9. Ontology

INFO 202 - 29 September 2008

Bob Glushko

Plan for INFO 202 Lecture #9

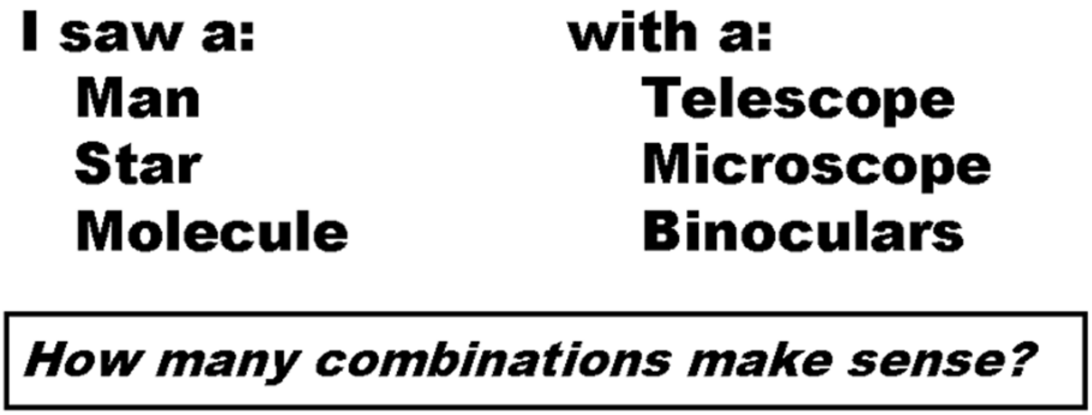
Introduction to ontology

A vocabulary for {lexical, conceptual} relationships

Indexes, thesauri, synonym rings

Topic maps

Making Sense [1]



I saw a:

- Man
- Star
- Molecule

with a:

- Telescope
- Microscope
- Binoculars

How many combinations make sense?

Making Sense [2]

"Bob saw the plane flying over Denver"
"Bob saw the mountains flying over Denver"

- What does "flying" refer to in each sentence?
- Where is "Bob" located?

Making Sense [3]

"How much is that doggy in the window?"

- Who is asking the question?
- What unit of measurement does "how much" refer to?
- Is the dog really "in" the window?

Language and Meaning

Words and sentence structure only hint at meaning

Meaning is constructed from all the clues or cues in the context of use -- common knowledge, assumptions, previous discourse, the present situation, and inferences from all of these

How much "context" and "common knowledge" must be represented / understood to make sense of what meaning is intended?

Two Solutions to the "Vocabulary Problem"

Furnas et al's solution (for people) was ...

The Artificial Intelligence solution (for computers) is to give an information system all the knowledge -- including "commonsense" -- that is needed to interpret every user's expressions in every context

A great deal of work in AI has been dedicated to building knowledge bases to support language understanding, reasoning, problem solving applications

The most famous / infamous effort is the Cyc project (http://www.cyc.com)

Cyc -- "Formalized Commonsense Knowledge"

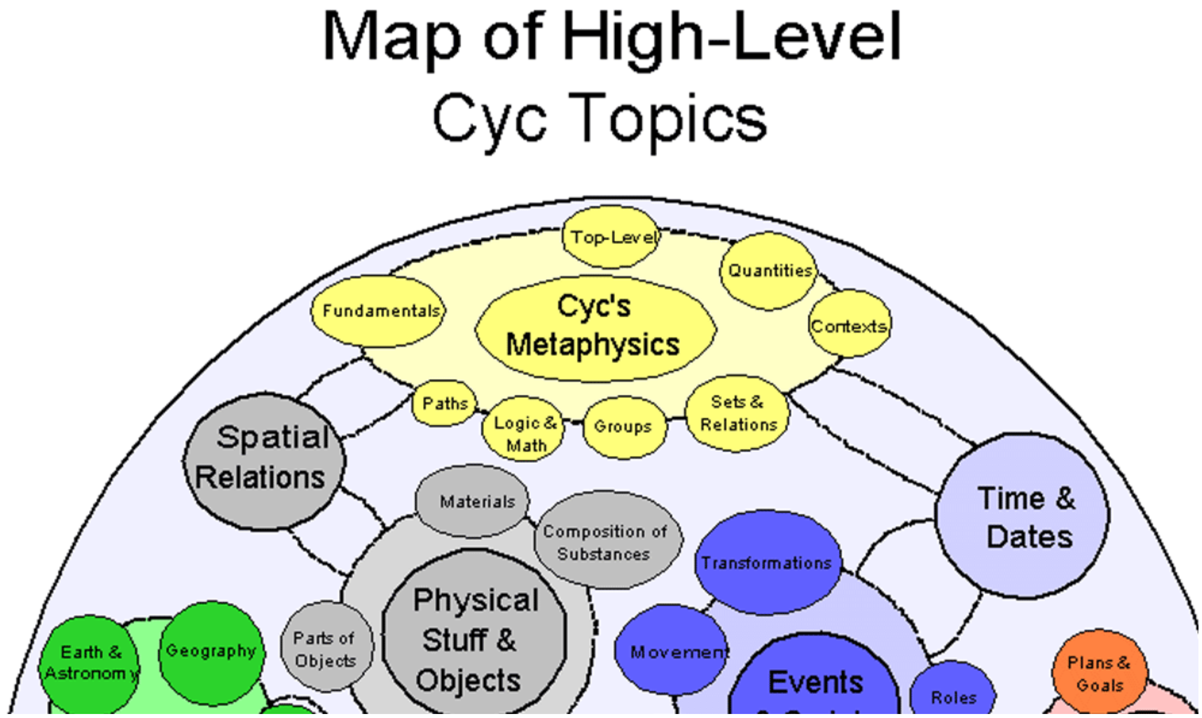
Cyc knows about 200,000 basic concepts and a few million human-entered assertions about the world -- "facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life"

