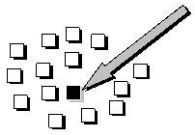
Thesaurus concepts

Leonard Will

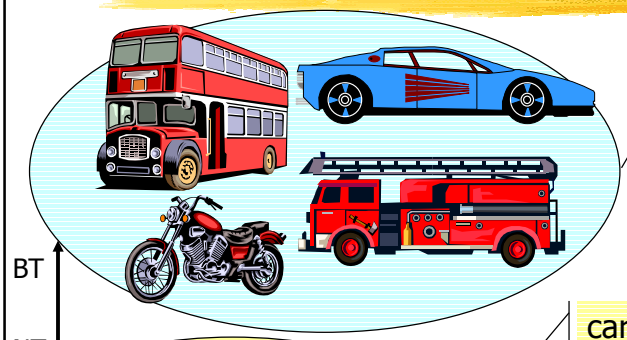


http://www.willpowerinfo.co.uk/

Concepts rather than words

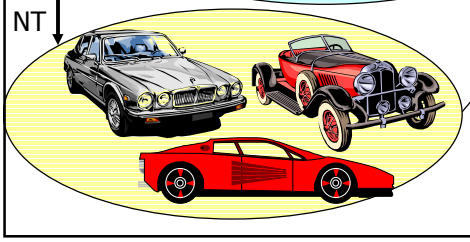
- **Concept**: a unit of thought, formed by mentally combining some or all of the characteristics of a concrete or abstract, real or imaginary object.
  - Concepts exist in the mind as abstract entities independent of terms used to express them [ANSI/NISO Z39.19]

Relationships are between concepts, not words



BT

NT

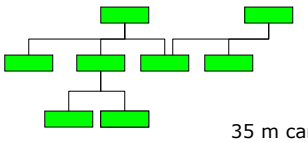


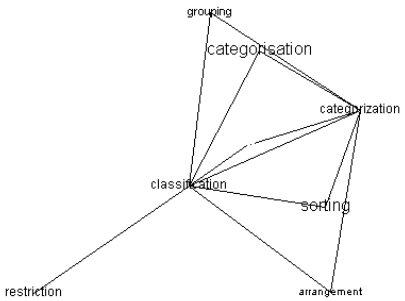
vehicles
road vehicles
conveyances
voitures
388.34
629.2

cars
automobiles
autos
private cars
388.342
629.222

Building blocks of all knowledge organisation schemes

- concepts
- relationships





35 m cameras
CC:H012
BT: film cameras
aqualungs
CC: D002
BT: diving equipment
camera accessories
CC: H002
BT: photographic equipment
NT: flash guns
light meters
tripods
RT: cameras

Relationships

- Paradigmatic, or *a priori* relationships: apply generally, independently of any specific document
  - shoes BT footwear
  - shoes RT shoemakers

A thesaurus can show these

- Syntagmatic, or *a posteriori* relationships: concepts that are related only in the context of a specific document
  - shoes : history
  - shoes : prices

A classification scheme can also show these

Ways of providing access to concepts

- Free text searching
- Classification schemes
- Subject headings
- Thesauri

Other current terms and techniques

- Ontologies
- Semantic nets
- Taxonomies
- Topic maps
- Citation links / usage counts (ISI, Amazon, Google)

Free text searching

- Characteristics
  - Depends on words in original documents
  - Can be simple matching of strings, words, phrases
  - Can use complex computer techniques
    - Word clusters and associations can be used
    - Infrequent terms given greater weight
    - Position of words can be significant

Free text searching

- Advantages
  - Little effort when creating records
  - Can retrieve items by unusual words, trade names, jargon, new terms
- Disadvantages
  - Inconsistent: depends on words used in original text
  - User has to think of alternative terms
  - May give high recall with low relevance

Software for automatic "classification"

- Autonomy: *categorizer*
  <http://www.autonomy.com/>
- Entrieva: *SemioTagger*
  <http://www.entrieva.com/entrieva/products/semiotagger.asp>
- Verity: *K2 enterprise*
  <http://www.verity.com/products/k2_enterprise/taxonomy.html>
- Vivisimo clustering engine
  <http://vivisimo.com/>

Classification *e.g. DDC, LC, ICONCLASS, SHIC, UDC*

- Characteristics
  - Concepts are arranged so that
    - Related concepts are near one another
    - There is a logical sequence of concepts
  - Primary grouping is usually by discipline
    - e.g. philosophy, religion, social science, science, technology, arts, literature, history
  - A symbolic notation is usually used to allow sorting

Hierarchical classification

300 Social sciences
360 Social problems and services; associations
363 Other social problems and services
363.1 Public safety programs
363.12 Transportation hazards
363.125 Highway and urban vehicular transportation
363.1256 Control of highway and urban vehicular transportation
363.12565 Investigation of specific vehicular and highway accidents

380 Commerce, communications, transportation
388 Transportation. Ground transportation
388.3 Vehicular transportation
388.34 Vehicles
388.341 Carts, wagons, carriages, stagecoaches
388.342 Automobiles
388.3423 Passenger automobiles for public transportation
388.34233 Buses

What is a facet?

