Knowledge Organization

Franz J. Kurfess

Computer Science Department
California Polytechnic State University
San Luis Obispo, CA, U.S.A.

Acknowledgements

*Some of the material in these slides was developed for a lecture series sponsored by the **European Community** under the **BPD program** with **Vilnius University** as host institution*

Logistics - Jan 29, 2013

- **Project**
  - Team repositories: TRAC Wiki, alternatives
    - will start grading of
      - project description, background and related work, difficulty, relevance
- **KB Nugget presentations**
  - who's presenting today?
  - Topics
  - Signup for Date & Time Slots via Semantic Media Wiki
- **Assignments**
  - A1: Concept Map
    - due Jan 31
  - A2: Ontology
    - Protégé tutorial, quiz, ontology submission
    - due Feb 14

Overview Knowledge Organization

- **Motivation, Objectives**
- **Chapter Introduction**
  - New topics, Terminology
- **Identification of Knowledge**
  - Object Selection
  - Naming and Description
- **Categorization**
  - Feature-based Categorization
  - Hierarchical Categorization
- **Knowledge Organization Methods**
  - Natural Language
  - Ontologies
- **Knowledge Organization Tools**
  - Editors, visualization tools, automated ontology construction
- **Examples**
- **Important Concepts and Terms**
- **Chapter Summary**

Motivation and Objectives

Motivation

- **effective utilization of knowledge depends critically on its organization**
  - quick access
  - identification of relevant knowledge
  - assessment of available knowledge
    - source, reliability, applicability
- **knowledge organization is a difficult task, and requires complementary skills**
  - expertise in the domain
  - knowledge organization skills
    - librarians

Objectives

- be able to identify the main aspects dealing with the organization of knowledge
- understand knowledge organization methods
- apply the capabilities of computers to support knowledge organization
- practice knowledge organization on small bodies of knowledge
- evaluate frameworks and systems for knowledge organization

Knowledge Organization

Identification of Knowledge

Object Selection; Naming and Description

Categorization

Feature-based Categorization; Hierarchical Categorization

Knowledge Organization Methods

Natural Language; Ontologies

Knowledge Organization Tools

Editors, visualization tools, automated ontology construction

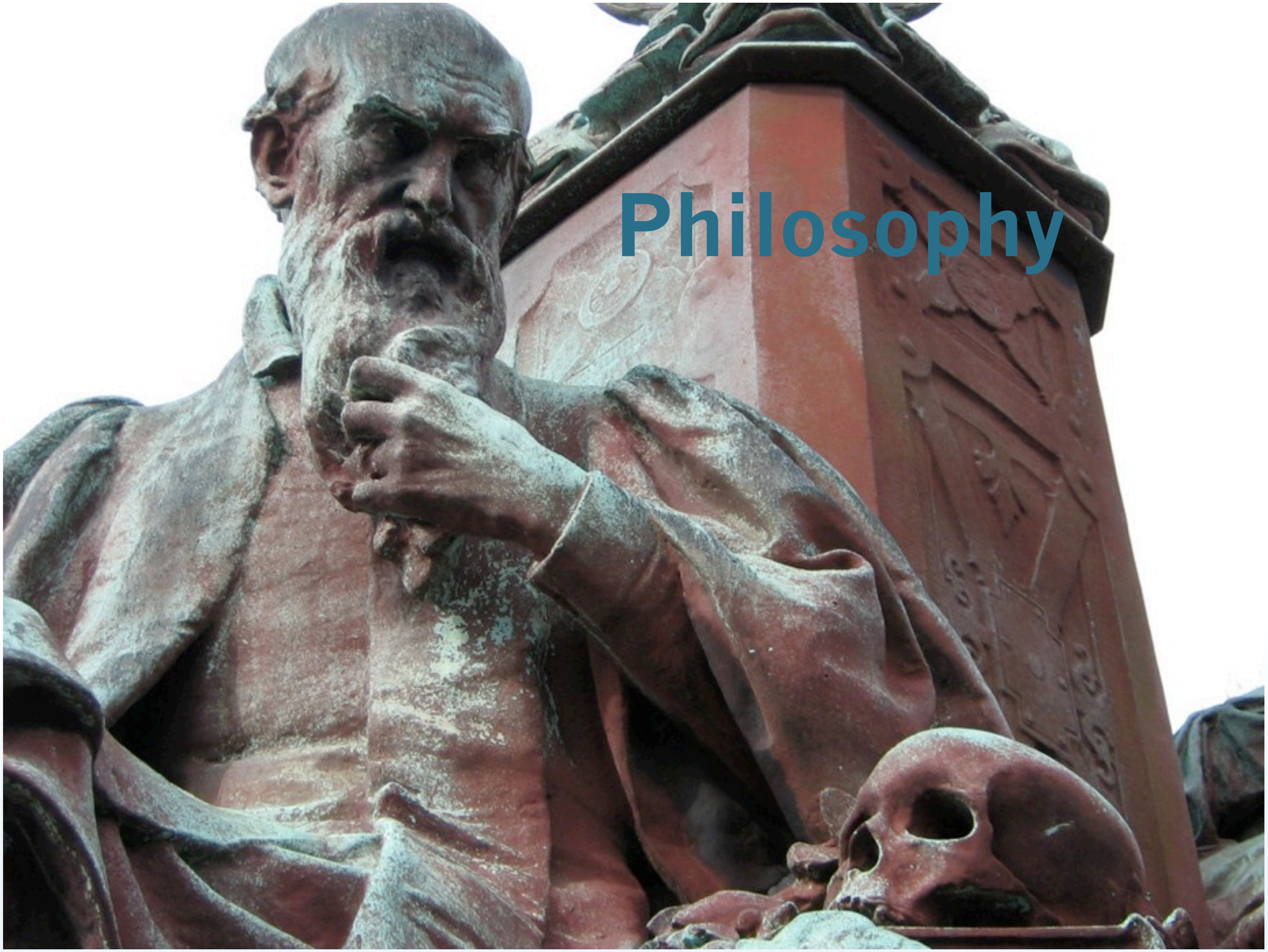
Examples

CAL POLY

Background

- Philosophy
- Epistemology
- Library Science

Philosophy



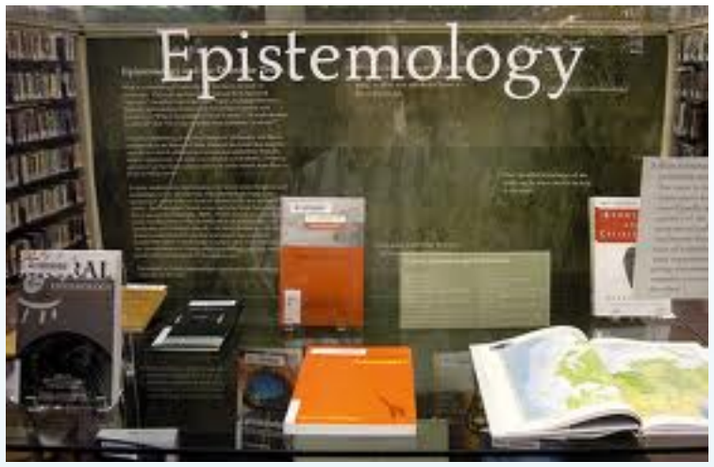
one of eight statues on University Avenue in Glasgow
Photo by liquidindian (Alan Miller)

http://images.cdn.fotopedia.com/flickr-427162166-original.jpg

Epistemology

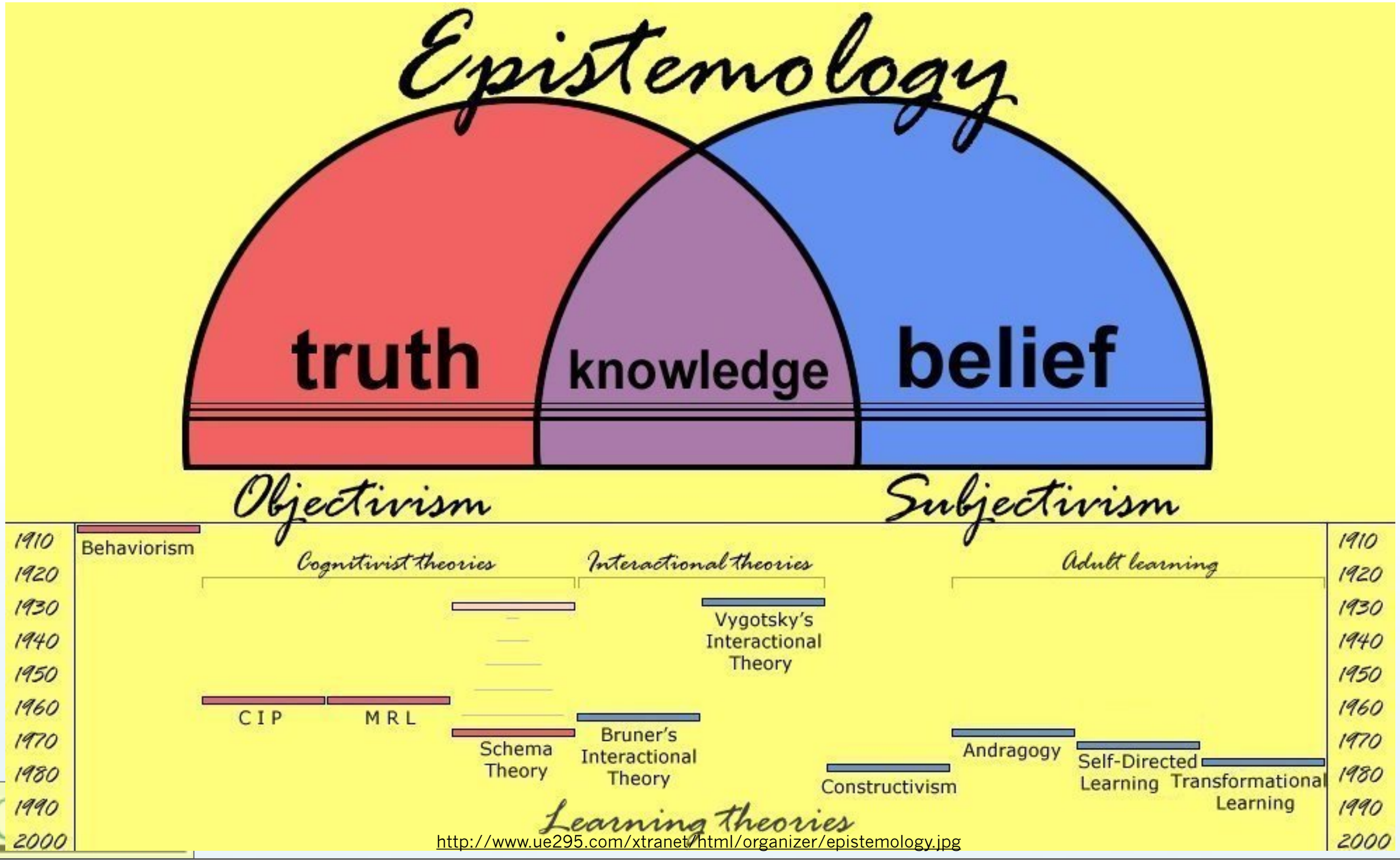
http://www.stagedive-magazine.de/joomla/images/stories/sanfrancisco/SF-thinker.JPG

- branch of philosophy concerned with the nature and scope (limitations) of knowledge



http://static.flickr.com/2321/2255746662_459e6b5c40.jpg

Epistomology and Learning Theories



http://www.ue295.com/xtranet/html/organizer/epistemology.jpg

Library Science

- **library catalog**
  - proxies for books and documents
  - multiple categories for single entities
  - index, keywords
- **categorization**
  - type of entity (book, article, thesis, special types)
  - Dewey Decimal System
- **service**
  - subject librarians

Duke Humfrey's Library (Photo by James Whitaker)

http://www.bodleian.ox.ac.uk/bodley

The card catalog in the nave of Sterling Memorial Library at Yale University. Picture by Henry Trotter, 2005.

http://commons.wikimedia.org/wiki/File:SML-Card-Catalog.jpg

Identification of Knowledge

- Object Selection
- Naming and Description

Object Selection

- **what constitutes a "knowledge object" that is relevant for a particular task or topic**
  - physical object, document, concept
- **how can this object be made available in the system**
- **example: library**
  - is it worth while to add an object to the library's collection
  - if so, how can it be integrated
    - physical document: book, magazine, report, etc.
    - digital document: file, data base, Web page, etc.

Naming and Description

- **names serve two important roles**
  - identification
    - ideally, a unique descriptor that allows the unambiguous selection of the object
    - often an ambiguous descriptor that requires context information
  - location
    - especially in digital systems, names are used as “address” for an object
- **names, descriptions and relationships to related objects are specified in listings**
  - dictionary, glossary, thesaurus, ontology, index

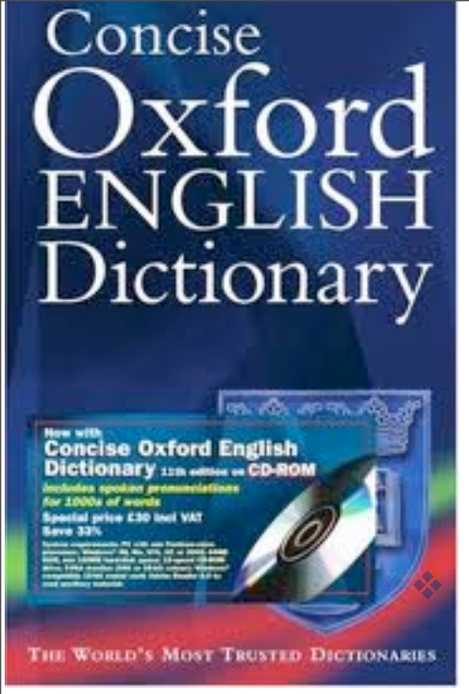
Knowledge Organization Methods

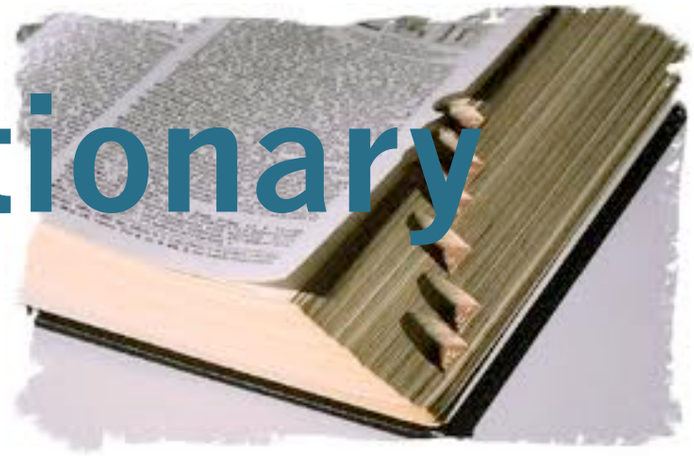
- **Naming and Description Devices**
  - index, glossary, dictionary, thesaurus, ontology
- **Natural Language (NL)**
  - Levels of NL Understanding
  - NL-based indexing
- **Categorization**
- **Ontologies**

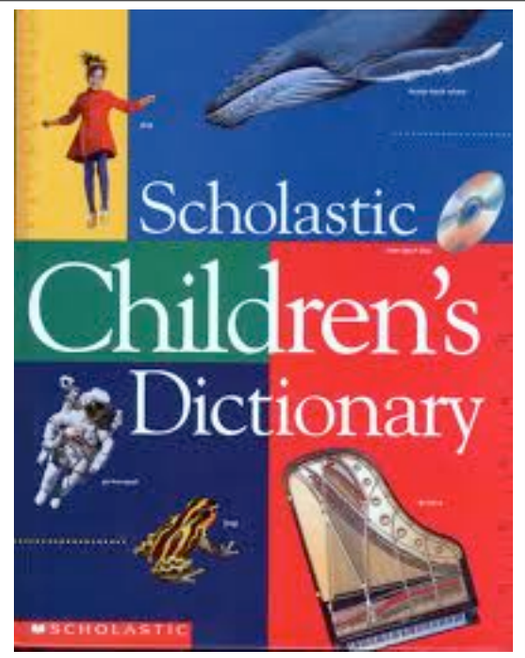
Naming and Description Methods

- **type**
  - dictionary, glossary, thesaurus
  - ontology
  - index
- **issues**
  - arrangement of terms
    - alphabetical, ordered by feature, hierarchical, arbitrary
  - purpose
    - explanation, unique identifier, clarification of relationships to other terms, access to further information

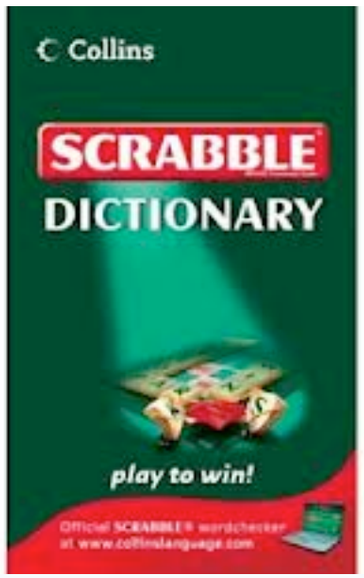
Dictionary

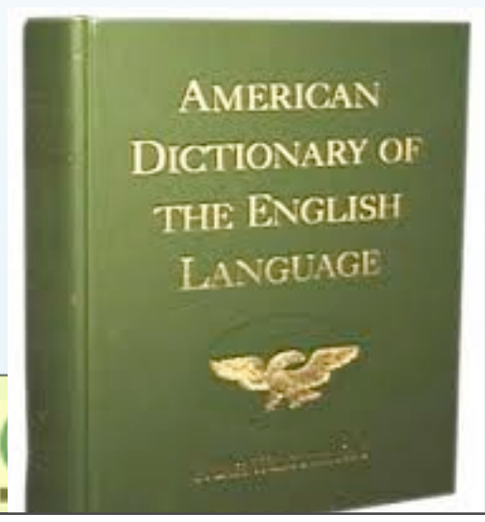


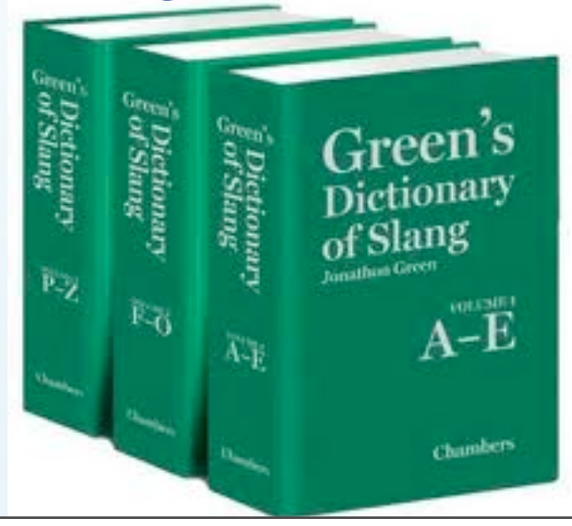


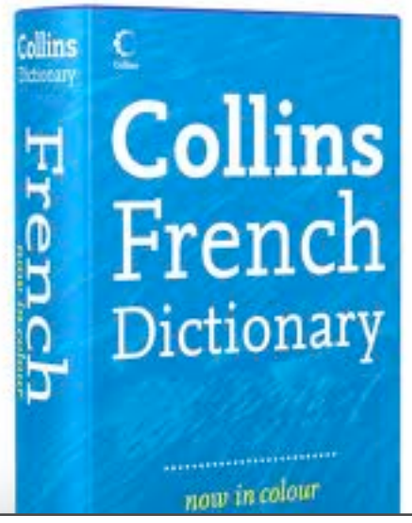


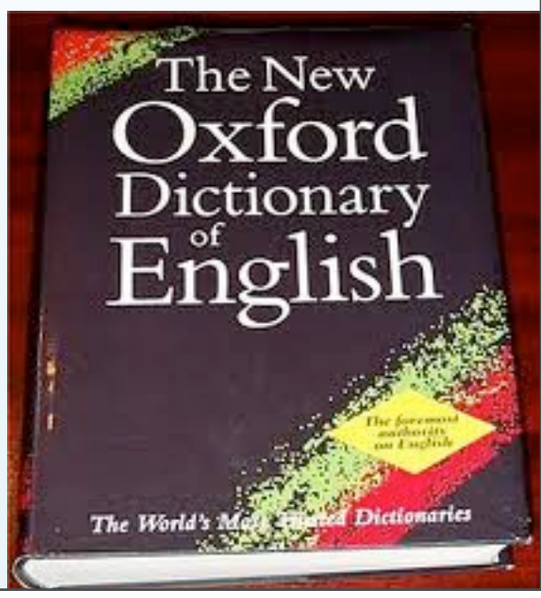
- list of words together with a short explanation of their meanings, or their translations into another language
- helpful for the identification of knowledge objects, and their distinction from related ones
- each entry in a dictionary may be considered an atomic knowledge object, with the word as name and "entry point"
  - may provide cross-references to related knowledge objects
- straightforward implementation in digital systems, and easy to integrate into knowledge management systems