The Cyc knowledge base consists of terms -- which constitute the vocabulary of CycL -- and assertions that relate those terms

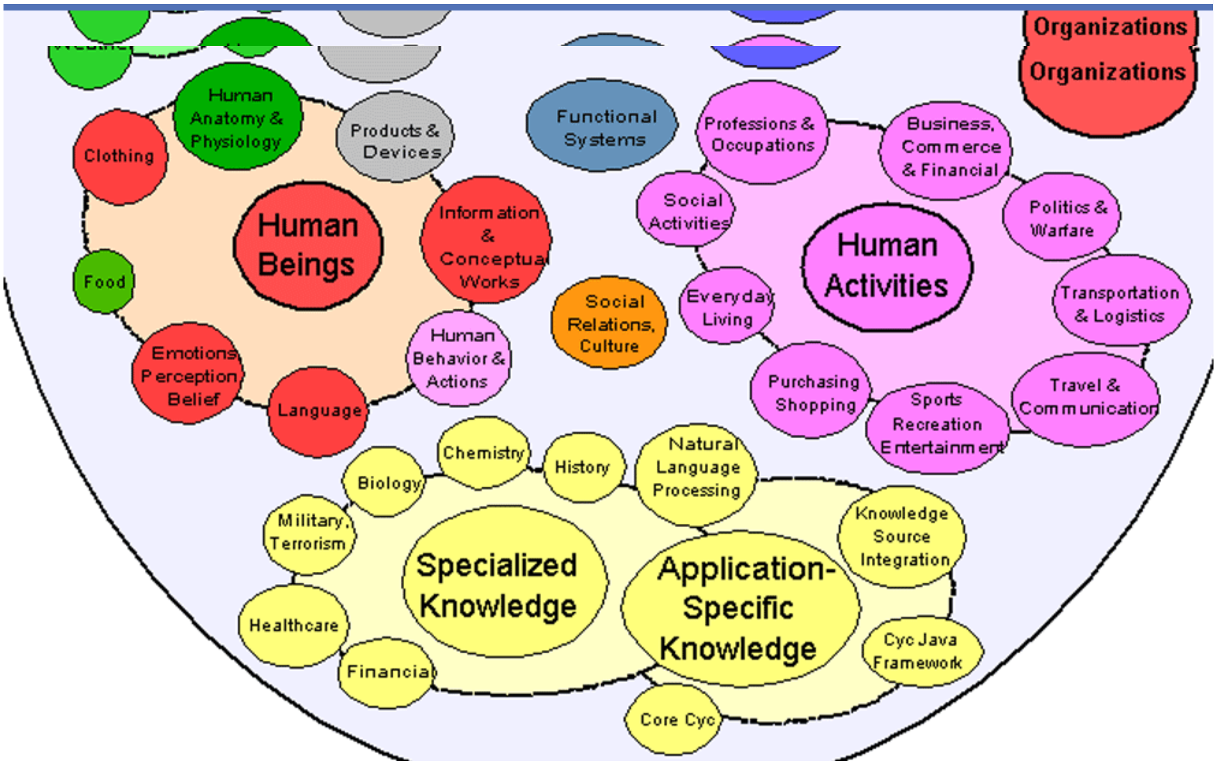
This kind of common sense is a pre-requisite for computers to achieve anything approaching human competence on natural language processing tasks (once you get outside of narrow, constrained domains)

The Cyc KB is divided into many (currently thousands of) "microtheories", each of which is essentially a bundle of assertions that share a common set of assumptions

What Cyc Knows About [1]



What Cyc Knows About [2]



Cyc Examples

Cyc can find the match between a user's query for "pictures of strong, adventurous people" and an image whose caption reads simply "a man climbing a cliff"

Cyc can notice if an annual salary and an hourly salary are inadvertently being added together in a spreadsheet

Cyc can combine information from multiple databases to guess which physicians in practice together had been classmates in medical school

Cyc Assertions About "Dog"

[Def] "A *BiologicalSpecies*
(scientific name 'Canis familiaris') that is a specialization of *CanineAnimal*

Each instance of *Dog*
is a canine animal that has either been bred to be a domestic pet (see *DomesticatedAnimal*) or is a wild canine animal that is not an instance of *Wolf*, *Fox*, or any other non-dog specialization of *CanineAnimal*

Note that although *Dog* and *Wolf* are considered distinct *BiologicalSpecies*, instances of the two can and do interbreed successfully. This species classification is therefore unusual, and in some circles, controversial."