(Sometimes called a fundamental facet)

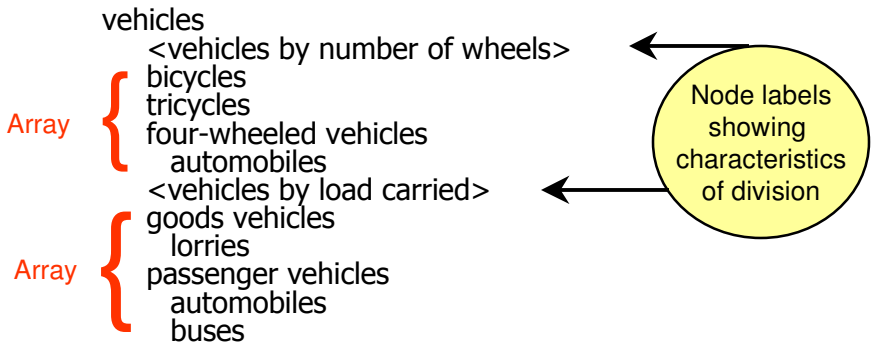
A high-level grouping of *concepts* of the same inherent category, e.g. activities, disciplines, people, materials, places, times. For example,

- *animals, mice, daffodils* and *bacteria* could all be members of a *living organisms* facet;
- *digging, writing* and *cooking* could all be members of an *activities* facet;
- *Paris, tropics* and *mountains* could all be members of a *places* facet.

What is an array?

(Sometimes called a subfacet)

A grouping of *concepts* within a *facet* by some stated *characteristic of division.*



Faceted classification

In a systematic or classified display, labels showing facet names introduce different types of concept:

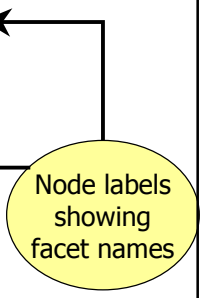
e.g. Classification for a transport collection

(places)
A1 cities
A2 villages
A3 motorways

(activities)
C1 accidents
C2 congestion
C3 overtaking

(people)
B1 drivers
B2 passengers
B3 traffic police
B4 planners

(vehicles)
D1 buses
D2 cars
D3 lorries



"Bus accidents on motorways" can be expressed as:
buses : accidents: motorways OR **D1C1A3**

Order of combining concepts

1 thing
2 kind
3 part
4 property
5 material
6 process
7 operation
8 patient (i.e. system operated on)
9 product
10 by-product
11 agent
12 space
13 time

So: **buses : accidents : motorways**
rather than: **accidents : motorways : buses**

Order of facets should be consistent within a discipline

Classification

- Advantages
  - Groups related material together, whatever the words or language used
  - A guided, logical sequence for browsing
  - A single preferred place for compound topics
- Disadvantages
  - Two-stage process: searcher must first find the relevant class
  - Notation may be complex when synthesised

Subject headings *e.g. Library of Congress Subject Headings (LCSH)*

Automobiles, Citroën (Firm)
Automobiles, Citroën (Firm) - Periodicals.
Automobiles, Classic
Automobiles - Classification - Periodicals
Automobiles - Cleaning
Automobiles - Climatic factors
Automobiles - Climatic factors - Congresses
Automobiles - Clutches
Automobiles - Clutches - Design and construction
Automobiles - Clutches - Maintenance and repair
Automobiles - Clutches - Maintenance and repair - Handbooks, manuals, etc.