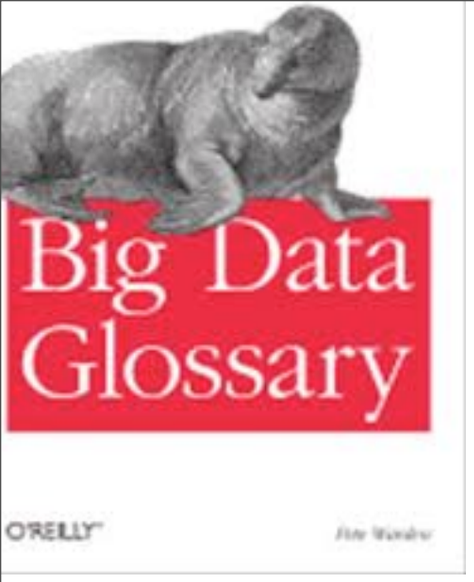




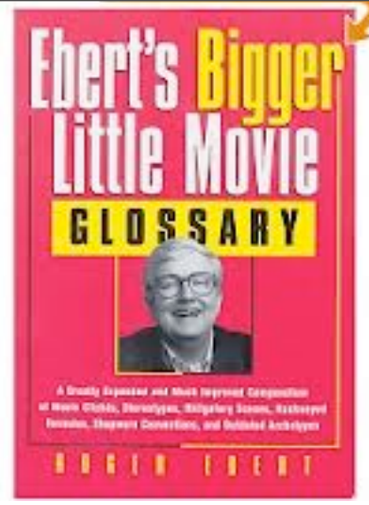




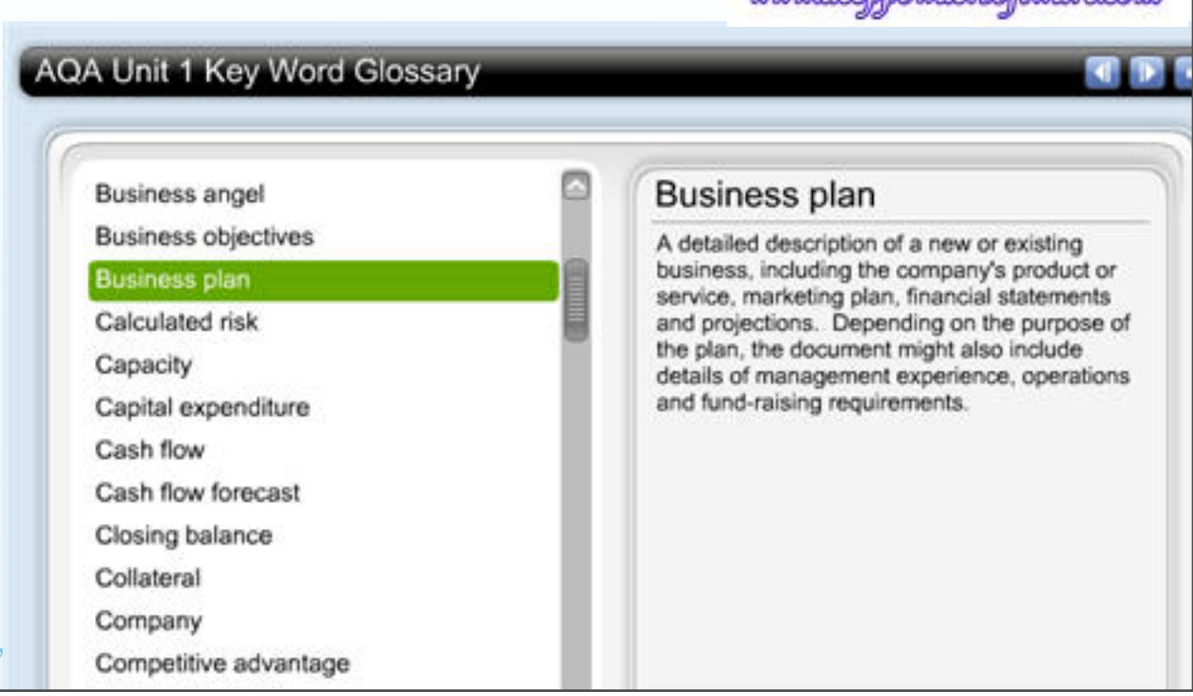
J. Kurfes

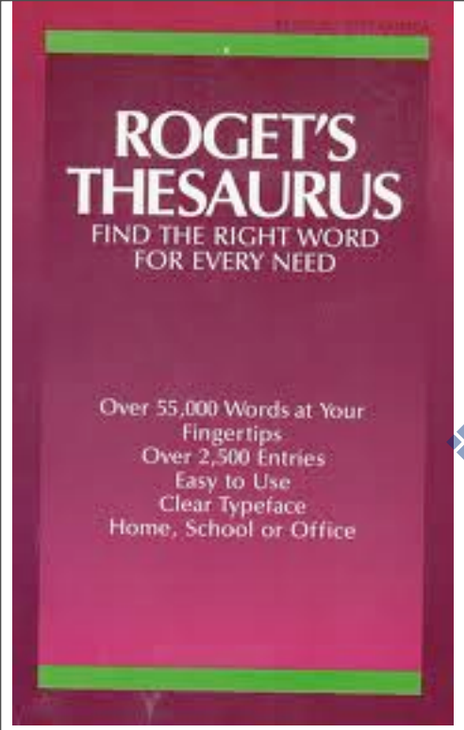


Glossary

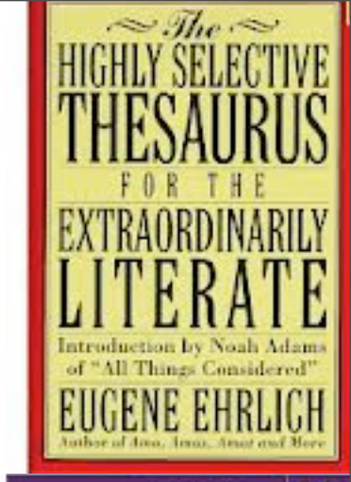


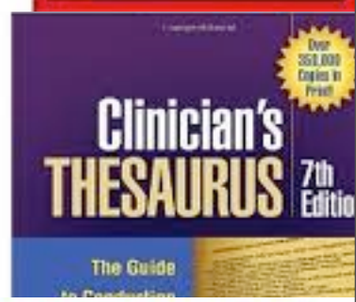
- **list of words, expressions, or technical terms with an explanation of their meanings**
  - usually restricted to a particular book, document, activity, or topic
- **provides a clarification of the intended meaning for knowledge objects**
- **otherwise similar to dictionary**



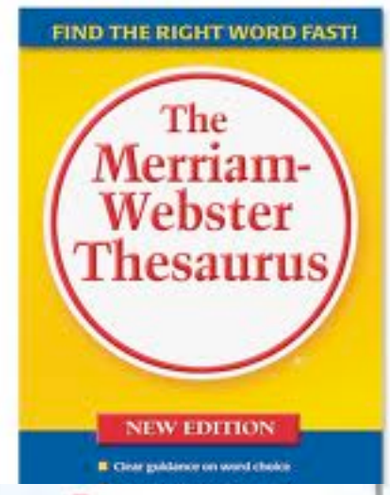


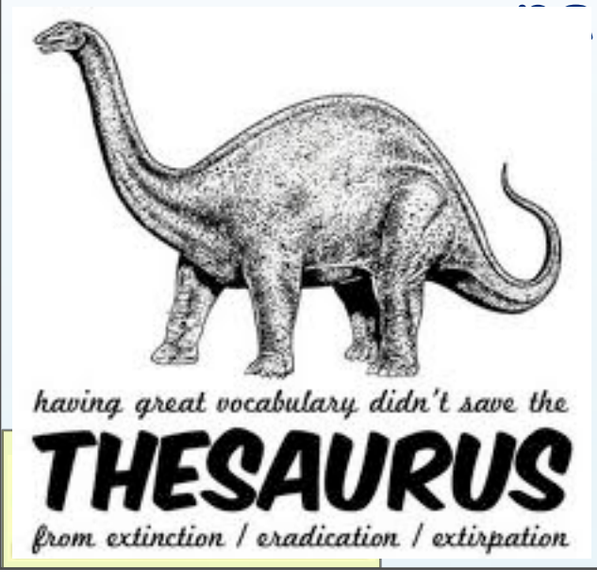
Thesaurus

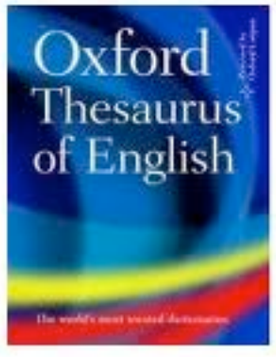




- **collection of synonyms (word sets with identical or similar meanings)**
  - frequently includes words that are related in some other way, e.g. antonyms (opposite meanings), homonyms (same pronunciation or spelling)
- **identifies and clarifies relationships between words**
  - not so much an explanation of their meanings
- **may be used to expand search queries in order to find relevant documents that may not contain a particular word**







Thesaurus Types

- knowledge-based
- linguistic
- statistical

[Liddy 2000]

Knowledge-based Thesaurus

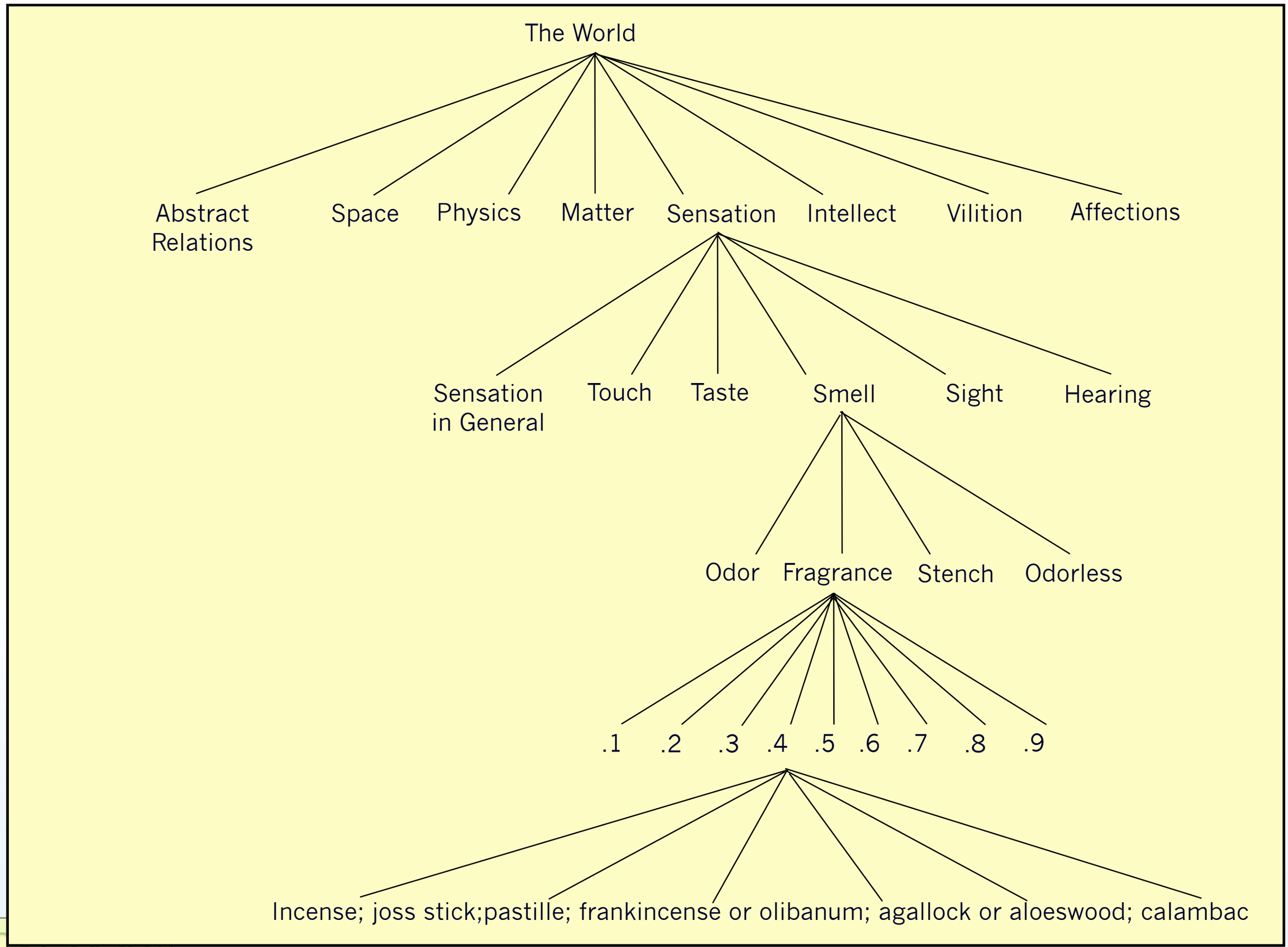
- manually constructed for a specific domain
- intended for human indexers and searchers
- contains
  - synonyms (“use for” UF)
  - more general (“broader term” BT)
  - more specific (“narrower” NT)
  - otherwise associated words (“related term” RT)
- example: “data base management systems”
  - UF data bases
  - BT file organization, management information systems
  - NT relational databases
  - RT data base theory, decision support systems

[Liddy 2000]

Linguistic Thesaurus

- **contains explicit concept hierarchies of several increasingly specified levels**
- **words in a group are assumed to be (near-) synonymous**
  - selection of the right sense for terms can be difficult
- **examples: Roget's, WordNet**
- **often used for query expansion**
  - synonyms (similar terms)
  - hyponyms (more specific terms; subclass)
  - hypernyms (more general terms; super-class)

[Liddy 2000]

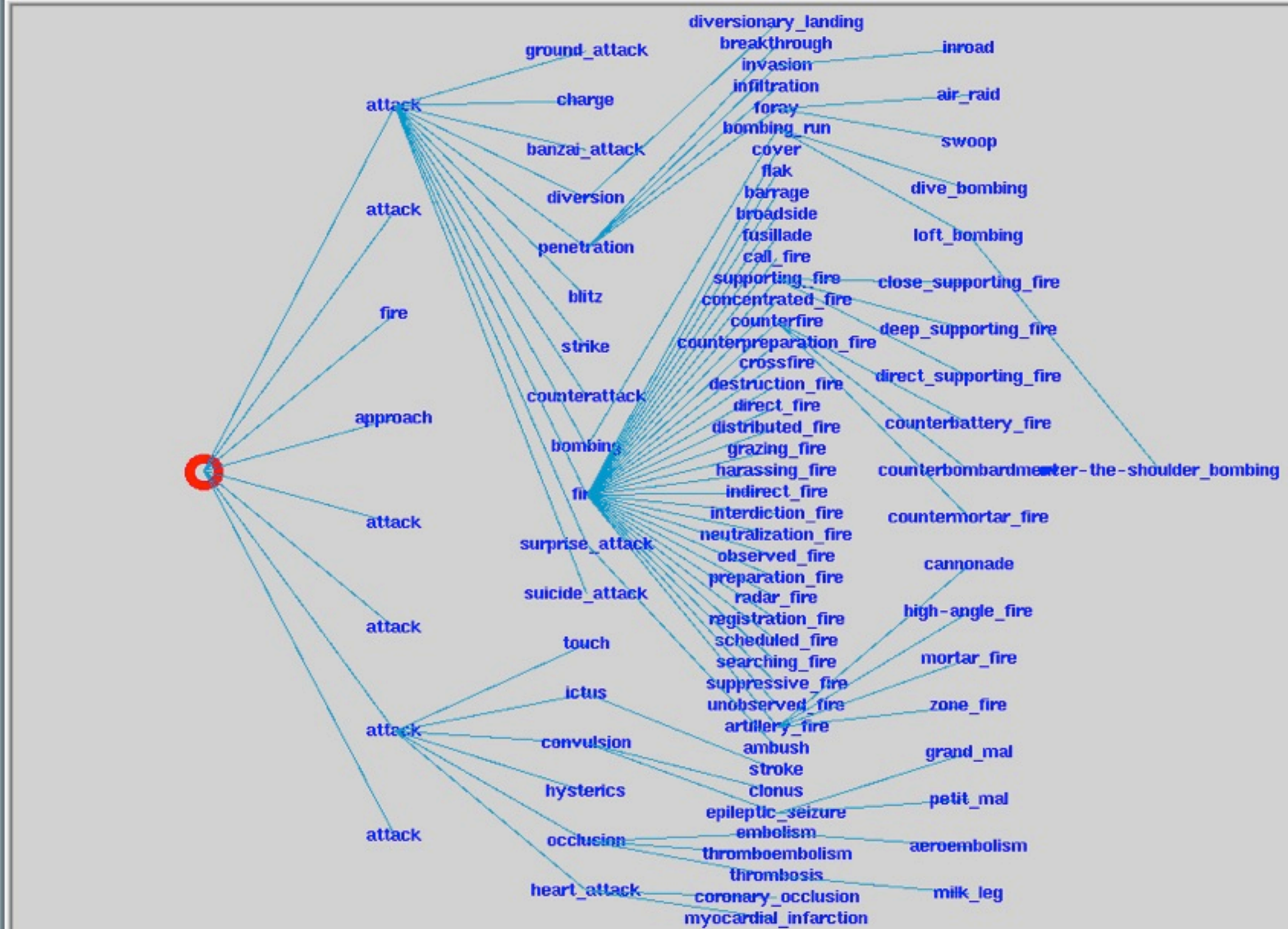

The World
Abstract Relations
Space
Physics
Matter
Sensation
Intellect
Vilition
Affections
Sensation in General
Touch
Taste
Smell
Sight
Hearing
Odor
Fragrance
Stench
Odorless
.1
.2
.3
.4
.5
.6
.7
.8
.9
Incense; joss stick;pastille; frankincense or olibanum; agallock or aloeswood; calambac