What is An Ontology?

An ontology defines the terms used to describe and represent an area of knowledge.

Ontologies are used by people, databases, and applications that need to share domain information.

Ontologies include computer-usable DEFINITIONS of basic concepts in the domain and the RELATIONSHIPS among them

They encode knowledge in a domain and also knowledge that spans domains to make that knowledge reusable.

Cyc attempts to be a "foundation" or "upper" ontology, because it includes general concepts common to all domains, but is primarily a "domain" or "lower" ontology because most of its concepts are quite specific

That's A Very Broad Definition

The word ontology has been used to describe artifacts with different degrees of structure that differ:

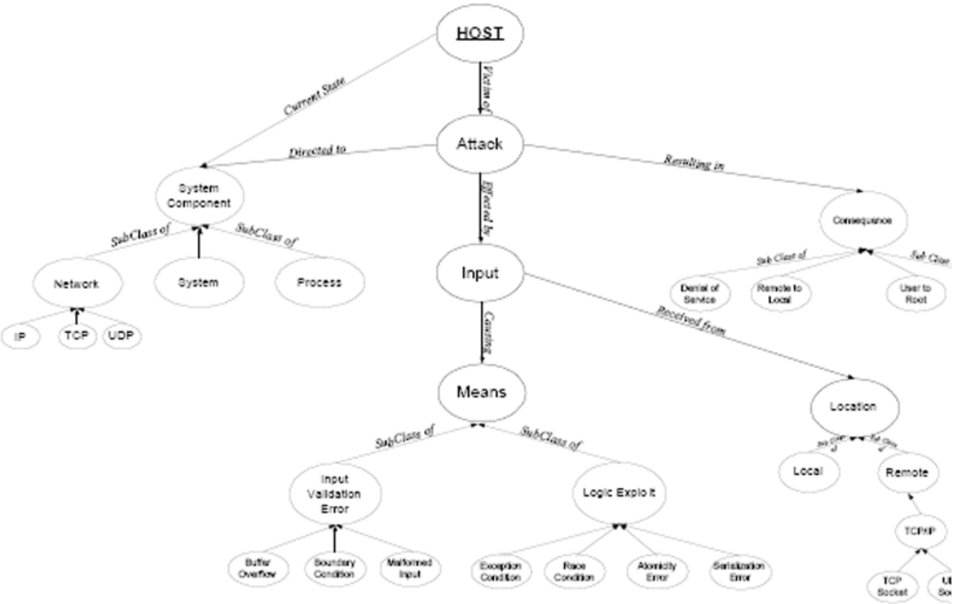
- ... according to how precisely the terms are defined
- ... according to how precisely the relationships among them are expressed

So the simplest ontology is a dictionary

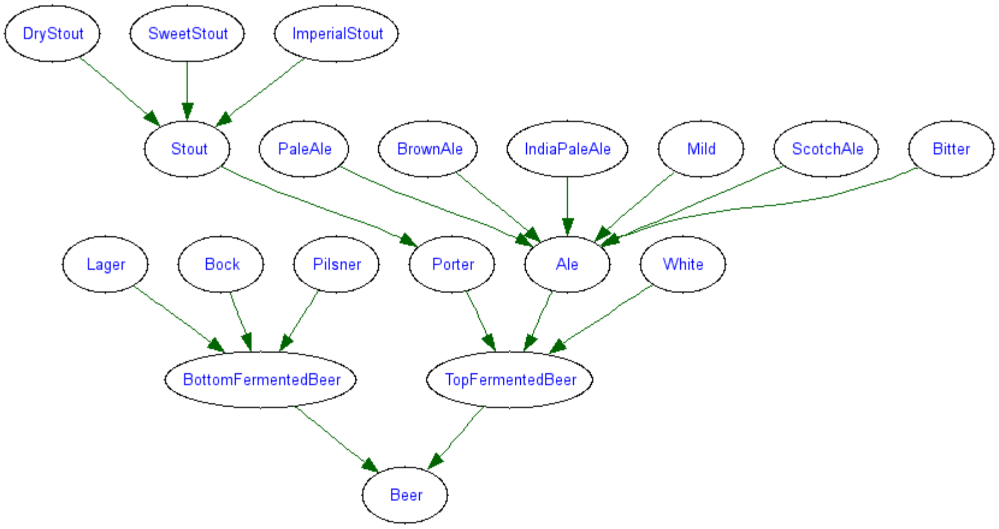
A thesaurus is a somewhat more complex ontology

