trial

CYC Social Vocabulary. Annotated List

Excerpted from http://www.cyc.com/cyc-2-1/vocab/social-vocab.html. Updated 1997-10-14, accessed 2001-4-15

#$controls : <#$Agent> <#$Individual>

(#$controls X Y) represents that assertion that agent X controls the object Y, in one of the following 2 senses: X can influence (prohibit, enable or constrain) the behavior of Y; or else X can at least influence (prohibit, enable or constrain) the behavior of other #$Agents in/concerning Y. For example, Fred may control his horse directly, forcing it to do things, or not do them; and he also could control the horse indirectly, by deciding who else has access to and use of that horse. Control of one agent over another agent is rarely total, of course, so this predicate is most likely to apply to a Y which is a non-living possession, and/or to apply in a very narrow context. X's control over Y is usually either actual (de facto) control or legal (de jure) control. It is usually #$cotemporal, meaning that some time slice of X controls the same temporal time slice of Y.
isa: #$BinaryPredicate #$CotemporalObjectsSlot
genlPreds: #$positiveVestedInterest #$cotemporal
some more specialized predicates: (1 additl more specialized public predicate, 4 unpubl. ones)

#$SocialBeing

The collection of beings whose existence is accepted by some social system. (Thus, the elements of #$SocialBeing will vary with social contexts.) Social beings are entities able to perform social roles in the system that recognizes them. #$SocialBeing includes elements of #$Organization (e.g., the #$QueensGuard) as well as the elements of #$LegalAgent (in that system), so, for example, in modern industrial social systems, the elements of #$LegalCorporation and #$Person are instances of #$SocialBeing.
isa: #$ExistingObjectType
genls: #$IntelligentAgent
some subsets: #$JudicialAgent #$MedicalCareProvider #$Family-SocialEntity #$LegalAgent #$Organization #$Court-Judicial #$MedicalCareProfessional #$MedicalCareOrganization #$GeopoliticalEntity #$SoleProprietorship #$Partnership #$LegalCorporation #$LegalGovernmentOrganization #$Person #$ManufacturingOrganization (plus 157 more public subsets, 1992 unpublished subsets)

#$affiliatedWith : <#$Agent> <#$Agent>

. . .

#$acquaintedWith : <#$IndividualAgent> <#$IndividualAgent>

(#$acquaintedWith AGENT1 AGENT2) means the #$IndividualAgent AGENT1 is acquainted with the #$IndividualAgent AGENT2 (in the minimal sense that AGENT1 has come into physical or conversational contact with AGENT2, or that they have somehow knowingly communicated with each other). This typically means that each #$IndividualAgent is aware of some facts about the other. In cases where one of the #$IndividualAgents is sentient, this typically includes the ability of this agent to recognize the other by appearance, voice, scent, or some other physical attribute.
isa: #$CotemporalObjectsSlot #$BinaryPredicate #$Predicate #$ReflexiveBinaryPredicate #$SymmetricBinaryPredicate
genlPreds: #$cotemporal
some more specialized predicates: #$boss #$siblings #$mate #$cohabitants #$likesAsFriend #$spouse #$cohabitingFamilyMembers #$loves #$friends (plus 11 unpubl. more spec. pred.)

#$Workplace

The collection of places where people customarily work (not the employing organizations). #$Workplace includes offices, restaurant buildings, construction sites, agricultural sites, the #$SpaceNeedle, etc. Some places may be #$Workplaces only during a small part of their existence (a piece of residential property while the house is being built, perhaps); some may almost always be #$Workplaces (grocery store buildings, office buildings, smithies, hospitals, etc.).
isa: #$ContactLocationType
genls: #$HumanlyOccupiedSpatialObject #$PhysicalContactLocation
some subsets: (10 unpublished subsets)

#$AdultFemalePerson

The collection of all women; i.e., #$Persons who are adult and female
isa: #$ExistingObjectType
genls: #$HumanAdult #$FemalePerson

#$HumanInfant

The collection of #$Persons in the infant stage of life. Functionally, this ends when the infant learns to walk (even just toddle) and/or talk (even a few words)... or, at latest, when the person's age greatly exceeds that at which most people develop those skills. Generally, this means that it spans the period from birth to about 12 - 18 months old. One of the subsets of this collection is #$NewbornBaby.
isa: #$ExistingObjectType #$TemporalObjectType
genls: #$HumanChild
some subsets: (3 unpublished subsets)

#$HumanChild

The collection of all #$Persons in the childhood stage of life. Functionally, this ends when the child begins to take responsibility for themselves, work, have children of their own,... or, at latest, when the person's age greatly exceeds that at which most people reach those milestones. Generally, this means that it spans the period from birth to teenage years. This is highly dependent on context, of course; childhood in Shakespeare's culture ended around age 12.
isa: #$ExistingObjectType #$TemporalObjectType
genls: #$JuvenileAnimal
some subsets: (1 more public subset, 8 unpublished subsets)

#$HumanAdult

The collection of human beings old enough to participate as independent, mature members of society. Since different societies have different age or maturity requirements for people to be considered adults, different axioms in various society-specific microtheories express these requirements. For most modern, Western, middle-class,... purposes, e.g., the current view is that anyone over 18 is an adult. In many cultures, adulthood occurs when one reaches puberty. Adulthood is #$contiguousAfter childhood; that is, a #$Person is a #$HumanChild for a while, and then is a #$HumanAdult.
isa: #$ExistingObjectType #$TemporalObjectType
genls: #$AdultAnimal #$Person
some subsets: #$AdultFemalePerson (plus 16 unpublished subsets)

Additional schemes

Bloom **Taxonomy of educational objectives** 1956 (1 copy in the cataloging laboratory) (LB17.B55.1956), a summary at
http://www.unesco.org/webworld/ramp/html/r8810e/r8810e0e.htm
http://websites.ntl.com/~james.atherton/learning/bloomtax.htm,
http://sweep.riv.csu.edu.au/td/bloom.html,
http://faculty.washington.edu/~krumme/guides/bloom.html

SOC **Standard Occupational Classification** 2000
Bureau of Labor Statistics (BLS) + other agencies
http://stats.bls.gov/soc/soc_home.htm
The SOC is augmented by the **Occupational Information Network (O*NET)**, a database with additional occupational titles, definitions, and features of occupations.
http://www.doleta.gov/programs/onet

CSDGM **Content Standard for Digital Geospatial Metadata** 1998
Federal Geographic Data Committee (FGDC)
http://www.fgdc.gov/metadata/contstan.html

ERIC **Education Resources Information Center Thesaurus**. 13th ed.
http://searcheric.org/

Yahoo

The Yahoo classification. Web pages www.yahoo.com