[Liddy 2000]

WNTool

Console | Database | Edit | Word Lookup | Help

Target Word: attack | Relation: Hyponym | Part of Speech: Noun | Class: Tops



Current Synset: (3571) attack,onslaught,onset,onrush | Definition: the beginning of an offensive; "the attack began at dawn"

Query Expansion in Search Engines

- look up each word in Word Net
- if the word is found, the set of synonyms from all Synsets are added to the query representation
- weigh each added word as 0.8 rather than 1.0
- **results better than plain SMART**
  - variable performance over queries
  - major cause of error: the use of ambiguous words' Synsets
- **general thesauri such as Roget's or WordNet have not been shown conclusively to improve results**
  - may sacrifice precision to recall
  - not domain specific
  - not sense disambiguated

[Liddy 2000, Voorhees 1993]

Statistical Thesaurus

- **automatic thesaurus construction**
  - classes of terms produced are not necessarily synonymous, nor broader, nor narrower
  - rather, words that tend to co-occur with head term
  - effectiveness varies considerably depending on technique used

[Liddy 2000]

Automatic Thesaurus Construction (Salton)

- **document collection based**
  - based on index term similarities
  - compute vector similarities for each pair of documents
  - if sufficiently similar, create a thesaurus entry for each term which includes terms from similar document

[Liddy 2000]

Sample Automatic Thesaurus Entries

408 dislocation
- junction
- minority-carrier
- point contact
- recombine
- transition

409 blast-cooled
- heat-flow
- heat-transfer

410 anneal
- strain

411 coercive
- demagnetize
- flux-leakage
- hysteresis
- induct
- insensitive
- magnetoresistance
- square-loop
- threshold

412 longitudinal
- transverse

[Liddy 2000]

Dynamic Automatic Thesaurus Construction

- **thesaurus short-cut**
  - run at query time
  - take all terms in the query into consideration at once
  - look at frequent words and phrases in the top retrieved documents and add these to the query
  - = automatic relevance feedback

[Liddy 2000]

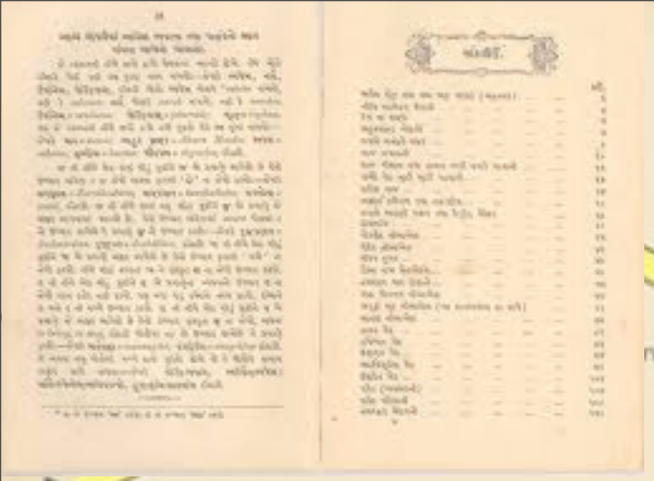
Expansion by Association Thesaurus

Query: ***Impact of the 1986 Immigration Law***

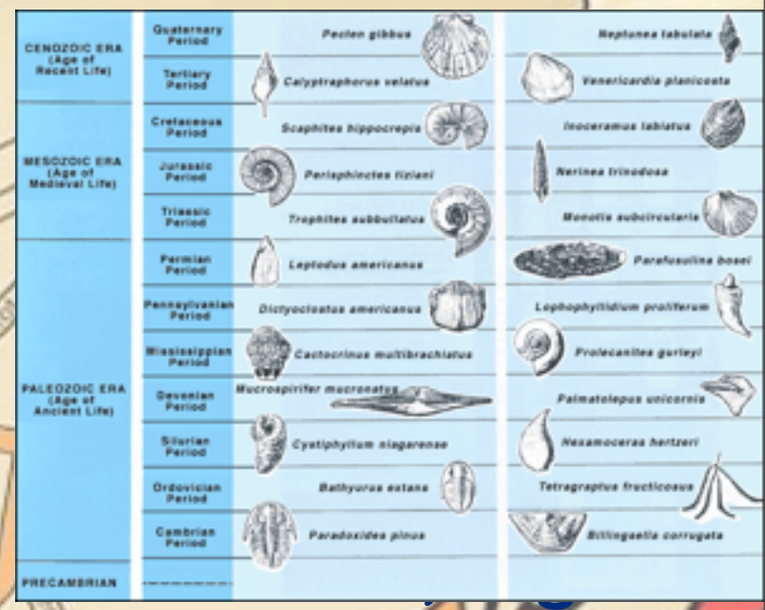
Phrases retrieved by association in corpus

- *illegal immigration*
- amnesty program
- immigration reform law
- editorial page article
- naturalization service
- civil fines
- new immigration law
- legal immigration
- employer sanctions
- *statutes*
- applicability
- seeking amnesty
- legal status
- immigration act
- undocumented workers
- guest worker
- sweeping immigration law
- undocumented aliens

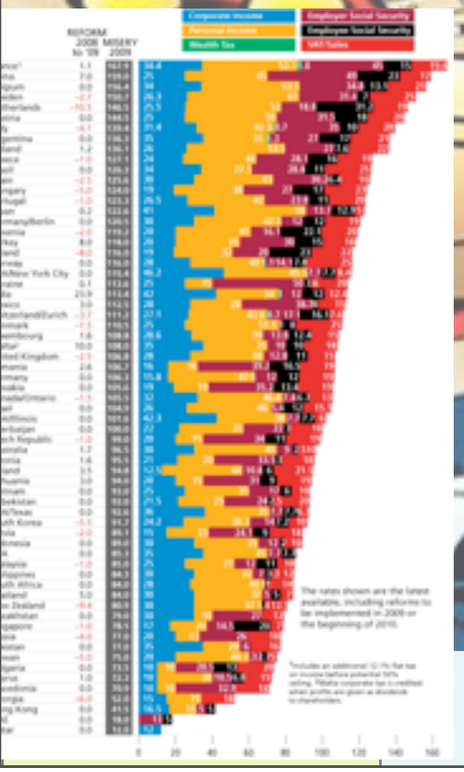
[Liddy 2000]

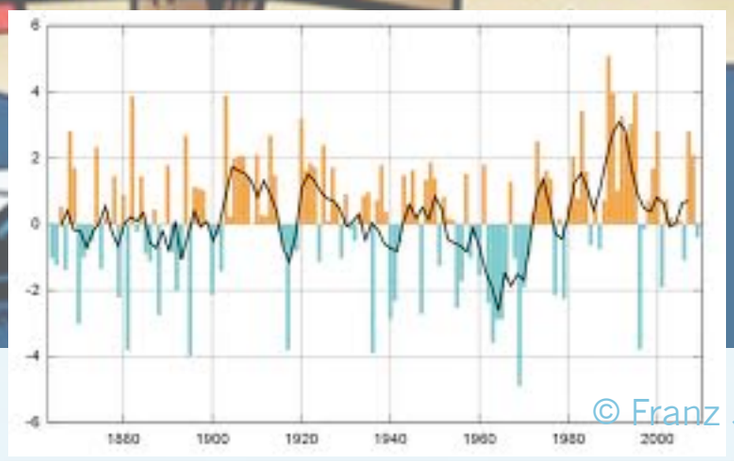


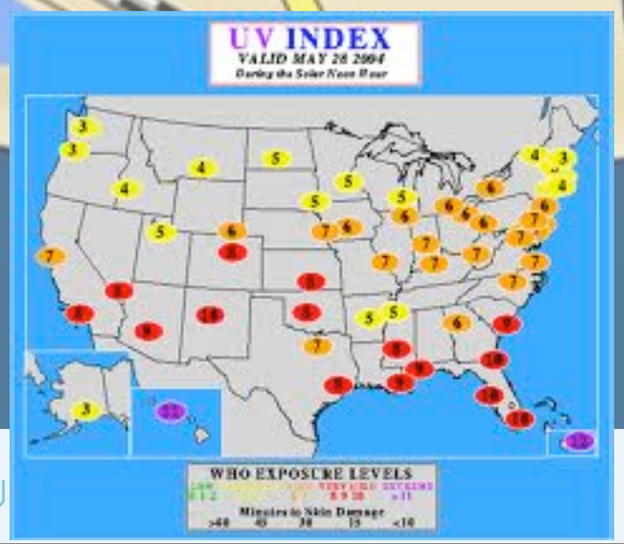
Index



- listing of words that appear in a set of with pointers to the locations where they appear
- provides a reference to further information concerning a particular word or concept
- constitutes the basis for computer-based search engines







Indexing

- **the process of creating an index from a set of documents**
  - one of the core issues in Information Retrieval
- **manual indexing**
  - controlled vocabularies, humans go through the documents
- **semi-automatic**
  - humans are in control, machines are used for some tasks
- **automatic**
  - statistical indexing
  - natural-language based indexing

Natural Language Methods

- Natural Language Processing
- Natural Language Understanding
- NLP-based Indexing

Natural Language Processing

- **a range of computational techniques for analyzing and representing naturally occurring texts**
  - at one or more levels of linguistic analysis
  - for the purpose of achieving human-like language processing
  - for a range of tasks or applications

[Liddy 2000]

Ontologies

- description
- “representational promiscuity”
- ontology types
- usage of ontologies
  - domain standards and vocabularies

- **ntology development**
  - development process
  - specification languages

Ontology

- **examines the relationships between words, and the corresponding concepts and objects**
  - in practice, it often combines aspects of thesaurus and dictionary
  - frequently uses a graph-based visual representation to indicated relationships between words
- **used to identify and specify a vocabulary for a particular subject or task**

Ontology in Computer Science

- related efforts since 1970s in AI, KBS
- more systematic approaches in the 1980s, 1990s
- formal definition as technical term by Tom Gruber, Stanford KSL
  - "An ontology is a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents. This definition is consistent with the usage of ontology as set of concept definitions, but more general. And it is a different sense of the word than its use in philosophy."
    - Gruber, T. (2001). "What is an Ontology?". Stanford University. Retrieved 2013-01-28.
    - see also Gruber, T. (2009). “Ontology” in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (Eds.), Springer-Verlag, 2009

The Notion of Ontology

- **ontology**
  explicit specification of a shared conceptualization that holds in a particular context

- **captures a viewpoint on a domain:**
  - taxonomies of species
  - physical, functional, & behavioral system descriptions
  - task perspective: instruction, planning

[Schreiber 2000]

Ontology Types

- domain-oriented
  - domain-specific
    - medicine => cardiology => rhythm disorders
    - traffic light control system
  - domain generalizations
    - components, organs, documents
- task-oriented
  - task-specific
    - configuration design, instruction, planning
  - task generalizations
    - problems solving, e.g. upml
- generic ontologies
  - “top-level categories”
  - units and dimensions

[Schreiber 2000]

Using Ontologies

- **ontologies needed for an application are typically a mix of several ontology types**
  - technical manuals
    - device terminology: traffic light system
    - document structure and syntax
    - instructional categories
  - e-commerce
- **raises need for**
  - modularization
  - integration
    - import/export
    - mapping

[Schreiber 2000]

Domain Standards and Vocabularies As Ontologies

- **example: Art and Architecture Thesaurus (AAT)**
- **contains ontological information**
  - AAT: structure of the hierarchy
- **structure needs to be "extracted"**
  - not explicit
- **can be made available as an ontology**
  - with help of some mapping formalism
- **lists of domain terms are sometimes also called "ontologies"**
  - implies a weaker notion of ontology
  - scope typically much broader than a specific application domain
  - example: domain glossaries, wordnet
  - contain some meta information: hyponyms, synonyms, text

[Schreiber 2000]

Ontology Specification

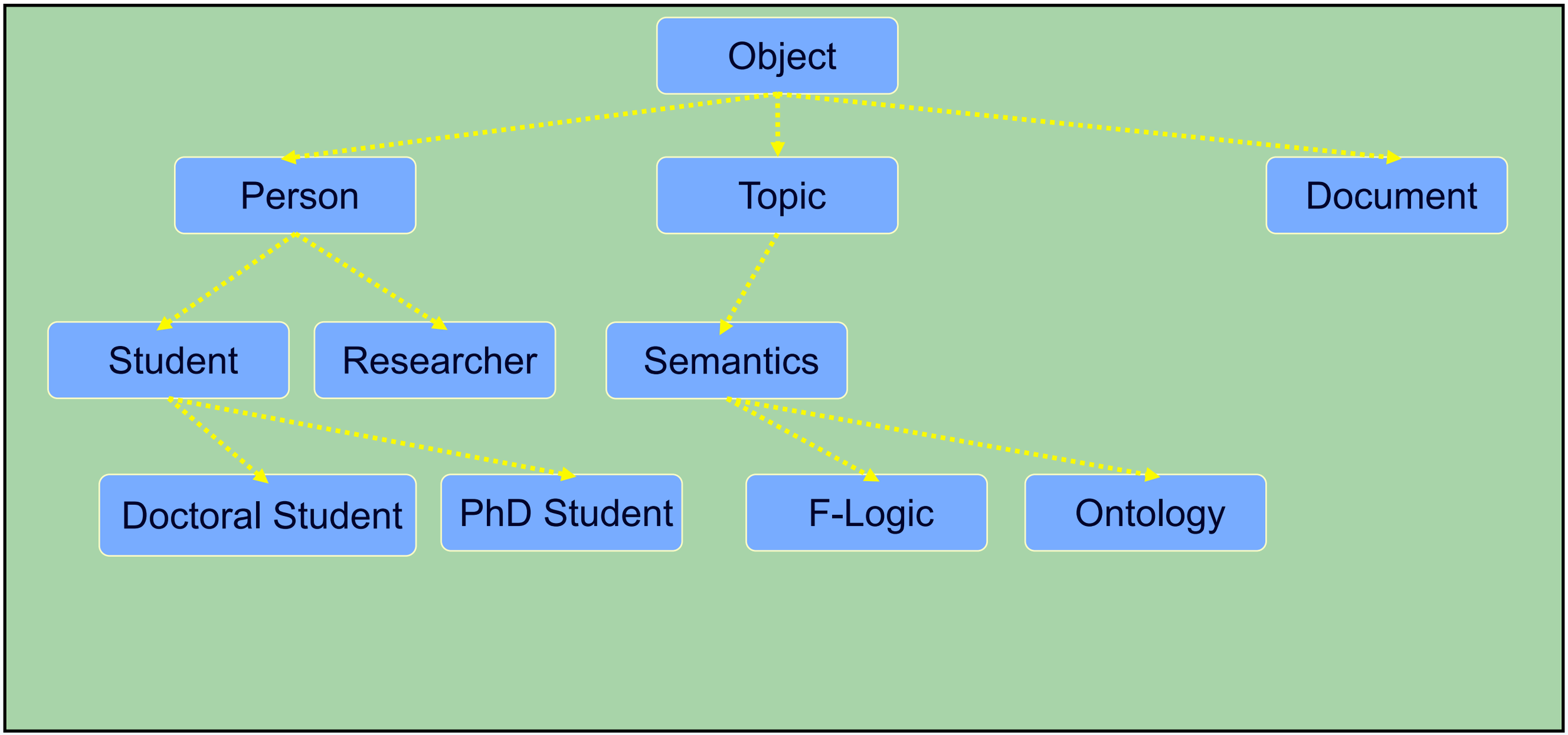
- **many different languages**
  - KIF
  - Ontolingua
  - Express
  - LOOM
  - UML
  - XML to the rescue: **Web Ontology Language** (**OWL**)
- **common basis**
  - class (concept)
  - subclass with inheritance
  - relation (slot)

[Schreiber 2000]

From Taxonomies to Ontologies

- **Taxonomy**
  - strict hierarchy
- **Thesaurus**
  - hierarchy plus synonyms and other relations between words
- **Topic Map**
  - additional relations between concepts
    - across the hierarchy
  - properties of concepts
- **Ontology**
  - rules specifying the structure of the concept space
  - instances of concepts

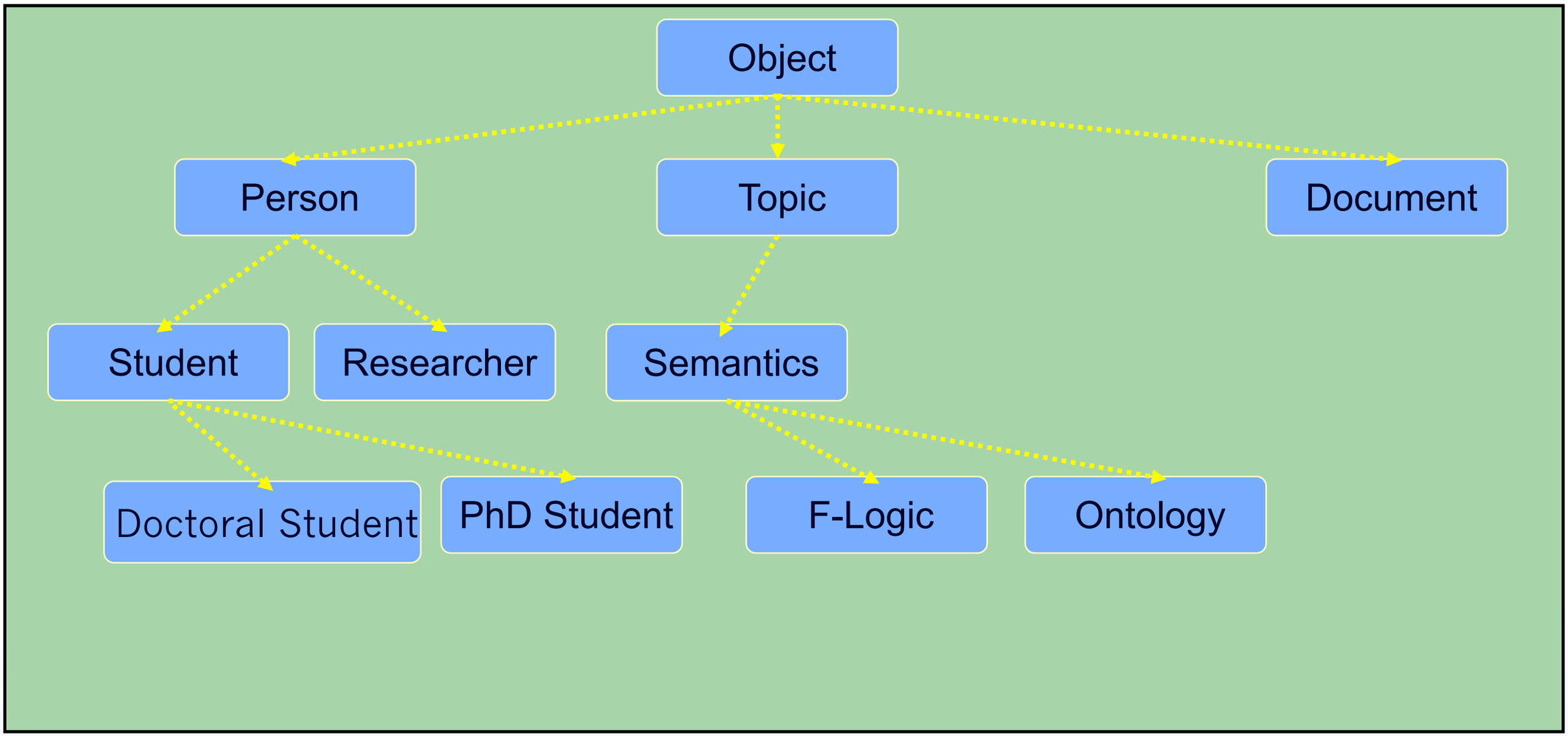
Taxonomy



Taxonomy: Segmentation, classification and ordering of elements into a classification system according to their relationships between each other