More complete ontologies are expressed using formal logic-based language

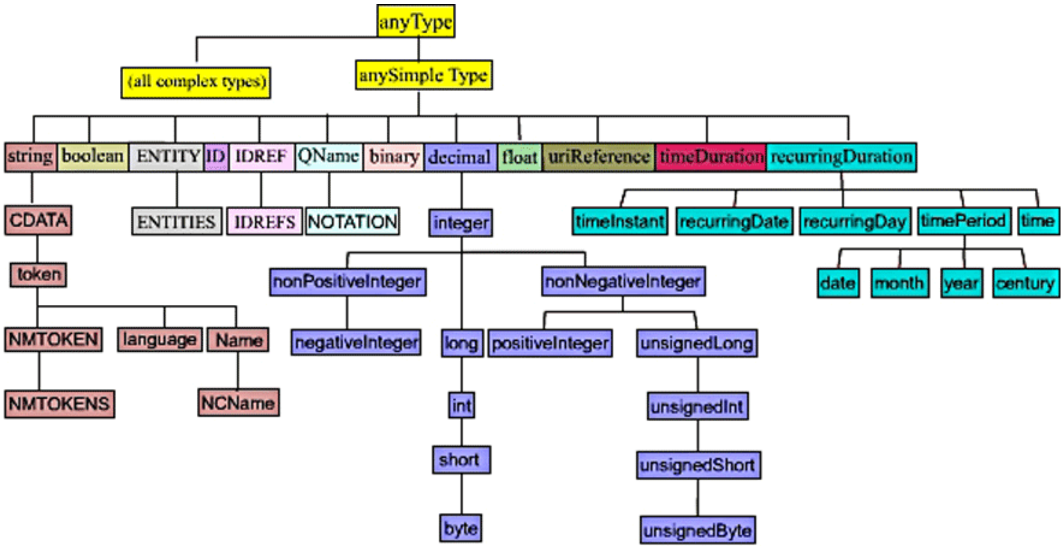
Ontology Example -- Computer Intrusion



Ontology Example -- Beer



Ontology Example -- W3C Datatypes



Rick Jellife's interactive datatype hierarchy -- each type is formally related to those around it by restriction relationships

Why Create an Ontology?

To share common understanding of the structure of information

To enable reuse of domain knowledge

To make domain assumptions explicit

To separate domain knowledge from operational knowledge

To analyze domain knowledge

Words and Concepts

A prototypical word is the minimal "meaning bearing" element of language

Words express concepts, but not all concepts are "lexicalized"

These "lexical gaps" differ from language to language

Whereas "conceptual gaps" -- the things we can't think of -- may be innate and universal

Relations Among Words

Polysemy

Synonymy

Antonymy

Relations Among Concepts

Hyponymy/Hypernymy

Meronymy/Holonymy

Polysemy

Many "word forms" (particular spelling patterns) are polysemous with multiple senses -- they are semantically ambiguous

- That dog has floppy *ears*
- She has a good *ear* for jazz.

These senses are established in the language and stored in a person's memory, and not be just possible uses

"bank" (financial) has related senses:

- a building (the bank on Shattuck)
- a specific financial firm (Wells Fargo)
- where money is kept (abstract notion)

Polysemy vs Metaphor

A polyseme is a word with multiple senses, but in which all the senses are related

Metaphor is a kind of nonlinear or figurative polysemy, where a sense is related but perhaps only on one or a few of the facets of a concept

- *swallow* a pill
- *swallow* an argument

Polysemy vs Homonymy

Two kinds of homoyms:

A HOMOGRAPH is a word with multiple senses, but for which the different senses are not conceptually related

- bank (financial sense)
- bank (river sense)

But what appears to be homography may be polysemy from a historical perspective...and native speakers sometimes disagree about whether two senses are polysemous or homonymous

HOMOPHONES are two words with the same pronunciation but different spellings and meanings

Synonymy

Synonyms are different word forms that can express the same concept

- cat, feline, Siamese cat

Absolute synonyms that can be substitutable for each other in every conceivable context probably don't exist

- {weep, sob, cry}-- differ in scale or degree
- "brave" implies physical, "courageous" implies moral

Propositional synonyms are more common - substitutability entails the same truth conditions

- She plays the {violin, fiddle}

Antonymy

Antonyms are lexical opposites

Some are "true antonyms" because they are inherently binary

- dead / alive, true / false, on / off

Others are "graded"

- long / short, hot / cold

Markedness: if one member of a pair is more restricted in its contexts it can stand out psychologically

- long is unmarked, short is marked

Hyponymy/Hyperonymy

The IS-A relationship -- a relationship between concepts that organizes the word nouns into a "lexical hierarchy"

Often used to situate "basic categories" with respect to superordinate and subordinate categories

- A robin is a hyponym of bird
- A bird is a hyponym of animal
- An animal is a hypernym of bird

A is a hyponym of B if A is a type of B

Co-hyponyms are mutually exclusive categories

A is a hypernym of B if B is a type of A

A Formula for Definitions

hyponym = {adjective+} hypernym {distinguishing clause+}

Robin = Migratory BIRD with clear melodious song, a reddish breast, gray or black upper plumage

Doesn't mention every characteristic of hyponym, only those needed to distinguish from other hyponyms

Meronymy/Holonymy

Meronymy defines Part/Whole relations

- Beak is a meronym of Bird
- Bark is a meronym of Tree

Holonyms are (approximately) the inverse of meronyms

- Tree is a holonym of Bark

Meronymy is transitive conceptually but not lexically

- The Knob is part of the Door
- The Door is part of the House
- but sounds odd to say "The Knob is part of the House"