Subject headings *e.g. Library of Congress Subject Headings (LCSH)*

- Advantages
  - Direct access through words
  - Browsing possible, if references followed
  - Widely available on bibliographic records, including MARC records for archives
- Disadvantages
  - Related items may be scattered (*animals, zoology*)
  - Later concepts in compound subjects may be hidden (e.g. *Great Britain - Civil defense*)
  - American spelling (e.g. *defense*)

A structured thesaurus

- Complements, does not replace, free text searching and classification
- A systematic development from alphabetical subject headings
- Particularly suitable for computer retrieval
- Purpose: to match concepts in a document with concepts in an enquiry

Choice of terms

- Many terms may be used for the same concept, e.g. *agriculture, farming, husbandry*
- Choose one term to represent this cluster
- There is no implication that the chosen term is "better" or "more correct" than the others
- Provide a scope note defining the concept
  SN *rearing of plants and animals on the land; includes market gardening, horticulture; excludes forestry, pet breeding*

Equivalence relationship

- USE / USE FOR [preferred / non-preferred terms]

agriculture
USE FOR *farming*
USE FOR *husbandry*

farming	USE	*agriculture*
husbandry	USE	*agriculture*

Preferred term substitution

Structure of terms

- Normally nouns and noun phrases
- Single concepts, not compound, e.g. *agricultural economics*
  → *agriculture + economics*
- Plurals for "count nouns" (answering "How many?"): *pigs, farms, children*
- Singulars for abstract concepts and "non-count nouns" (answering "How much?"): *peace, poverty, water, energy, economics*

Distinguish different terms with the same spelling

banks (earthworks)
banks (financial institutions)

cranes (birds)
cranes (lifting equipment)

Perth (Australia)
Perth (Scotland)

Hierarchical relationship (1)

BT / NT [broader term / narrower term]

crofts
BT *farms*

dairy farms
BT *farms*

farms
NT *crofts*
dairy farms

Hierarchical relationship (2)

- Relationship should be generic / specific
  - A *croft* IS A KIND OF *farm*
  - *crofts* BT *farms* ✔
  - A *farm* IS NOT A KIND OF *agriculture*
  - *farms* BT *agriculture* ✘
- Part / whole relationships in a few restricted cases
  - *Paris* BT *France* ✔
  - *blades* BT *knives* ✘

Polyhierarchy (1)

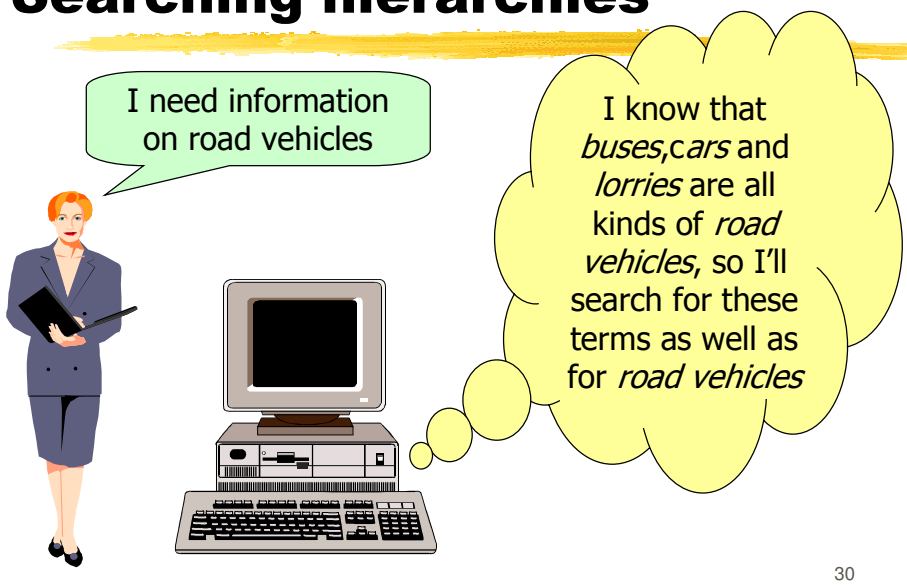
A term may have more than one broader term, so can appear in several places in a tree structure

France
BT *European Union*
EEC countries
French speaking countries
Mediterranean countries
OECD countries
Western Europe

Polyhierarchy (2)

vehicles
- *road vehicles*
  - *buses*
  - *cars*
  - *lorries*
- *passenger vehicles*
  - *buses*
  - *cars*
- *goods vehicles*
  - *lorries*