[Hotho, Sure, 2003]

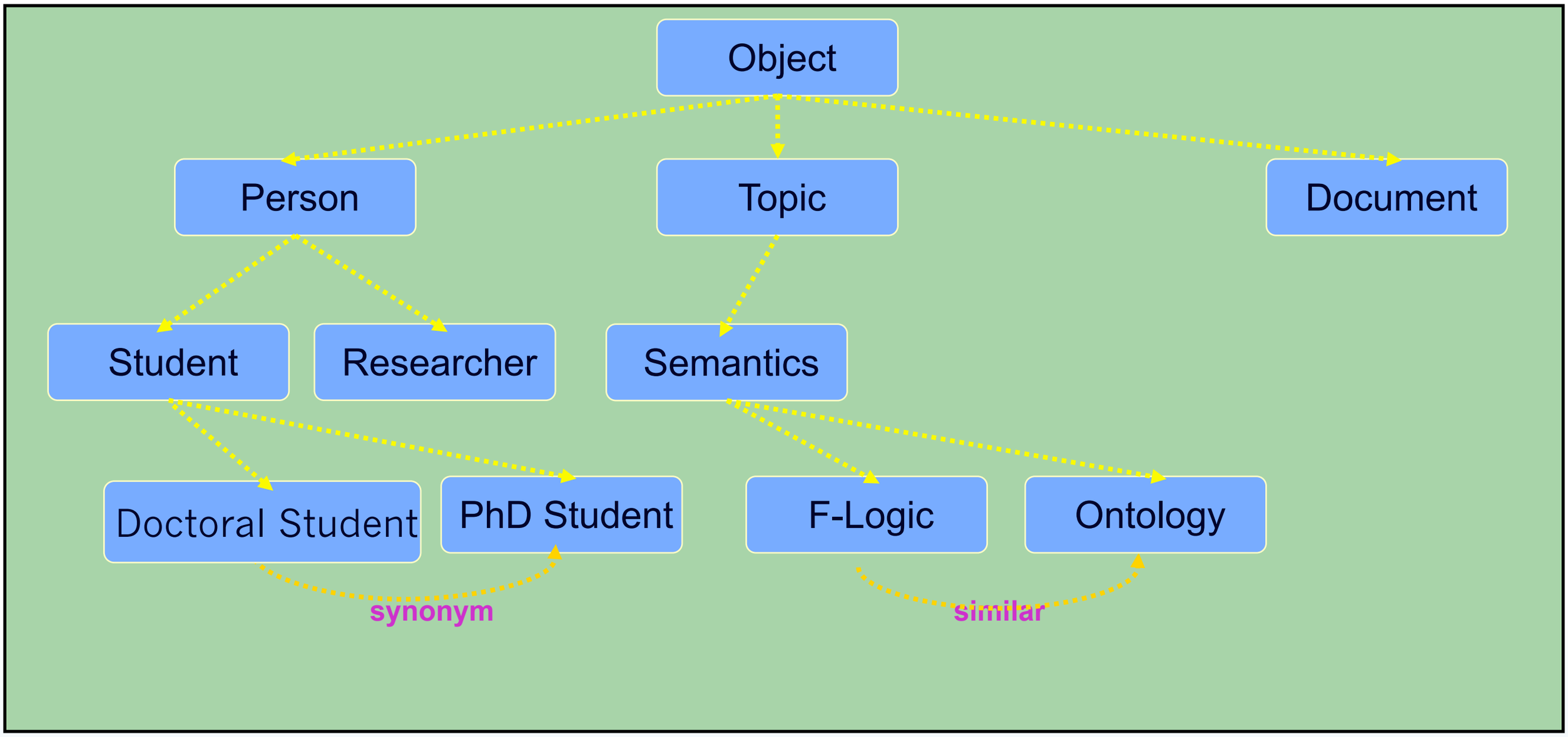
Thesaurus



- Terminology for specific domain
- Graph with primitives, 2 fixed relationships (similar, synonym), sometimes additional relationships (antonym, homonym, ...)
- originated from bibliography

[Hotho, Sure, 2003]

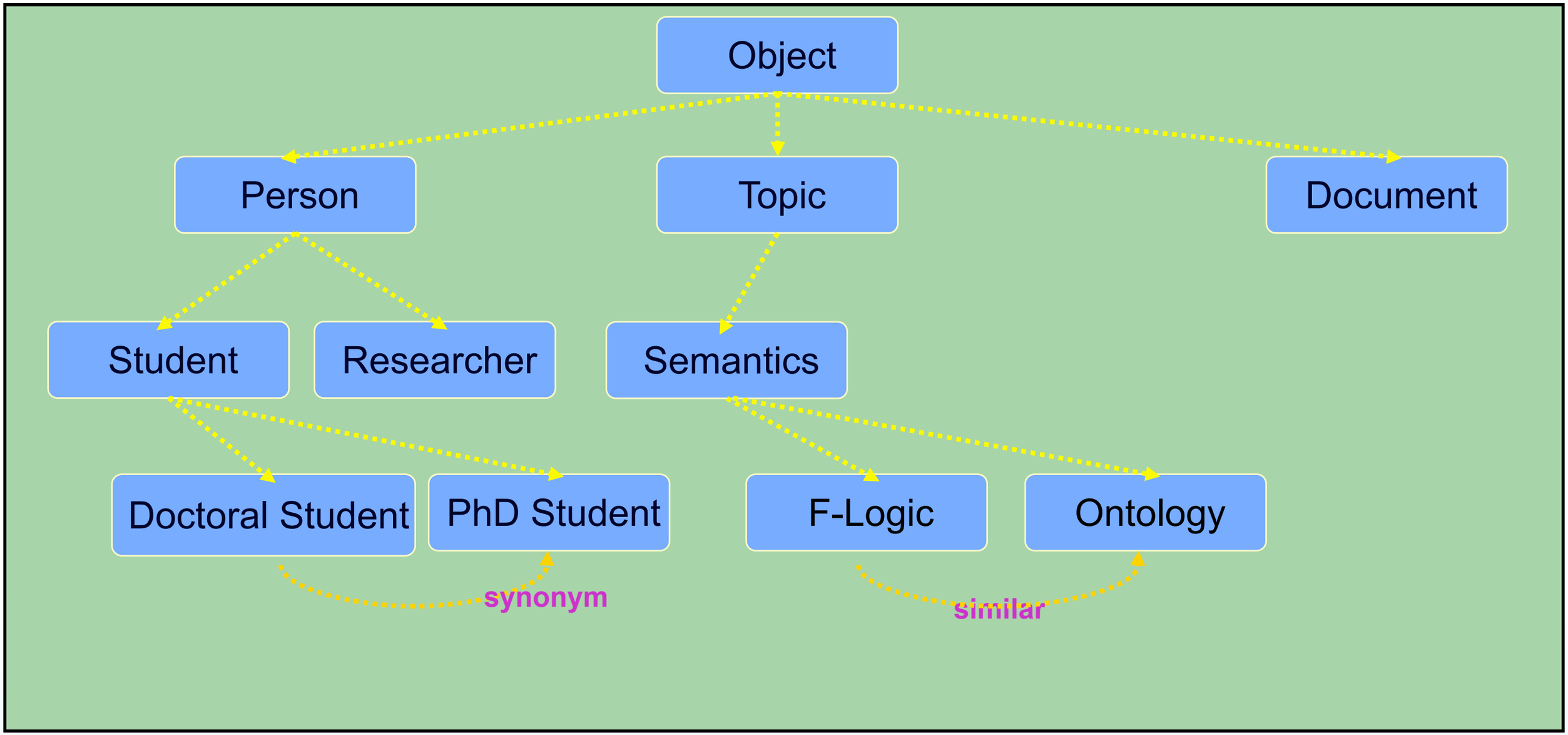
Thesaurus



- Terminology for specific domain
- Graph with primitives, 2 fixed relationships (similar, synonym), sometimes additional relationships (antonym, homonym, ...)
- originated from bibliography

[Hotho, Sure, 2003]

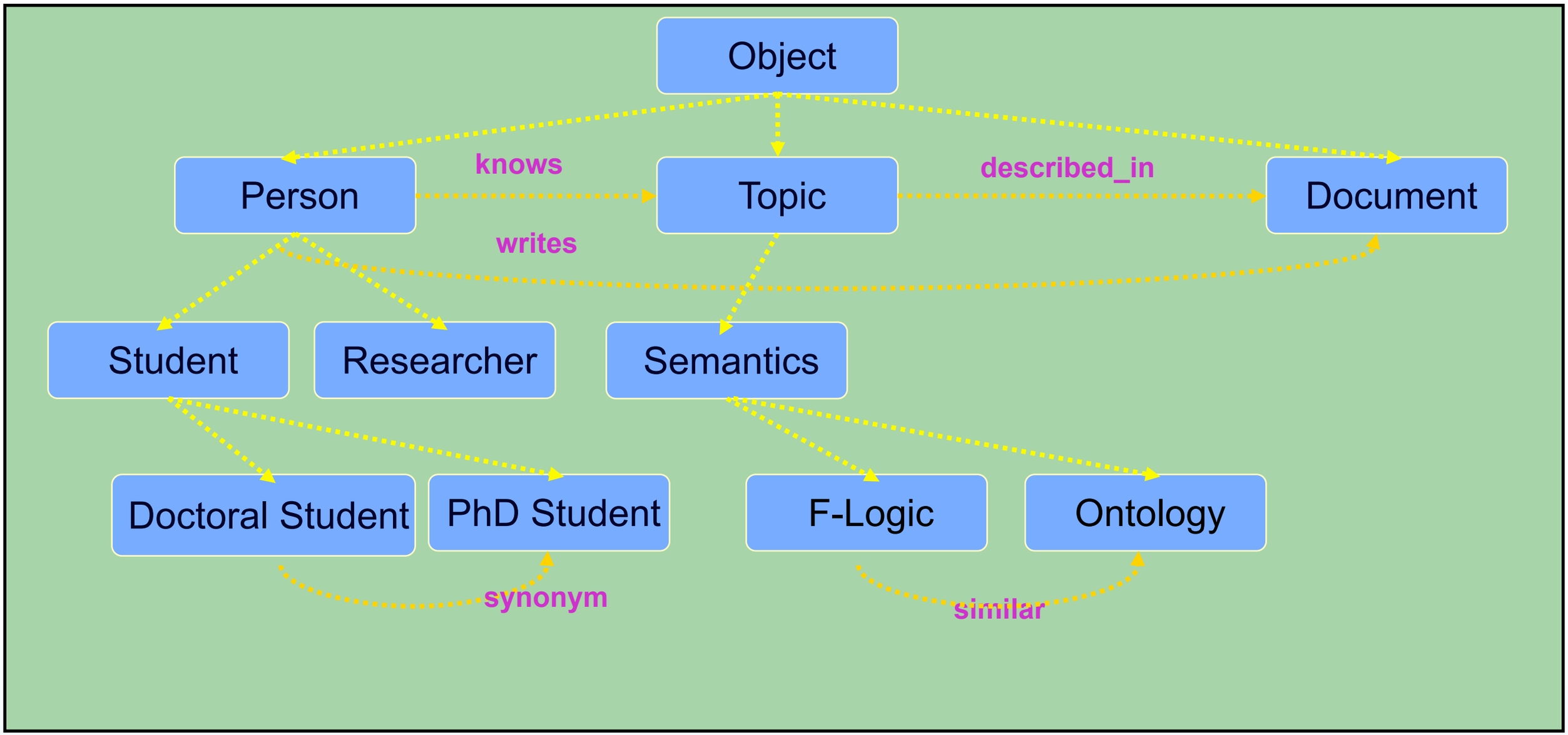
Topic Map



- Topics (nodes), relationships and *occurrences* (to documents)
- ISO-Standard
- typically for navigation and visualization

[Hotho, Sure, 2003]

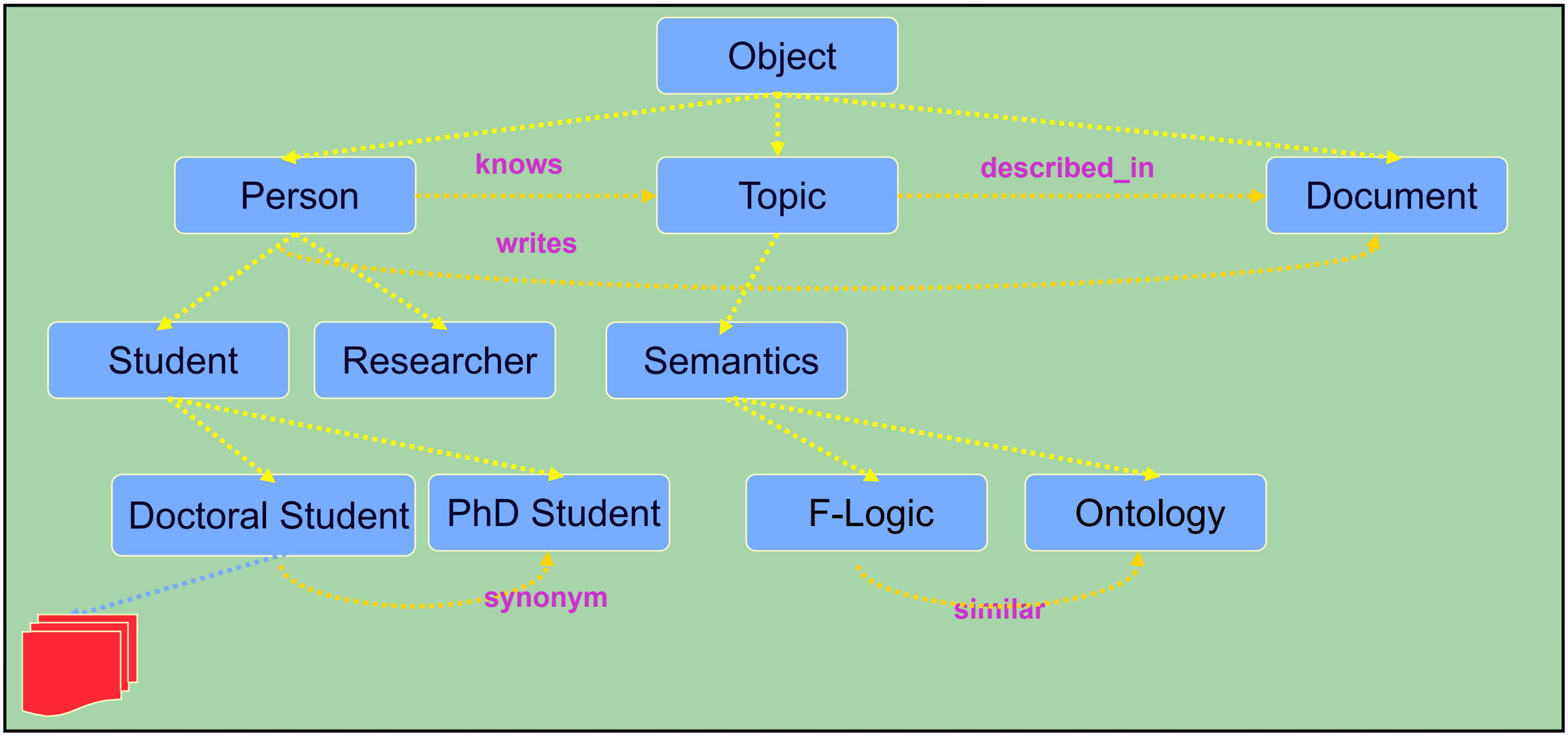
Topic Map



- Topics (nodes), relationships and *occurrences* (to documents)
- ISO-Standard
- typically for navigation and visualization

[Hotho, Sure, 2003]