Indexes

A "map" to the knowledge contained in a text or collection of texts

Consists of a list of (names of) topics and references to occurrences of those topics

Topics can be arranged / decomposed hierarchically

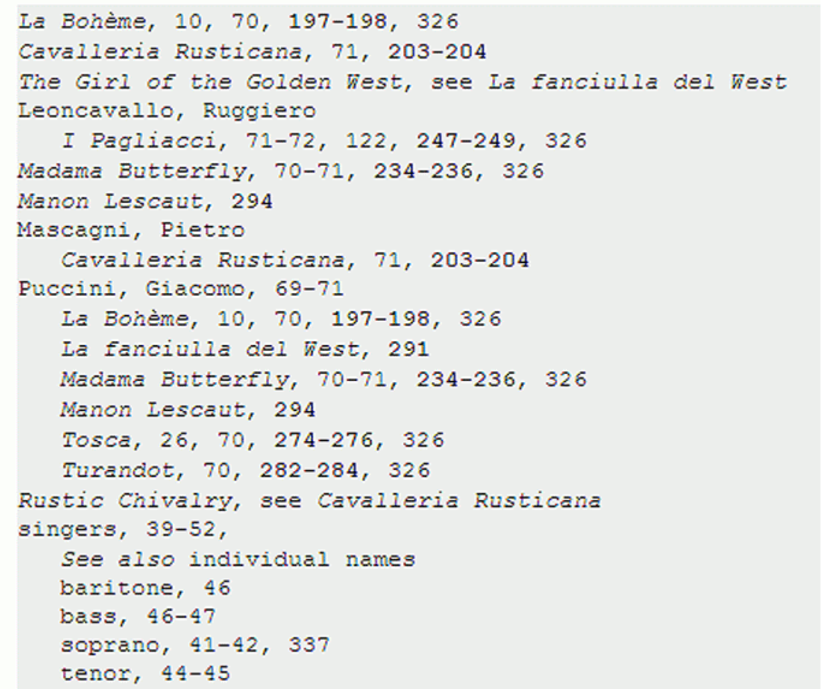
Topics can be associated with other topics

Topics and references can be "typed"

Topics and references can be "aliased"

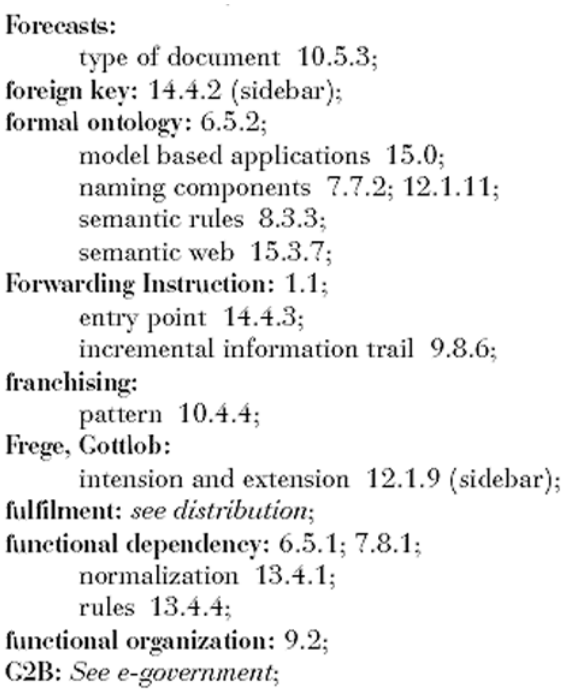
Typographic conventions can be used to reinforce these distinctions and relationships

Index - Opera (from TAO of Topic Maps)



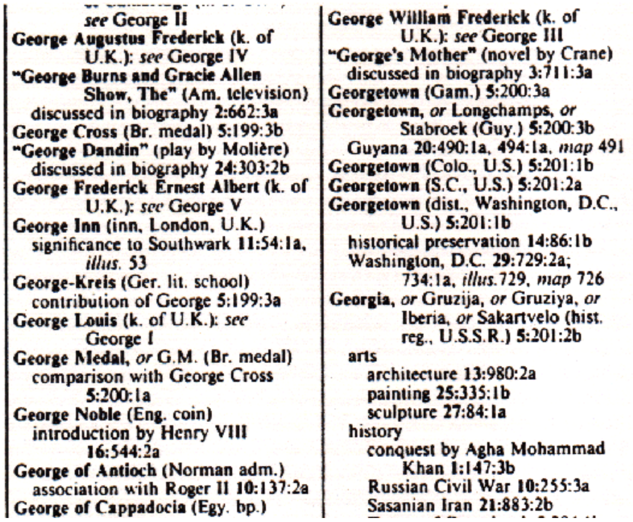
- *La Bohème*, 10, 70, 197-198, 326
- *Cavalleria Rusticana*, 71, 203-204
- *The Girl of the Golden West*, see *La fanciulla del West*
- Leoncavallo, Ruggiero
  - *I Pagliacci*, 71-72, 122, 247-249, 326
- *Madama Butterfly*, 70-71, 234-236, 326
- *Manon Lescaut*, 294
- Mascagni, Pietro
  - *Cavalleria Rusticana*, 71, 203-204
- Puccini, Giacomo, 69-71
  - *La Bohème*, 10, 70, 197-198, 326
  - *La fanciulla del West*, 291
  - *Madama Butterfly*, 70-71, 234-236, 326
  - *Manon Lescaut*, 294
  - *Tosca*, 26, 70, 274-276, 326
  - *Turandot*, 70, 282-284, 326
- *Rustic Chivalry*, see *Cavalleria Rusticana*
- singers, 39-52,
  - *See also* individual names
  - baritone, 46
  - bass, 46-47
  - soprano, 41-42, 337
  - tenor, 44-45