Searching hierarchies



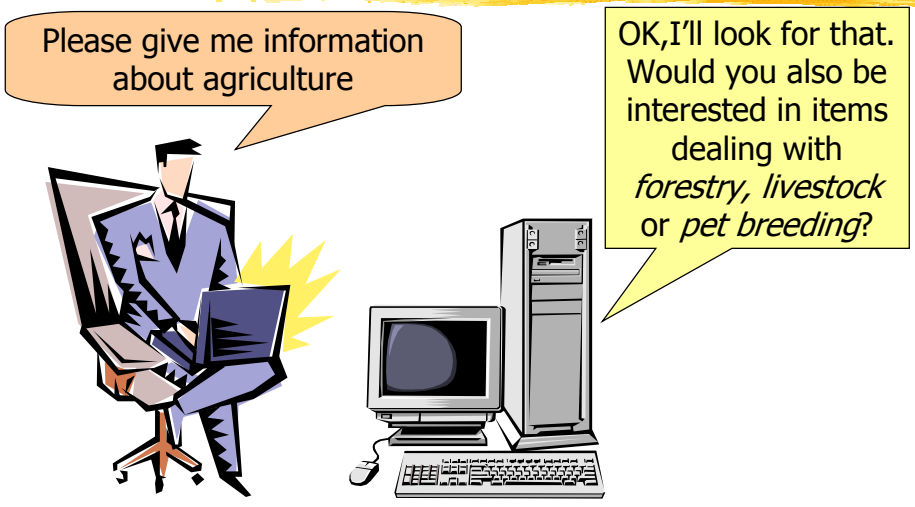
Associative relationship

RT / RT [related term / related term]

Use when someone searching for one term is likely to be interested in items indexed with the other

agriculture	RT	*farms*
farms	RT	*agriculture* *livestock*
livestock	RT	*farms*

Searching related terms



Microthesauri

Groups of terms, and hierarchies of terms, on related topics, e.g. education, science, culture

- *one tree for each facet in each category*

770: PHOTOGRAPHY

(fields of work)	(people)	(equipment)
. photography	. models (people)	. photo equipment
. . colour photography	. photographers	. . cameras

797.23: DIVING

(fields of work)	(people)	(equipment)
. diving	. divers	. diving equipment
. . scuba diving		. . aqualungs
. . snorkel diving		. . diving suits
		. . . dry suits

The same concepts viewed in different ways

Searching : thesaurus view

- If you know what you want
- Like a gazetteer or index
- Get quickly to individual concepts
- Usually arranged by facet
- Show paradigmatic relationships
- Combine concepts when searching

Browsing : classification view

- If you need to survey a topic
- Like a map or contents page
- See related concepts together
- Usually arranged by discipline
- Show syntagmatic and paradigmatic relationships
- See compound topics as pre-combined subject strings

Use of a thesaurus

- Software support for formulating questions
  - Browsing the thesaurus to choose terms
  - Combining terms with AND, OR, NOT and ( )
- A thesaurus as a search aid with unindexed material
  - Allows searching on terms linked to the term asked for

Thesaurus creation and management

- Standards
  - BS/ISO standards give helpful guidance
- Software
  - Many packages available
  - Best if integrated with database used for cataloguing
- Cooperative thesaurus development and use
  - DIY is a major and continuing task

Thesaurus development never ends

- It is an ongoing task
- It needs a knowledgeable thesaurus editor
- It needs cooperation and input from indexers and users
- User feedback

Examples of subject interfaces and thesauri (1)

- http://www.getty.edu/research/conducting_research/vocabularies/aat/
- http://www.getty.edu/research/conducting_research/vocabularies/tgn/
- http://www.getty.edu/research/conducting_research/vocabularies/ulan/
- http://www.picturethesaurus.gov.au/
- http://www.english-heritage.org.uk/thesaurus/thes_splash.htm
- http://eat.epicurious.com/
- http://hitite.adlibsoft.com/index.html
- http://www.arkive.org/
- http://www.mda.org.uk/
- http://www.sp2000.org/
- http://www.ukat.org.uk/

Examples of subject interfaces and thesauri (2)

- http://hilt.cdlr.strath.ac.uk/Sources/thesauri.html
- http://www.oudegracht4.nl/treasures/
- http://www.holm.demon.co.uk/shic.htm
- http://www.iconclass.nl/
- http://www.loc.gov/
- http://www.willpowerinfo.co.uk/

Lists of existing thesauri

- Controlled vocabularies resource guide
  http://sky.fit.qut.edu.au/~middletm/cont_voc.html
- Controlled vocabularies, thesauri and classification systems available in the WWW.
  http://www.lub.lu.se/metadata/subject-help.html
- *word*HOARD
  http://www.mda.org.uk/wrdhrd1.htm

References

- **British standard guide to establishment and development of monolingual thesauri** / British Standards Institution. - 1st rev. - London : BSI, 1987. - 32p ; 30cm. - (BS5723:1987) (ISO2788-1986)
- **British standard guide to establishment and development of multilingual thesauri** / British Standards Institution. - London : BSI, 1985. - 63p ; 30cm. - (BS6723:1985) (ISO5964-1985)
- **Guidelines for the construction, format, and management of monolingual thesauri** / National Information Standards Organization (U.S.). - Bethesda, MD : NISO Press, 1994 . - (ANSI/NISO Z39:19-1993). <http://www.techstreet.com/cgi-bin/detail?product_id=52601>