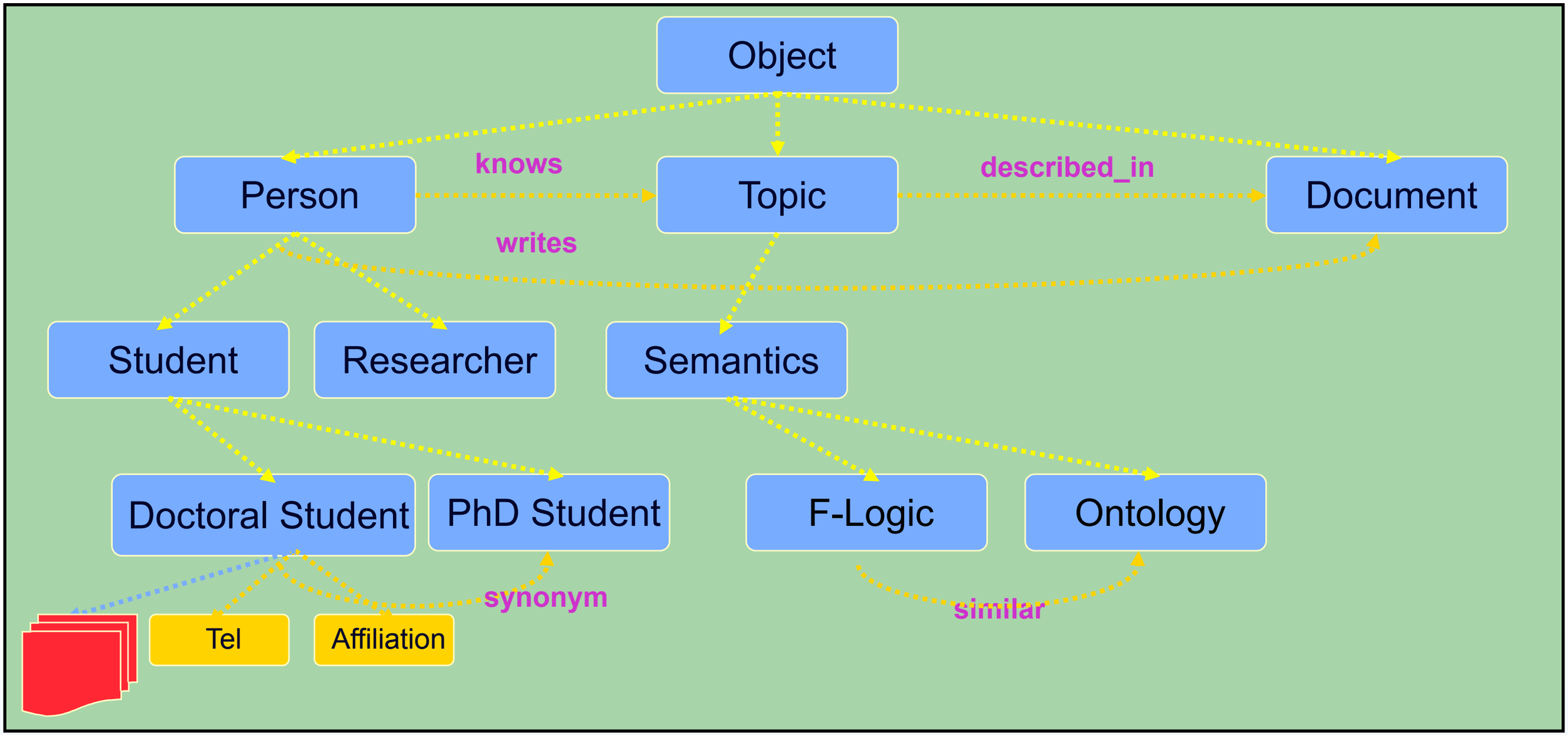
Topic Map



- Topics (nodes), relationships and *occurrences* (to documents)
- ISO-Standard
- typically for navigation and visualization

[Hotho, Sure, 2003]

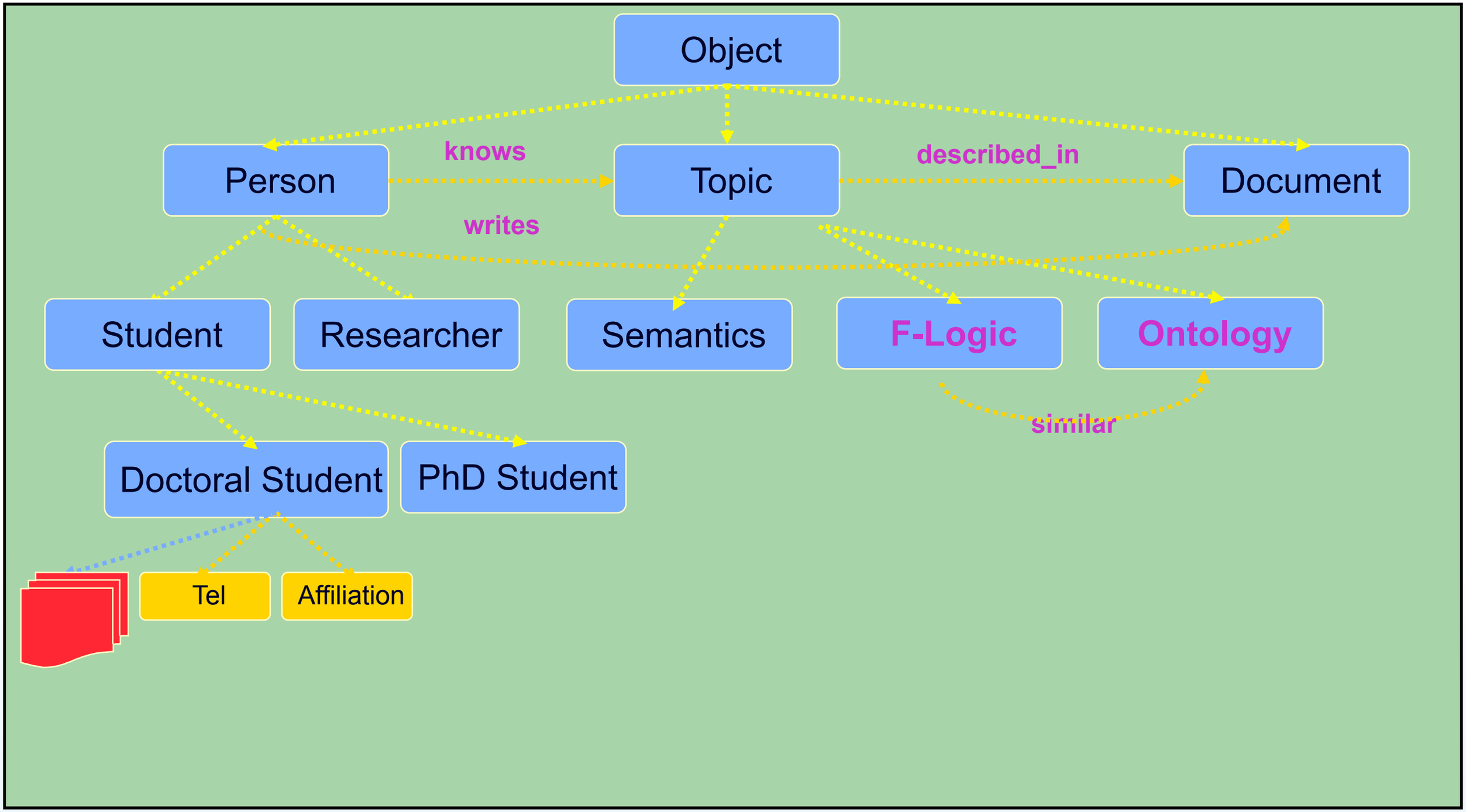
Topic Map



- Topics (nodes), relationships and *occurrences* (to documents)
- ISO-Standard
- typically for navigation and visualization

[Hotho, Sure, 2003]

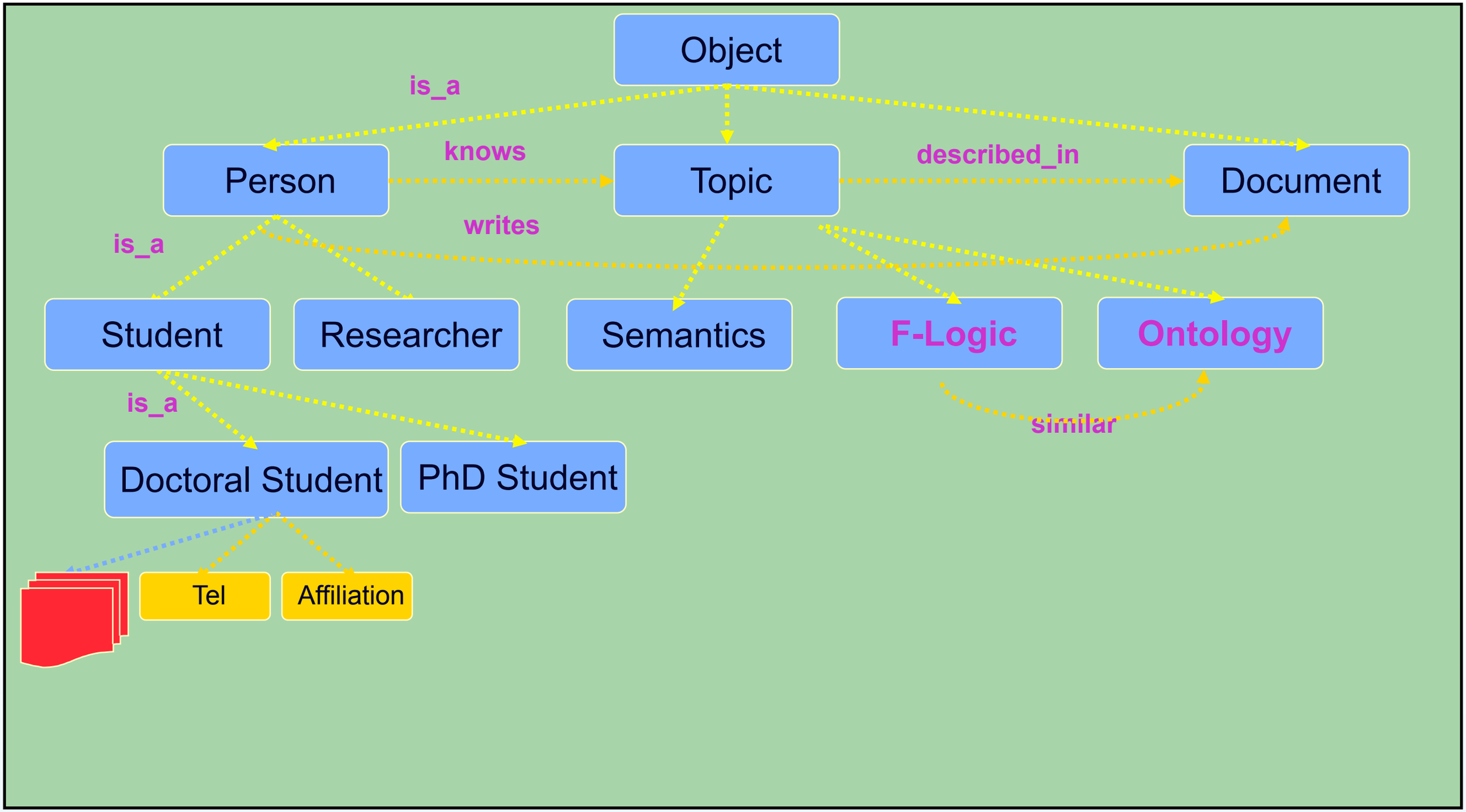
Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

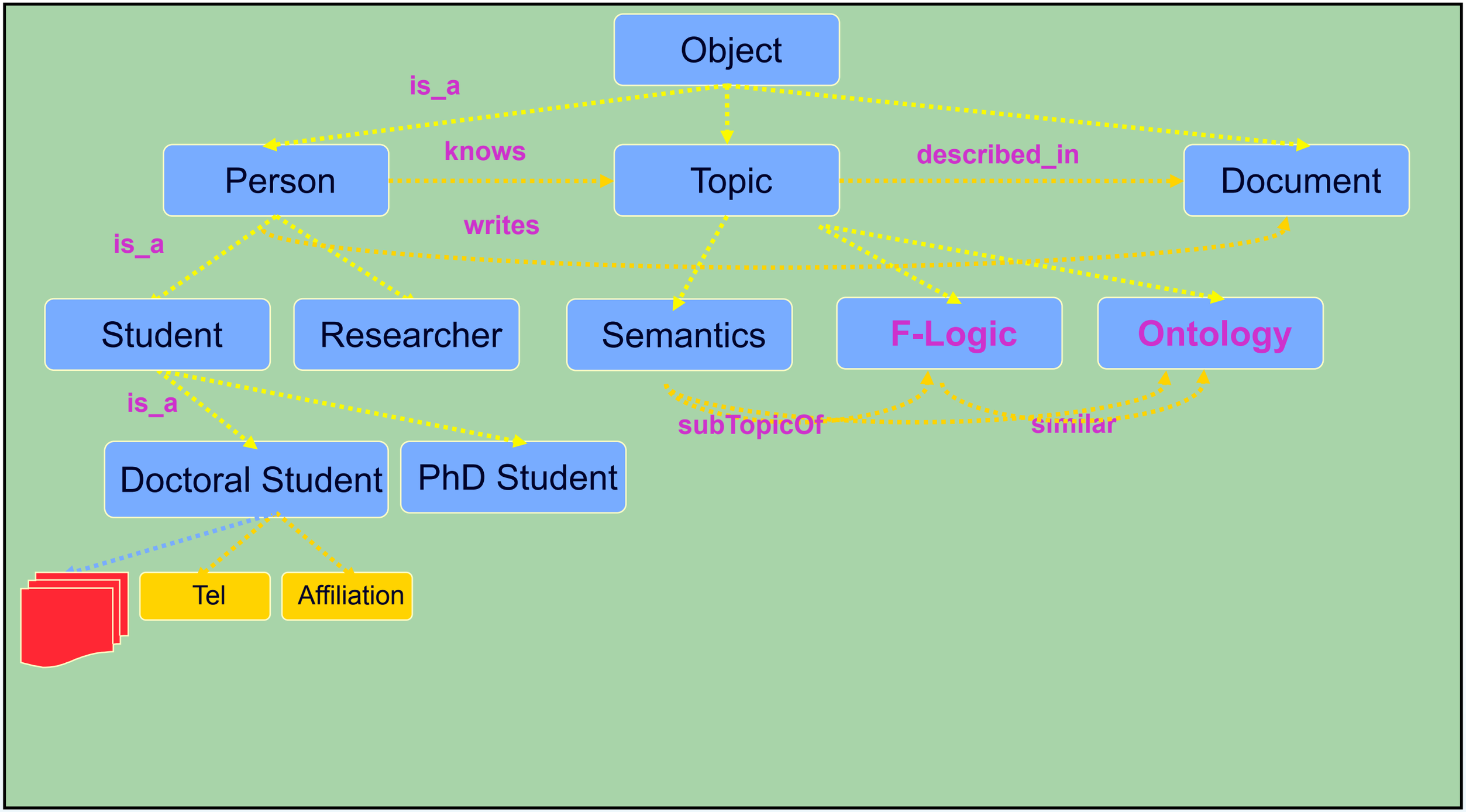
Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

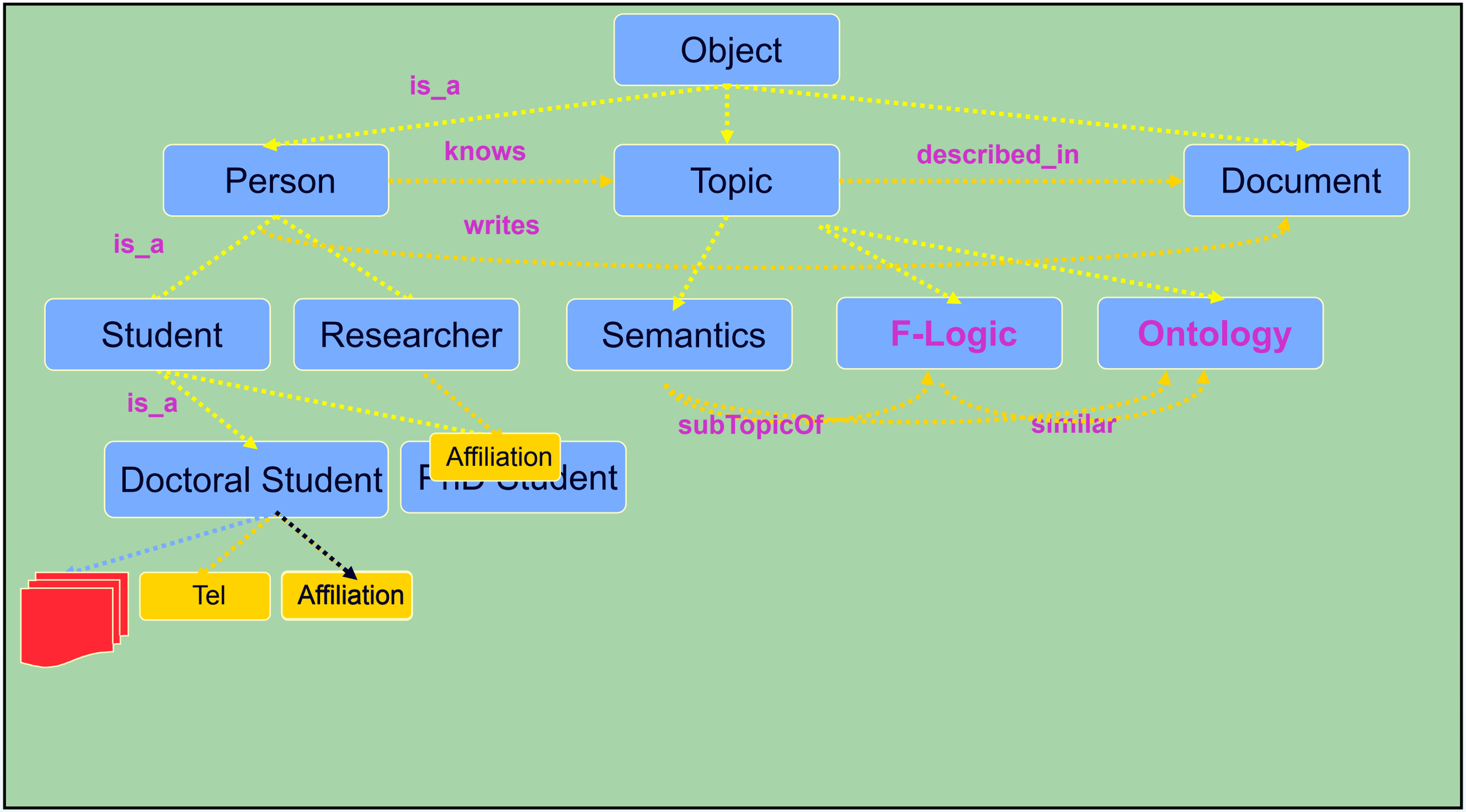
Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