Index - Document Engineering



- **Forecasts:**
  - type of document 10.5.3;
- **foreign key:** 14.4.2 (sidebar);
- **formal ontology:** 6.5.2;
  - model based applications 15.0;
  - naming components 7.7.2; 12.1.11;
  - semantic rules 8.3.3;
  - semantic web 15.3.7;
- **Forwarding Instruction:** 1.1;
  - entry point 14.4.3;
  - incremental information trail 9.8.6;
- **franchising:**
  - pattern 10.4.4;
- **Frege, Gottlob:**
  - intension and extension 12.1.9 (sidebar);
- **fulfilment:** *see distribution*;
- **functional dependency:** 6.5.1; 7.8.1;
  - normalization 13.4.1;
  - rules 13.4.4;
- **functional organization:** 9.2;
- **G2B:** *See e-government*;

Index - Encyclopedia Americana



see George II
George Augustus Frederick (k. of U.K.): *see* George IV
"George Burns and Gracie Allen Show, The" (Am. television)
discussed in biography **2:662:3a**
George Cross (Br. medal) **5:199:3b**
"George Dandin" (play by Molière)
discussed in biography **24:303:2b**
George Frederick Ernest Albert (k. of U.K.): *see* George V
George Inn (inn, London, U.K.)
significance to Southwark **11:54:1a,** *illus.* 53
George-Kreis (Ger. lit. school)
contribution of George **5:199:3a**
George Louis (k. of U.K.): *see* George I
George Medal, *or* G.M. (Br. medal)
comparison with George Cross **5:200:1a**
George Noble (Eng. coin)
introduction by Henry VIII **16:544:2a**
George of Antioch (Norman adm.)
association with Roger II **10:137:2a**
George of Cappadocia (Egy. bp.)

George William Frederick (k. of U.K.): *see* George III
"George's Mother" (novel by Crane)
discussed in biography **3:711:3a**
Georgetown (Gam.) **5:200:3a**
Georgetown, *or* Longchamps, *or* Stabroek (Guy.) **5:200:3b**
Guyana **20:490:1a, 494:1a,** *map* 491
Georgetown (Colo., U.S.) **5:201:1b**
Georgetown (S.C., U.S.) **5:201:2a**
Georgetown (dist., Washington, D.C., U.S.) **5:201:1b**
historical preservation **14:86:1b**
Washington, D.C. **29:729:2a;** 734:1a, *illus.*729, *map* 726
Georgia, *or* Gruzija, *or* Gruziya, *or* Iberia, *or* Sakartvelo (hist. reg., U.S.S.R.) **5:201:2b**
arts
architecture **13:980:2a**
painting **25:335:1b**
sculpture **27:84:1a**
history
conquest by Agha Mohammad Khan **1:147:3b**
Russian Civil War **10:255:3a**
Sasanian Iran **21:883:2b**

Thesauri

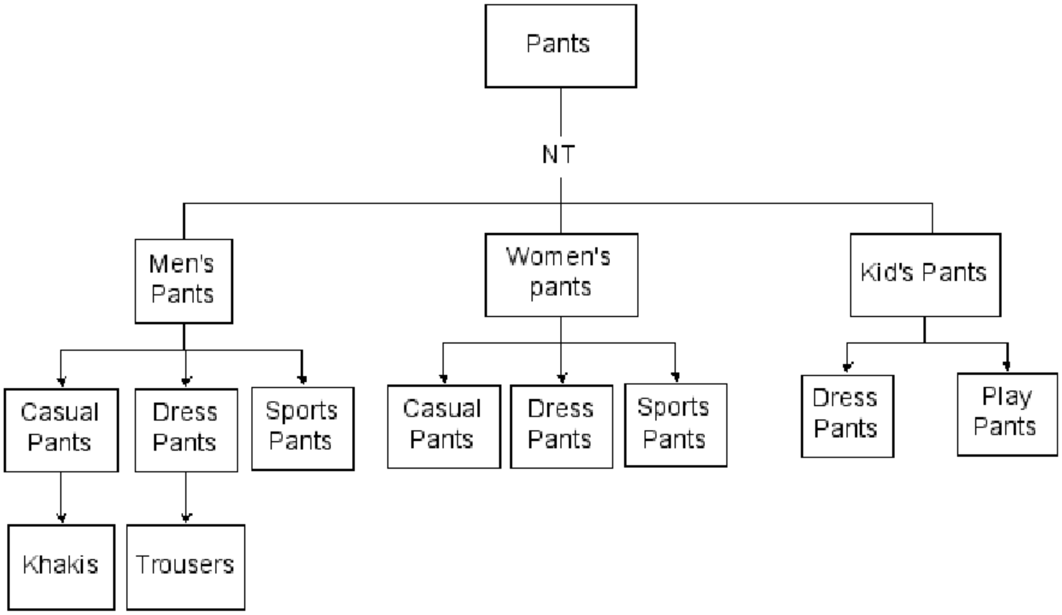
A THESAURUS is a tool for leading cataloguers or searchers to the "right" or "good" terms of a controlled vocabulary