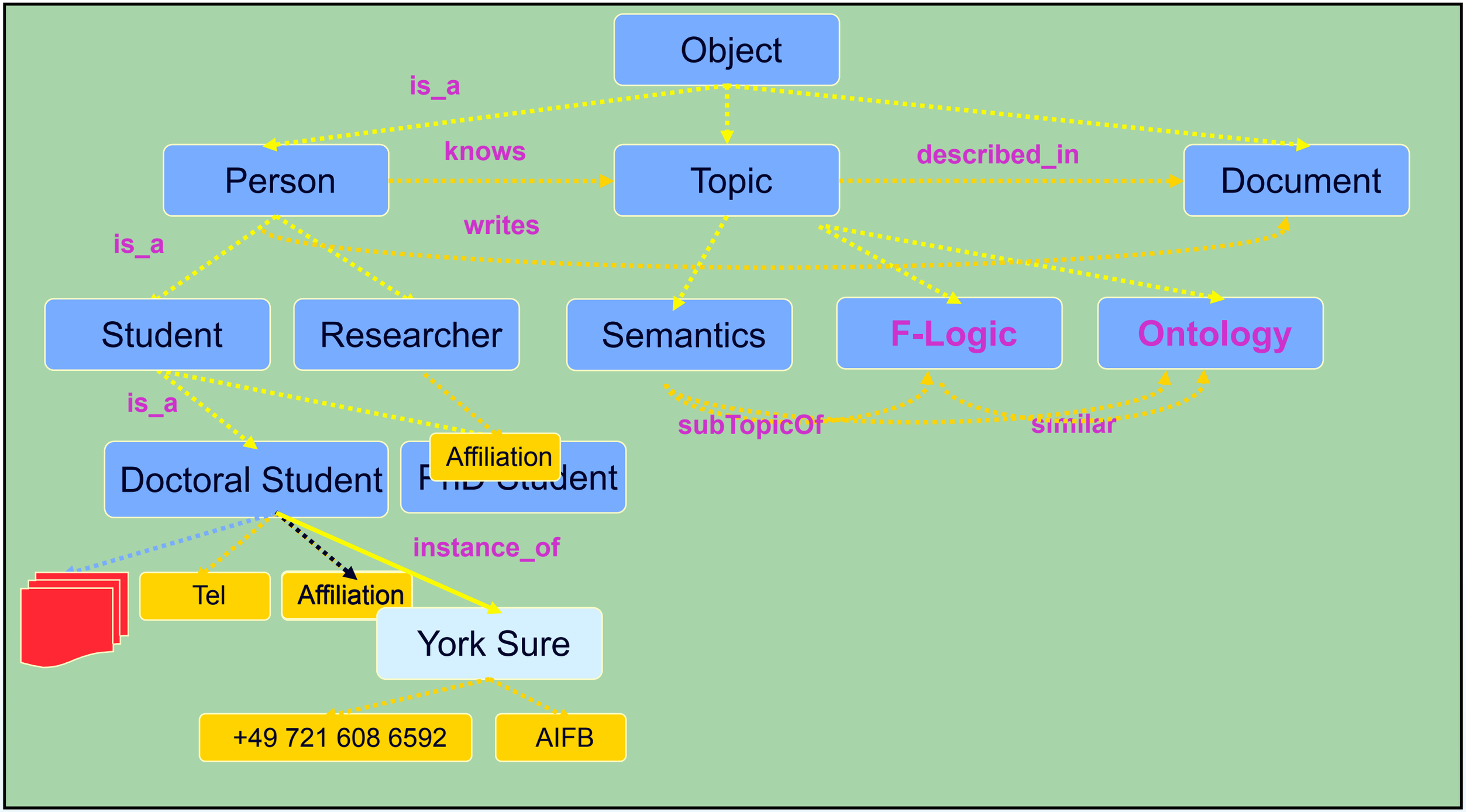
Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

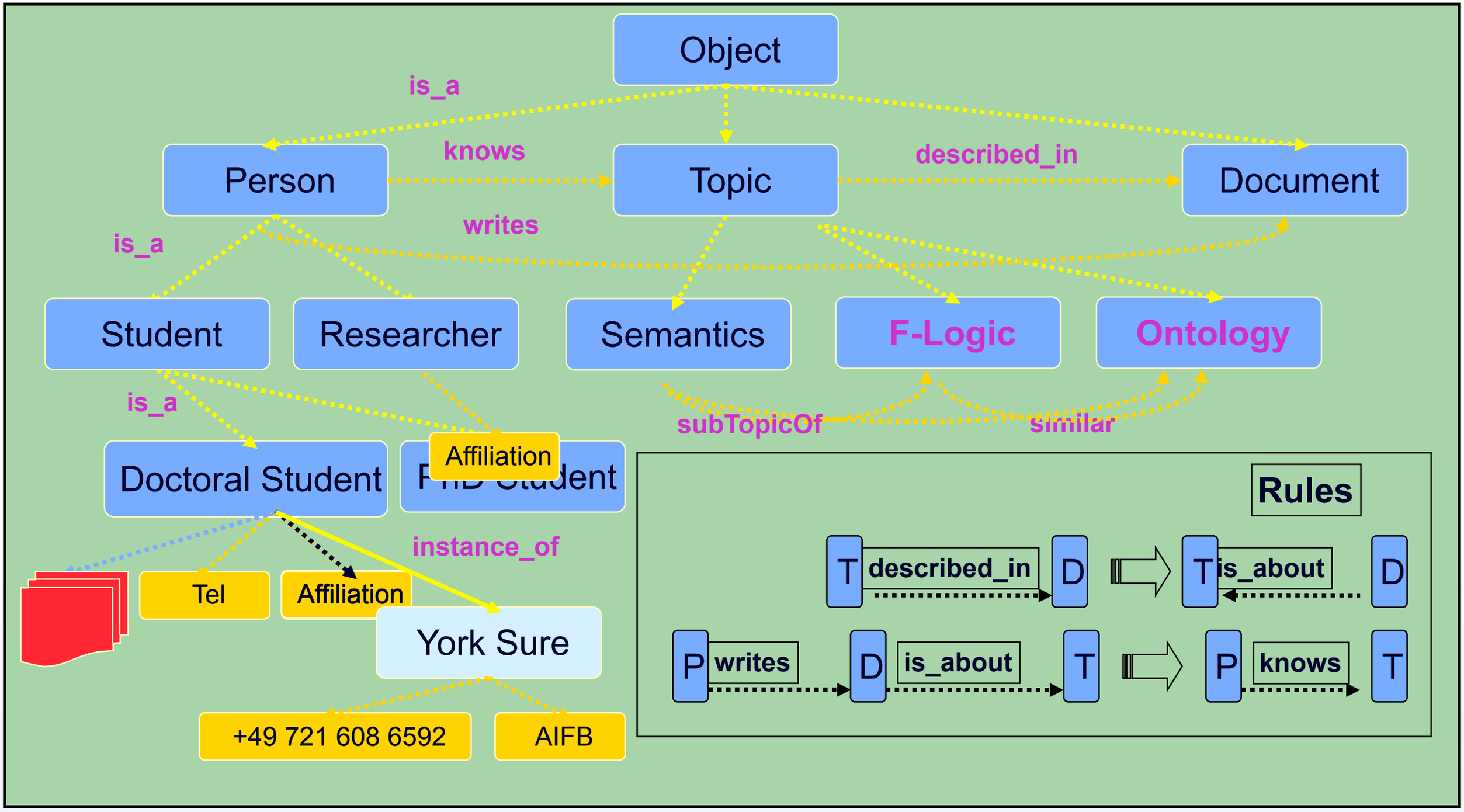
Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

Ontology



- Representation Language: Predicate Logic (F-Logic)
- Standards: RDF(S); OWL

[Hotho, Sure, 2003]

Knowledge Organization Examples

Knowledge Organization Examples

- ad-hoc via diagrams
- concept-form-referent triangle
- ontology mind map
- comparison on knowledge organization methods
  - taxonomy, thesaurus, topic map, ontology
- examples of ontologies

http://keg.cs.tsinghua.edu.cn/persons/tj/Reports/Pswmp-Jie-Tang.ppt

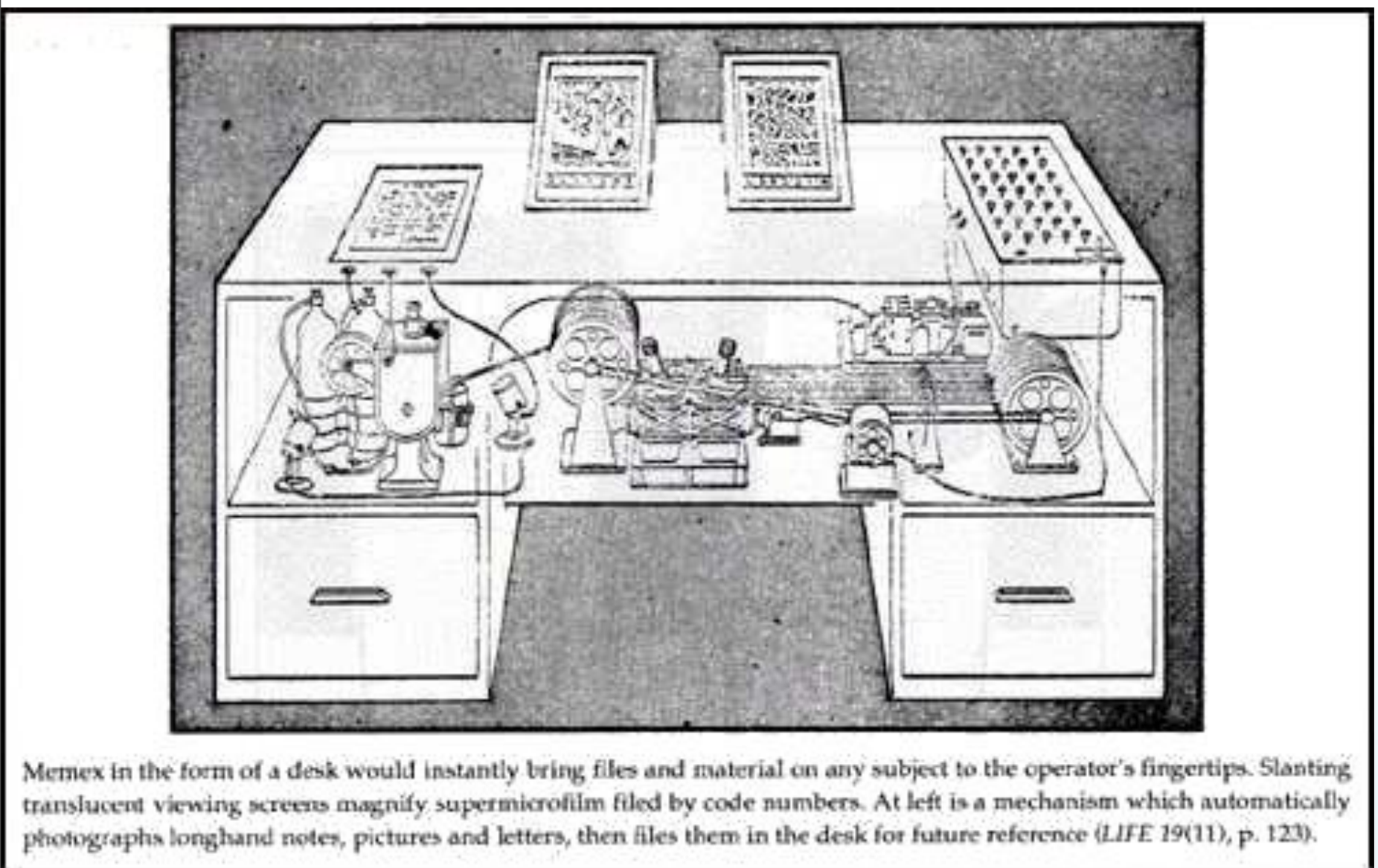
Vannevar Bush: Memex

- **hypothetical information storage device**
  - described in an article in the Atlantic magazine, July 1945
- **sort of mechanized private file and library**
- **enlarged supplement to an individual's memory**
- **memex may stand for "memory extender" or a combination of "memory" and "index"**

http://www.theatlantic.com/magazine/print/1945/07/as-we-may-think/3881/

Memex

Memex

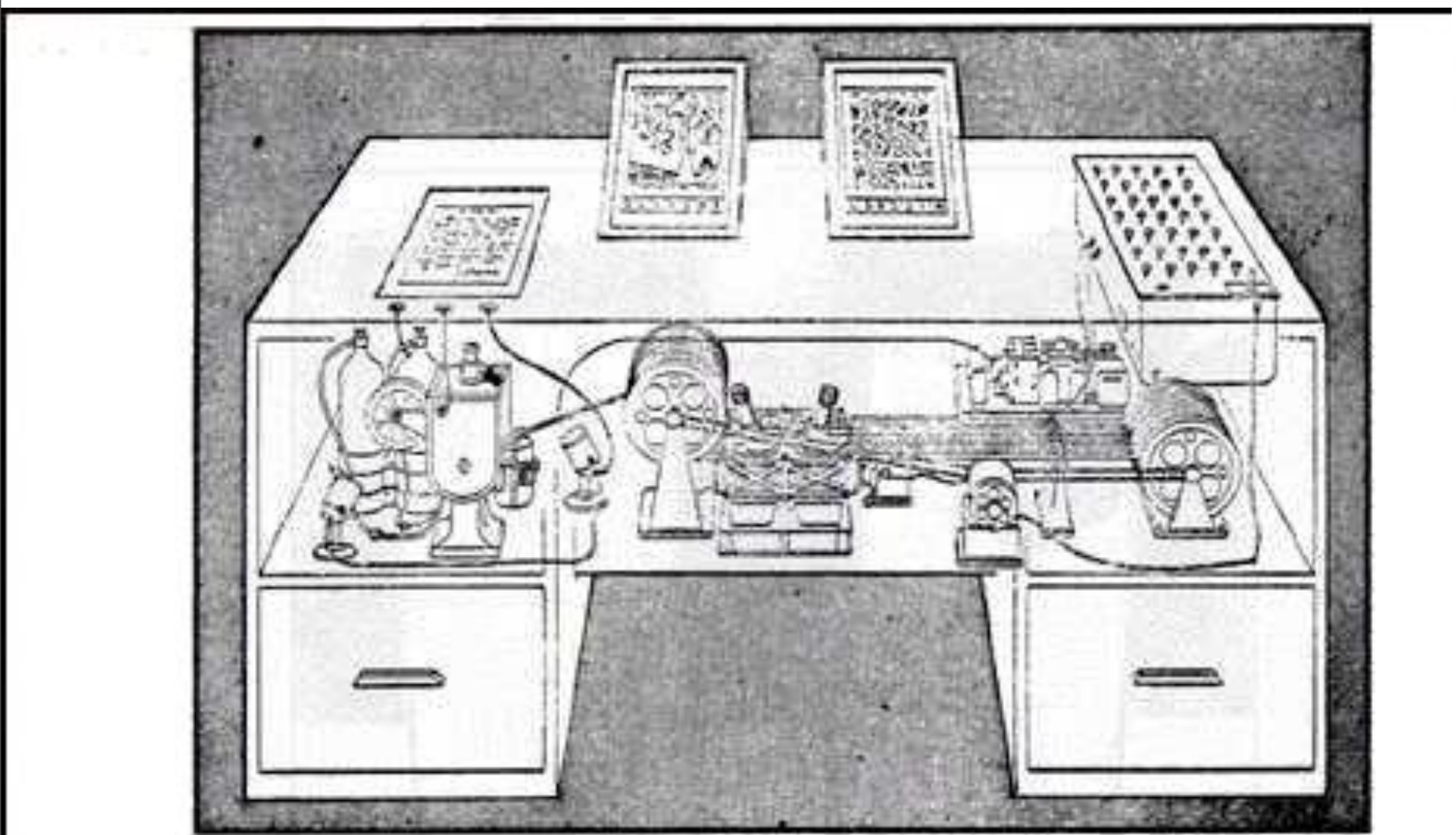


Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Drawing of Bush's theoretical Memex machine (Life Magazine, November 19, 1945)

http://www.kerryr.net/images/pioneers/gallery/memex_lg.jpg

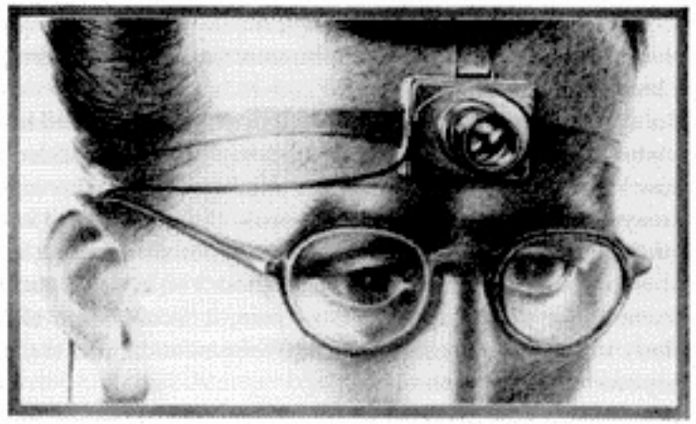
Memex



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Drawing of Bush's theoretical Memex machine (Life Magazine, November 19, 1945)

http://www.kerryr.net/images/pioneers/gallery/memex_lg.jpg

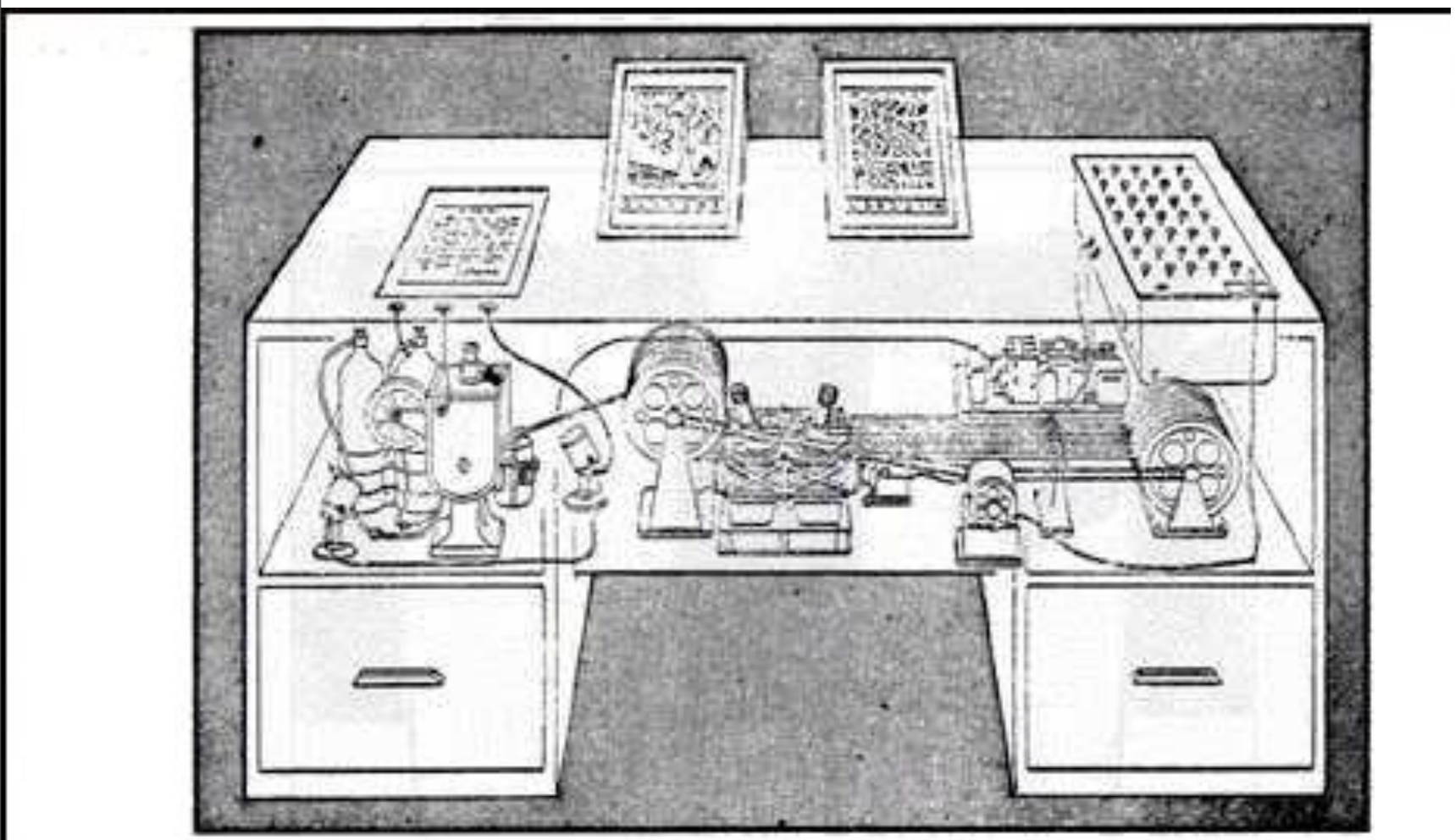


A scientist of the future records experiments with a tiny camera fitted with universal-focus lens. The small square in the eyeglass at the left sights the object (*LIFE* 19(11), p. 112).

MEMEX head camera

http://www.acmi.net.au/AIC/headcam.gif

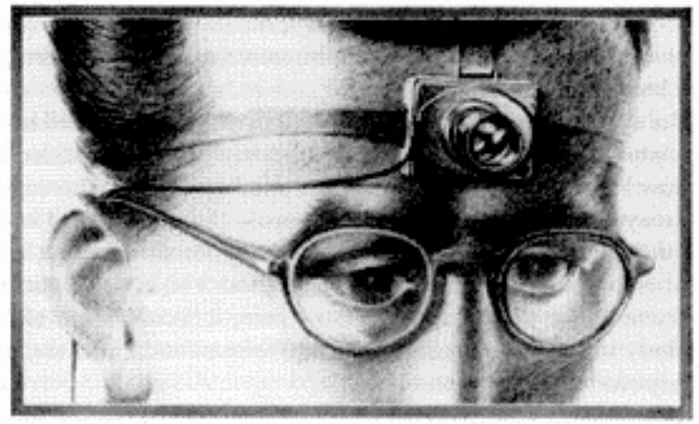
Memex



Memex in the form of a desk would instantly bring files and material on any subject to the operator's fingertips. Slanting translucent viewing screens magnify supermicrofilm filed by code numbers. At left is a mechanism which automatically photographs longhand notes, pictures and letters, then files them in the desk for future reference (*LIFE* 19(11), p. 123).

Drawing of Bush's theoretical Memex machine (Life Magazine, November 19, 1945)

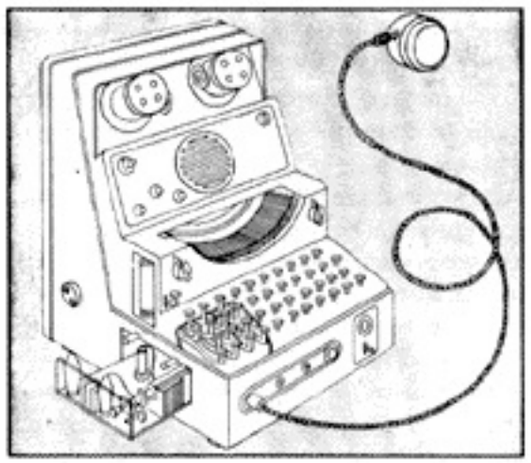
http://www.kerryr.net/images/pioneers/gallery/memex_lg.jpg



A scientist of the future records experiments with a tiny camera fitted with universal-focus lens. The small square in the eyeglass at the left sights the object (*LIFE* 19(11), p. 112).

MEMEX head camera

http://www.acmi.net.au/AIC/headcam.gif



Supersecretary of the coming age, the machine contemplated here would take dictation, type it automatically and even talk back if the author wanted to review what he had just said. It is somewhat similar to the Voder seen at the New York World's Fair. Like all machines suggested by the diagrams in this article, it is not yet in existence (*LIFE* 19(11), p. 114).

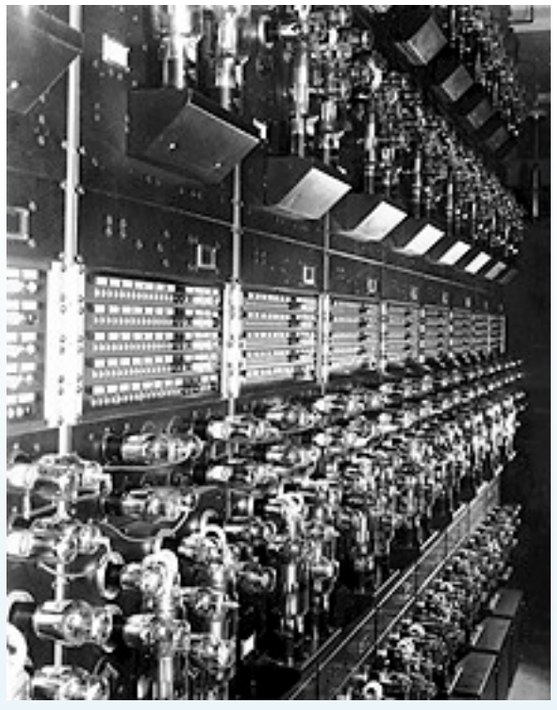
Vannavar Bush's MEMEX voice input output device

http://www.acmi.net.au/AIC/voice.gif

Vannevar Bush

Vannevar Bush seated at a desk. This portrait is credited to "OEM Defense", the Office for Emergency Management (part of the United States Federal Government) during World War II; it was probably taken some time between 1940 and 1944.

source: http://lcweb2.loc.gov/cgi-bin/query/r?pp/PPALL:@field(NUMBER+@1(cph+3a37339)



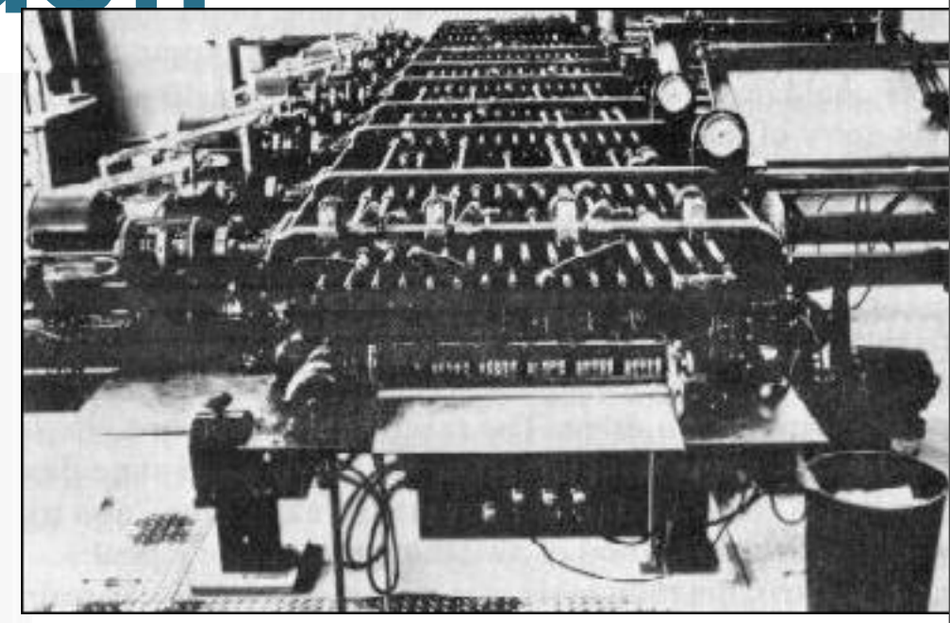
Rockefeller Differential Analyzer
http://www.eecs.mit.edu/AY95-96/events/bush/gif/vb2

http://www.eecs.mit.edu/AY95-96/events/bush/photos.html

Vannevar Bush

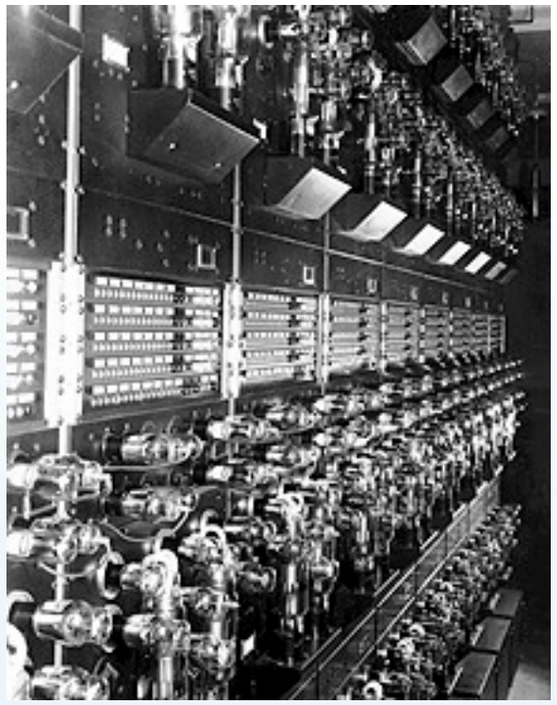
Vannevar Bush seated at a desk. This portrait is credited to "OEM Defense", the Office for Emergency Management (part of the United States Federal Government) during World War II; it was probably taken some time between 1940 and 1944.

source: http://lcweb2.loc.gov/cgi-bin/query/r?pp/PPALL:@field(NUMBER+@1(cph+3a37339)



Closer view of the Differential Analyser

http://www.kerryr.net/images/pioneers/gallery/diff_analyser3_lg.j



Rockefeller Differential Analyzer

http://www.eecs.mit.edu/AY95-96/events/bush/gif/vb2

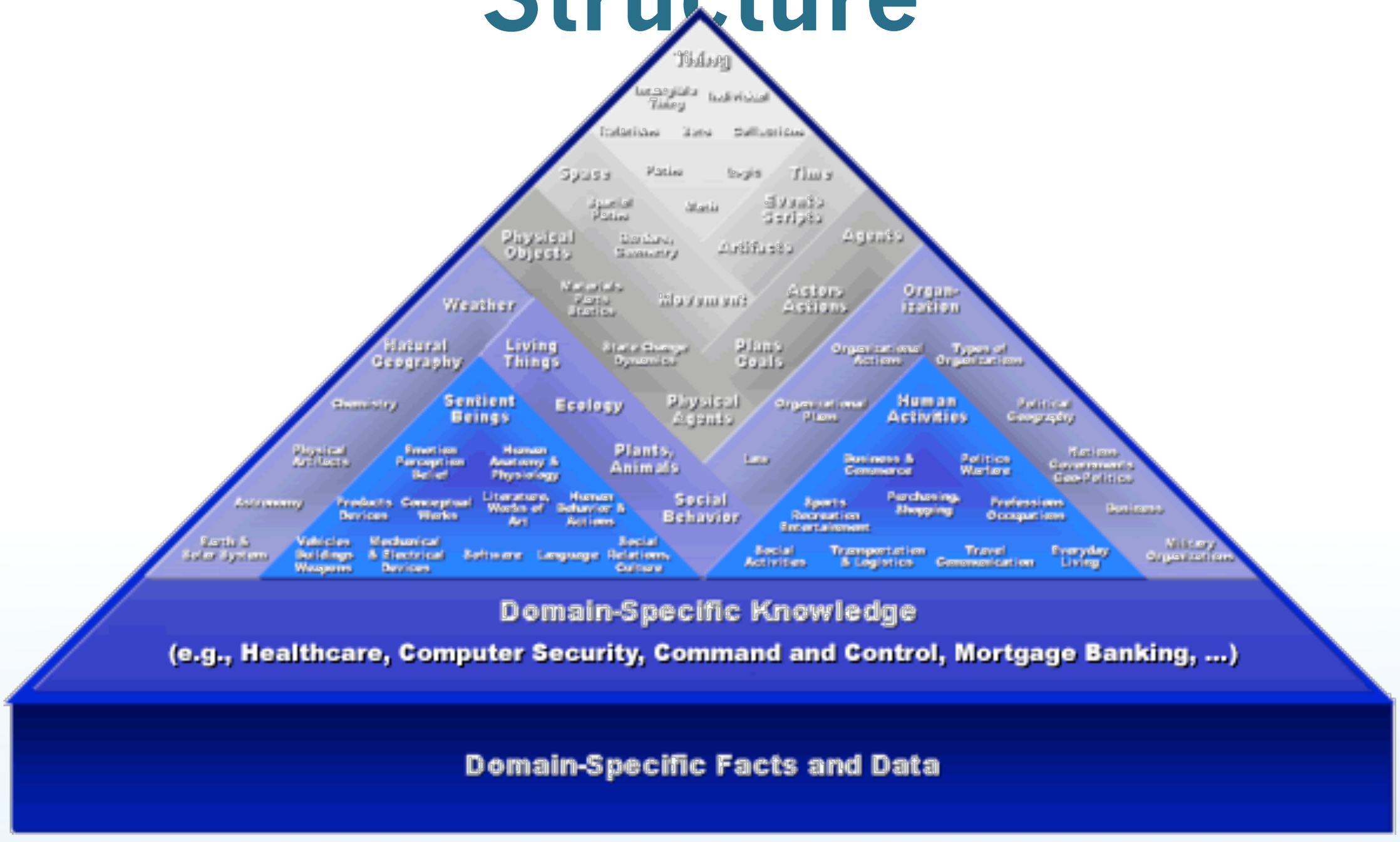
http://www.eecs.mit.edu/AY95-96/events/bush/photos.html

Gordon Bell's CyberAll

- **Personal Digital Store**
  - Microsoft Research MyLifeBits project
    - http://research.microsoft.com/en-us/projects/mylifebits/default.aspx
  - inspired by Vannevar Bush's Memex vision
- **encodes, stores, and allows easy retrieval of a person's information**
  - professional documents
    - books, articles, tech reports, work documents, email, ...
  - personal documents
    - letters, notes, shopping lists, ...

 Bell G (January 2001) A personal

Cyc Knowledge Base Structure

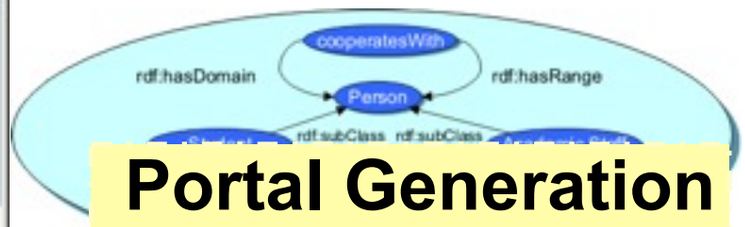


Follow the link below for an interactive version that shows more information about the categories (requires JavaScript, and may not work in all browsers):
http://www.cyc.com/cyc/images/cyc/technology/whatiscyc_dir/whatdoescycknow

OntoWeb.org

[Hotho, Sure, 2003]

OntoWeb.org



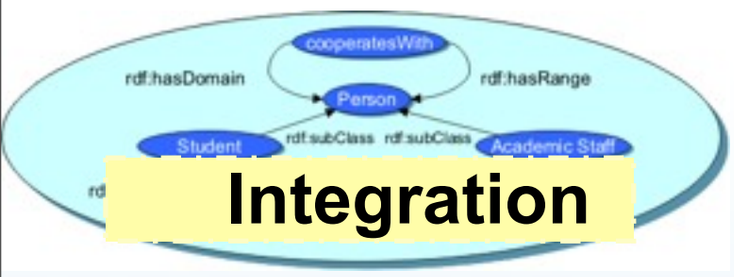
Portal Generation

Navigation

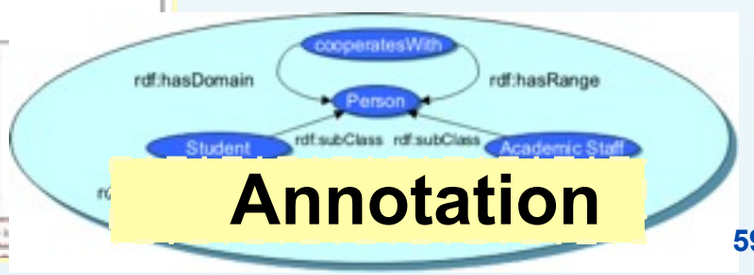
Query/Serach

Content

Collect metadata from participating partners



Integration

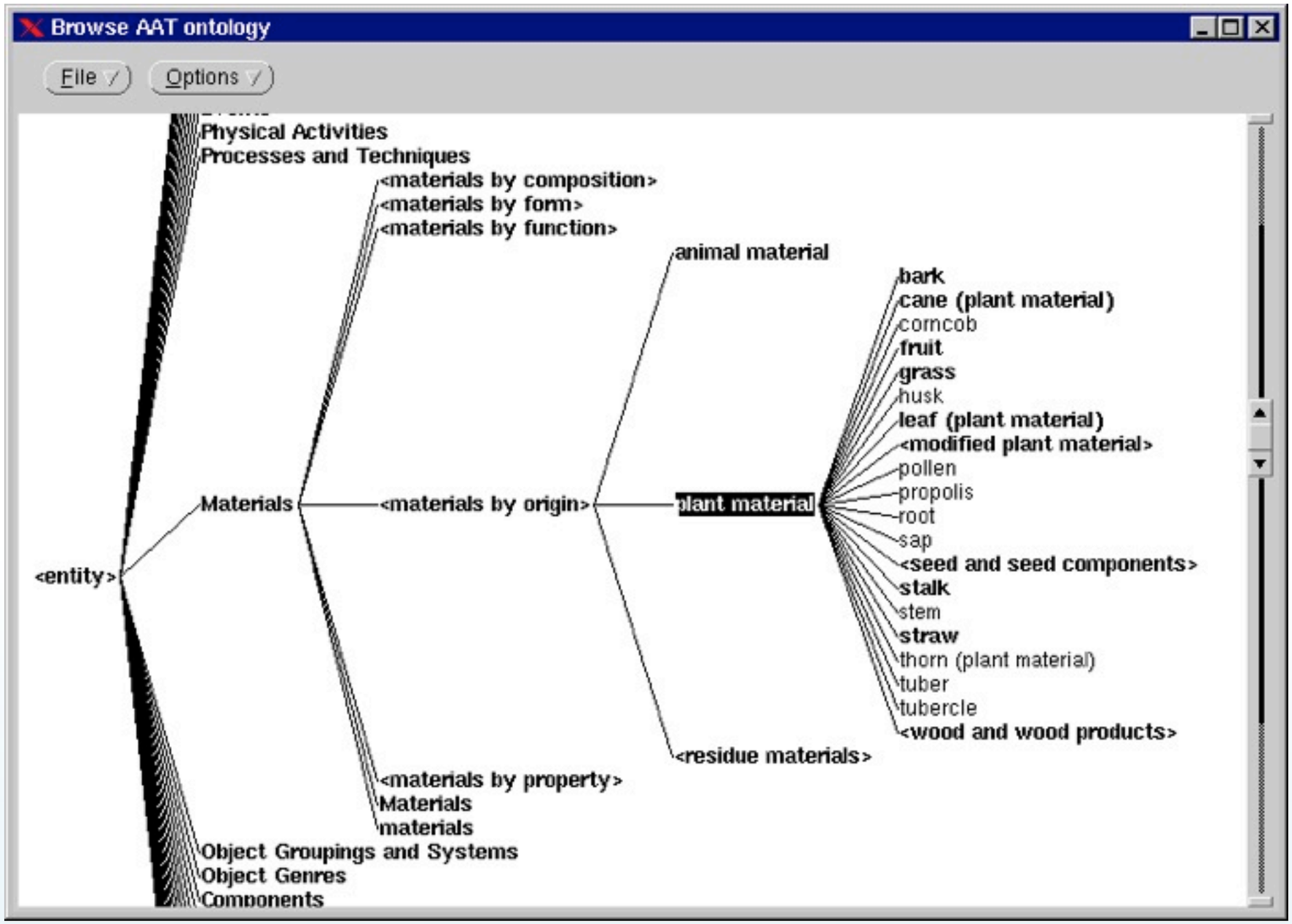


Annotation

[Hotho, Sure, 2003]