It is a collection of (usually single) vocabulary terms annotated with lexical relationships to indicate terms that are:

- Preferred (UF "used for")
- Broader (BT "broader term")
- Narrower (NT "narrower term")
- Related (RT "related term" or "see also")

USE in a thesaurus refers the reader from a variant term to a preferred term; the inverse of UF

Thesaurus Example (Graphical Format)



Pants

NT

Men's Pants

Women's pants

Kid's Pants

Casual Pants

Dress Pants

Sports Pants

Casual Pants

Dress Pants

Sports Pants

Dress Pants

Play Pants

Khakis

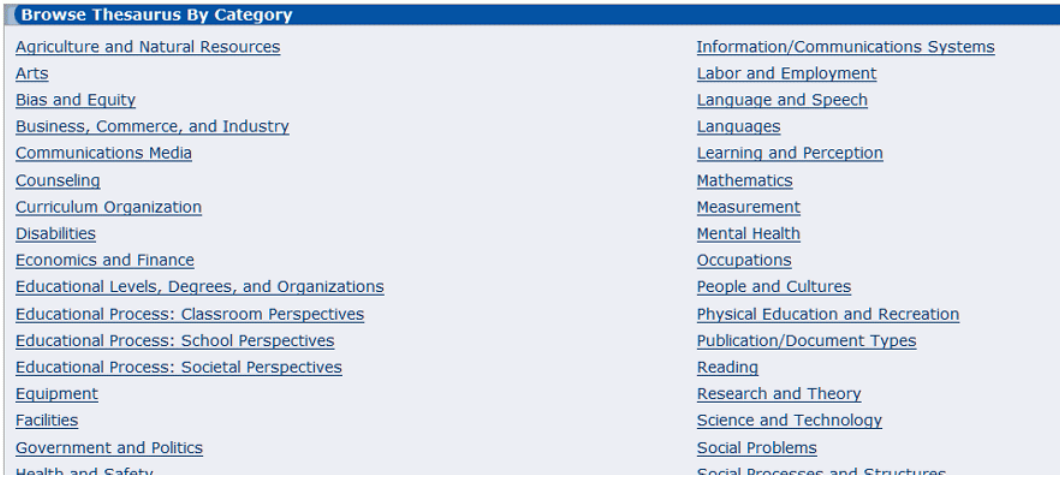
Trousers

Thesaurus Example (Textual Format)

```
        Women's Pants
BT Pants
NT Casual Pants
NT Dress Pants
        Jeans
BT Pants
NT Levis
NT Wranglers
NT Sports Pants
UF Waist Overalls
RT Denim
RT Overalls
```

Thesaurus Example - ERIC [1]

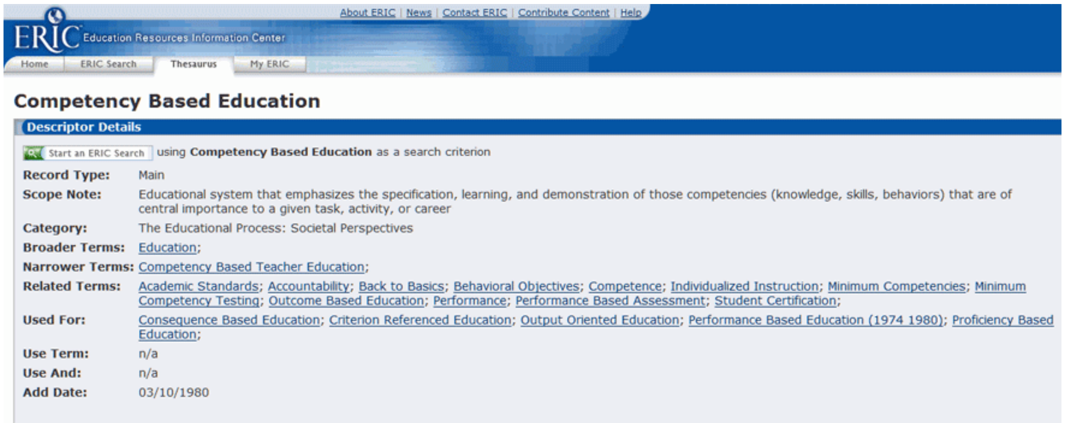
http://www.eric.ed.gov/
-- Education Resources Information Center (ERIC) is a digital library of education-related resources, sponsored by the Institute of Education Sciences of the U.S. Department of Education.



Browse Thesaurus By Category

- Agriculture and Natural Resources
- Arts
- Bias and Equity
- Business, Commerce, and Industry
- Communications Media
- Counseling
- Curriculum Organization
- Disabilities
- Economics and Finance
- Educational Levels, Degrees, and Organizations
- Educational Process: Classroom Perspectives
- Educational Process: School Perspectives
- Educational Process: Societal Perspectives
- Equipment
- Facilities
- Government and Politics
- Health and Safety
- Information/Communications Systems
- Labor and Employment
- Language and Speech
- Languages
- Learning and Perception
- Mathematics
- Measurement
- Mental Health
- Occupations
- People and Cultures
- Physical Education and Recreation
- Publication/Document Types
- Reading
- Research and Theory
- Science and Technology
- Social Problems
- Social Processes and Structures

Thesaurus Example - ERIC [2]


About ERIC | News | Contact ERIC | Contribute Content | Help
ERIC Education Resources Information Center
Home
ERIC Search
Thesaurus
My ERIC
Start an ERIC Search

Competency Based Education

Descriptor Details

Start an ERIC Search using **Competency Based Education** as a search criterion

Record Type:	Main
Scope Note:	Educational system that emphasizes the specification, learning, and demonstration of those competencies (knowledge, skills, behaviors) that are of central importance to a given task, activity, or career
Category:	The Educational Process: Societal Perspectives
Broader Terms:	Education;
Narrower Terms:	Competency Based Teacher Education;
Related Terms:	Academic Standards; Accountability; Back to Basics; Behavioral Objectives; Competence; Individualized Instruction; Minimum Competencies; Minimum Competency Testing; Outcome Based Education; Performance; Performance Based Assessment; Student Certification;
Used For:	Consequence Based Education; Criterion Referenced Education; Output Oriented Education; Performance Based Education (1974 1980); Proficiency Based Education;
Use Term:	n/a
Use And:	n/a
Add Date:	03/10/1980

Synonym Rings

"If you need to know about cow farming, you're probably also searching for cattle ranching, beef (or dairy) production, and Kuhbauernhof, whether you know it or not." (Tim Bray)

A synonym ring connects a series of terms together and treats them all as equivalent for search purposes

It is a weaker mechanism of vocabulary control than an authority file or thesaurus because it doesn't designate a term as the preferred or normative form

Recommended Types of Synonyms for Rings

Scientific terms versus popular use terms: acetylsalicylic acid, aspirin; lilioceris, lily beetle

Variant spellings: cancelled, canceled; honor, honour

Abbreviations (Initialisms, acronyms, apocopations, nicknames)

Topic Maps

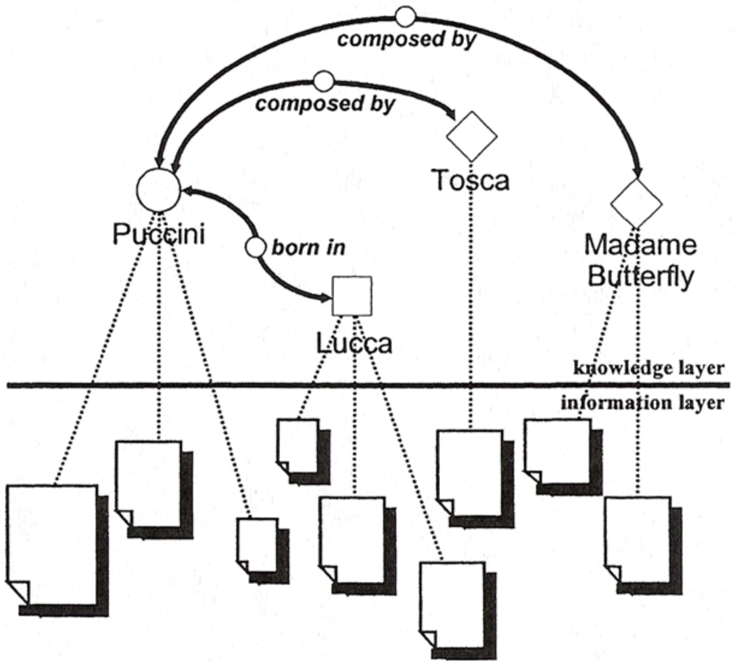
A recent invention... designed to support the distributed management of information and "knowledge"

Motivation is "merging the indexes" of printed and digital information collections

Two-layer model: an "information layer" consisting of "topics" and "associations" and a "knowledge layer" that are linked together by "occurrences"

(Try the "Omnigator" - generic topic map browser - at http://www.ontopia.net/omnigator)

Topic Maps - Two Layers



Assignment 4

Designing a faceted classification system to organize household "tools" using 10 instances provided to you

Test the scope and robustness of your system with additional instances provided to you by someone else

Use Facetmap (http://facetmap.com/)

Turn in your facetmap and a report about your experiences by October 6

Readings for INFO Lecture #10

R. Glushko, "Modeling Methods and Artifacts for Crossing the Data/Document Divide"

K. Thomas, "XML in the Pharmaceutical Industry: Structured Product Labeling"

Tim Bray, "On Language Creation"