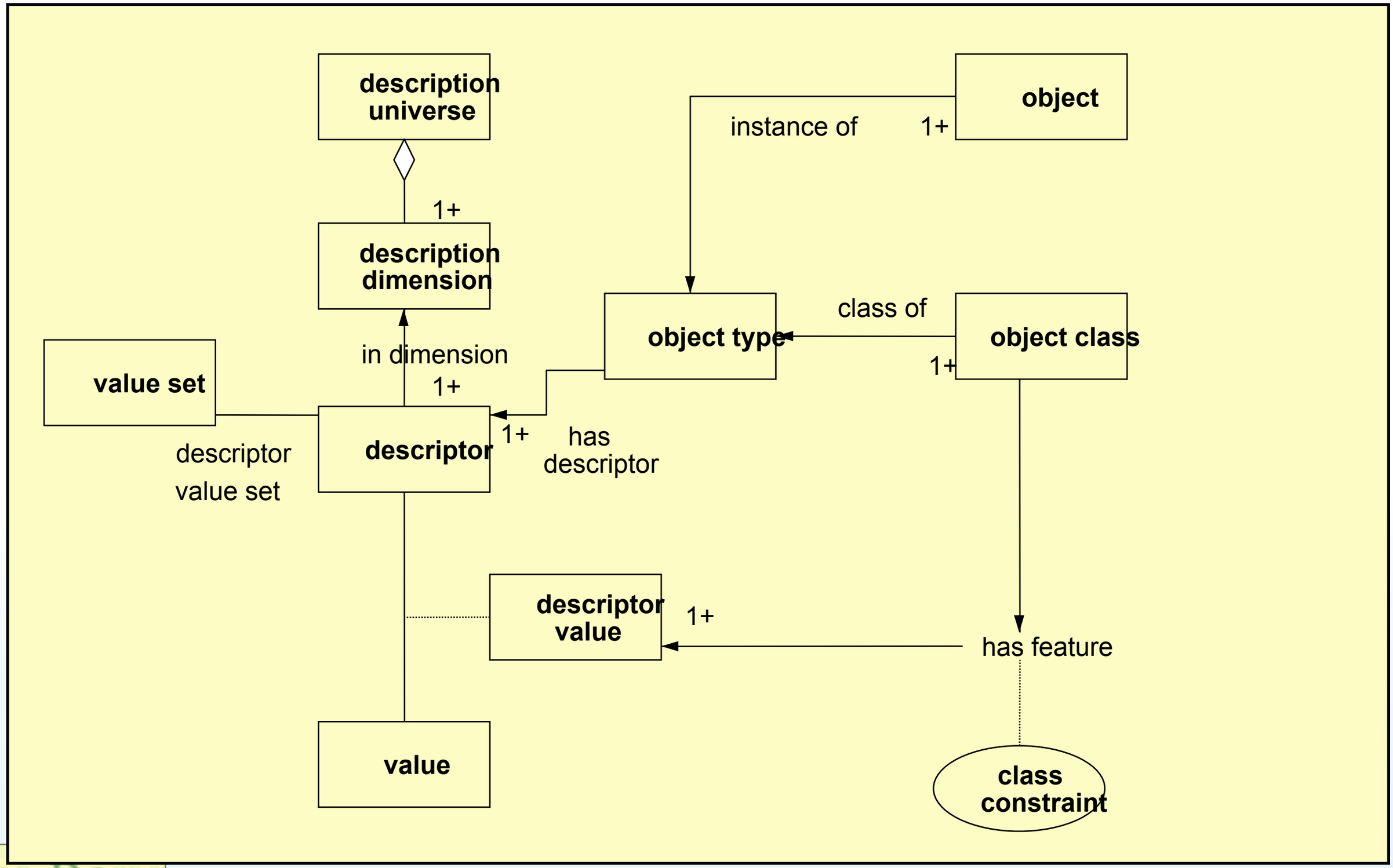
Art & Architecture Thesaurus

used for indexing stolen art objects in European police databases



[Schreiber 2000]

AAT Ontology



[Schreiber 2000]

ArnetMiner.org— Academic Researcher Social Network

Jie Tang, Jing Zhang, Limin Yao, Duo Zhang, and Mingcai Hong
Knowledge Engineering Group, DCST, Tsinghua University
{tangjie, zhangjing, ylm, zhangduo, hmc}@keg.cs.tsinghua.edu.cn

Arnetminer
Http://www.arnetminer.org

Association Finding

Expert Finding

Person Search

Survey Paper Finding

Publication Search

Social Network Mining

- Expert Finding
- Association Finding
- Survey Paper Finding
- Sub-topic Finding
- Hot-topic Finding

Social Network Search

- Person Search
- Publication Search
- Conference Search

Arnetminer

Sub-topic Finding

Conference Search

Hot-topic Finding

Social Network Storage

RNKB

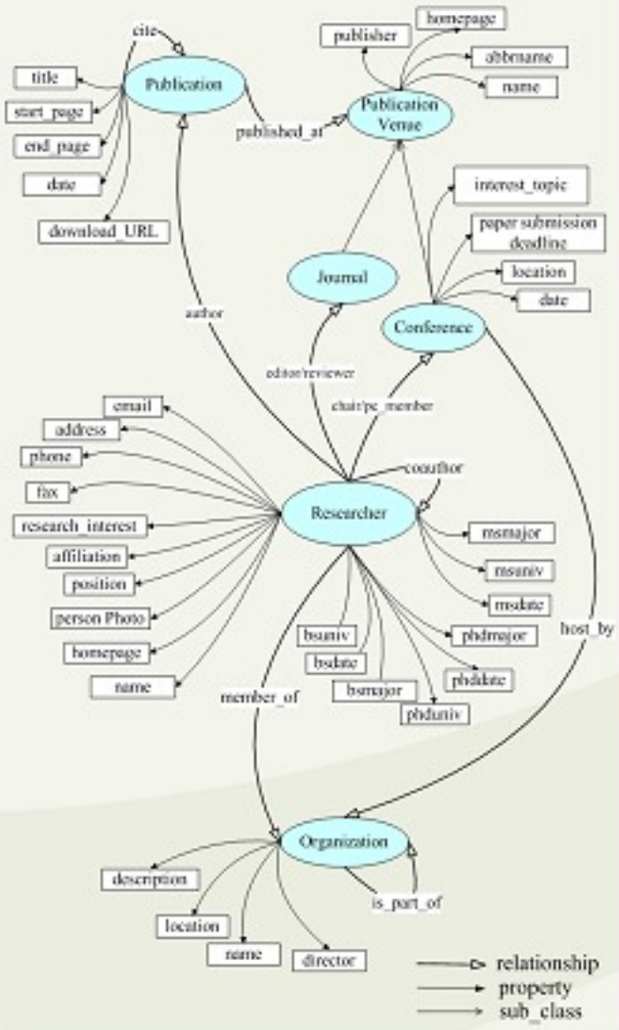
- Access
- Indexing
- Storage

Social Network Extraction

- Name Disambiguation
- Integration
- Profiling

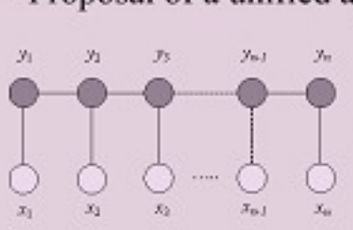
Technique Issues

Metadata



ArnetMiner advances four points:

- Proposal of a unified approach to researcher profiling based on conditional random fields.



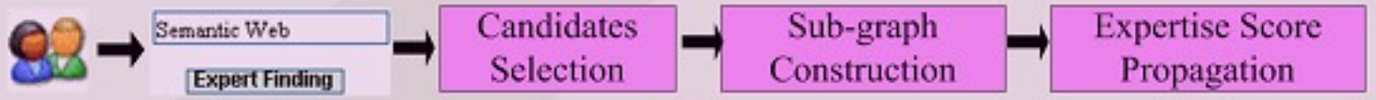
$$p(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{e\in\{E^m,E^o,E^m\}_j}\lambda_j f_j(e,y|_e,x)+\sum_{v\in V_k}\mu_k s_k(v,y|_v,x)\right)$$

- Proposal of a constraint-based probabilistic model to name disambiguation.

$$p(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}D(x_i,y_k)+\sum_{i,j\neq i}\{D(x_i,x_j)\sum_{c_k\in C}[w_k c_k(x_i,x_j)]\}\right)$$

C_r	W_r	Constraint Name	Description
c_1	w_1	CoOrg	$a_i^{[3]}.affiliation = a_j^{[3]}.affiliation$
c_2	w_2	CoAuthor	$\exists r, s>0,\ a_i^{[s]}=a_j^{[s]}$
c_3	w_3	Citation	p_i cites p_j or p_j cites p_i
c_4	w_4	CoEmail	$a_i^{[3]}.email = a_j^{[3]}.email$
c_5	w_5	Feedback	Constraints from user feedback
c_6	w_6	t-CoAuthor	one common author in t extension

- Proposal of a score-and-propagate approach to expert finding

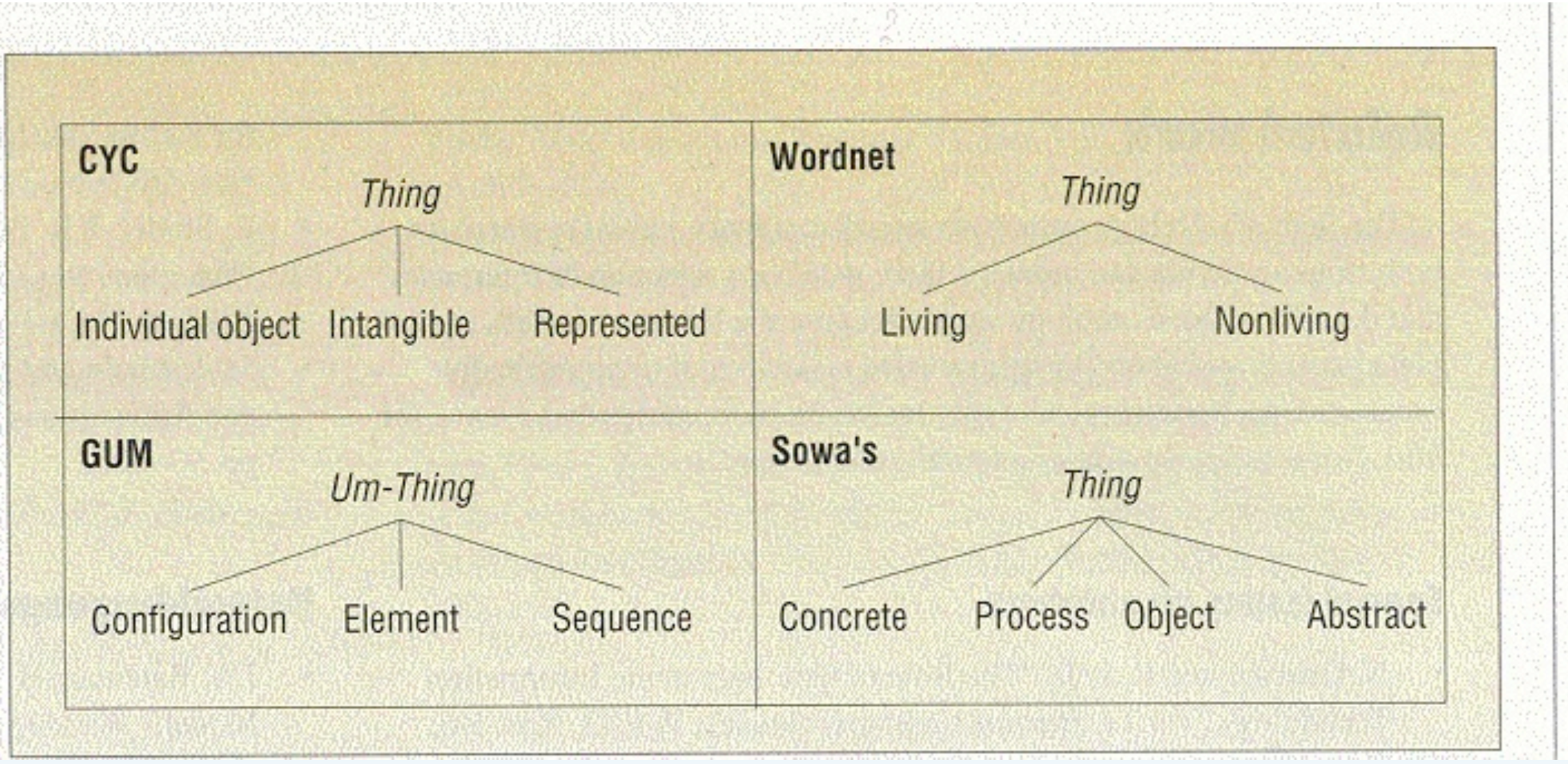


- Proposal of an efficient approach to association search.

* Other features are developed based on NLP and Text Mining, for example: Key-Phrase Extraction (e.g., research interest finding), Classification based ranking (e.g., survey paper finding), Hierarchical clustering (e.g., sub-topic finding), etc.

KEG, TSINGHUA, CHINA

Top-level Categories: Many Different Proposals



Chandrasekaran et al. (1999)

[Schreiber 2000]

Rama Hoetzlein - Quanta System

- **Quanta - The Organization of Human Knowledge: Systems for Interdisciplinary Research**

  Rama Hoetzlein; Master's Thesis, University of California Santa Barbara, June 2007

  - http://www.rchoetzlein.com/quanta/

Linked Data

- **entities identified by URIs**
- **people and agents can refer to these entities**
  - typically via http
- **information about entities**
  - structured according to standards such as RDF/XML
- **links to other, related entities**

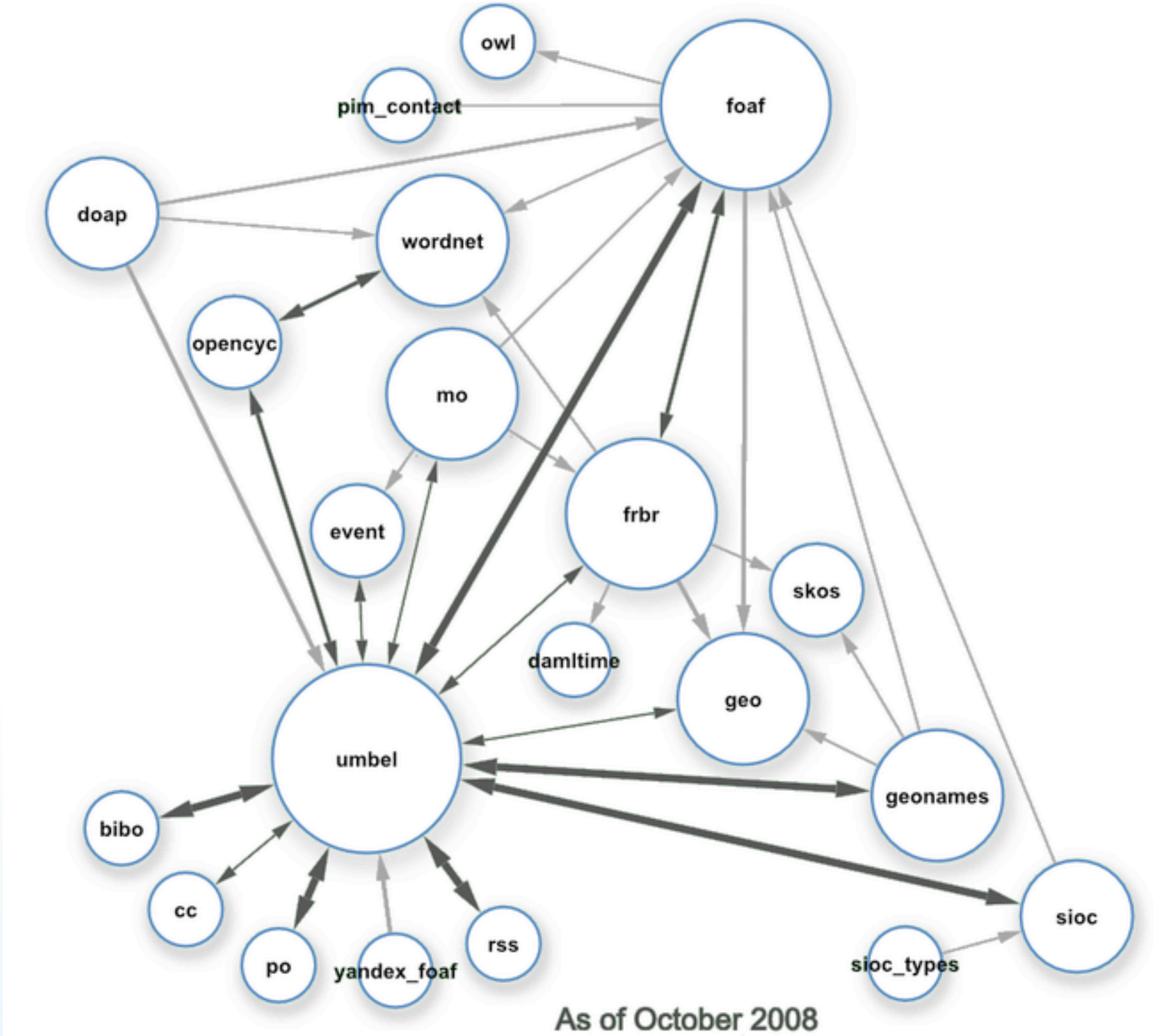
Tim Berners-Lee on the next Web. Talk at the TED 2009 conference, http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html or http://video.ted.com/talks/podcast/TimBerners-Lee_2009_480.mp4

Tom Heath and Christian Bizer (2011) *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. http://linkeddatabook.com/book

DOI: 10.2200/S00334ED1V01Y201102WBE001
ISBN: 9781608454303 (paperback)
ISBN: 9781608454310 (ebook)

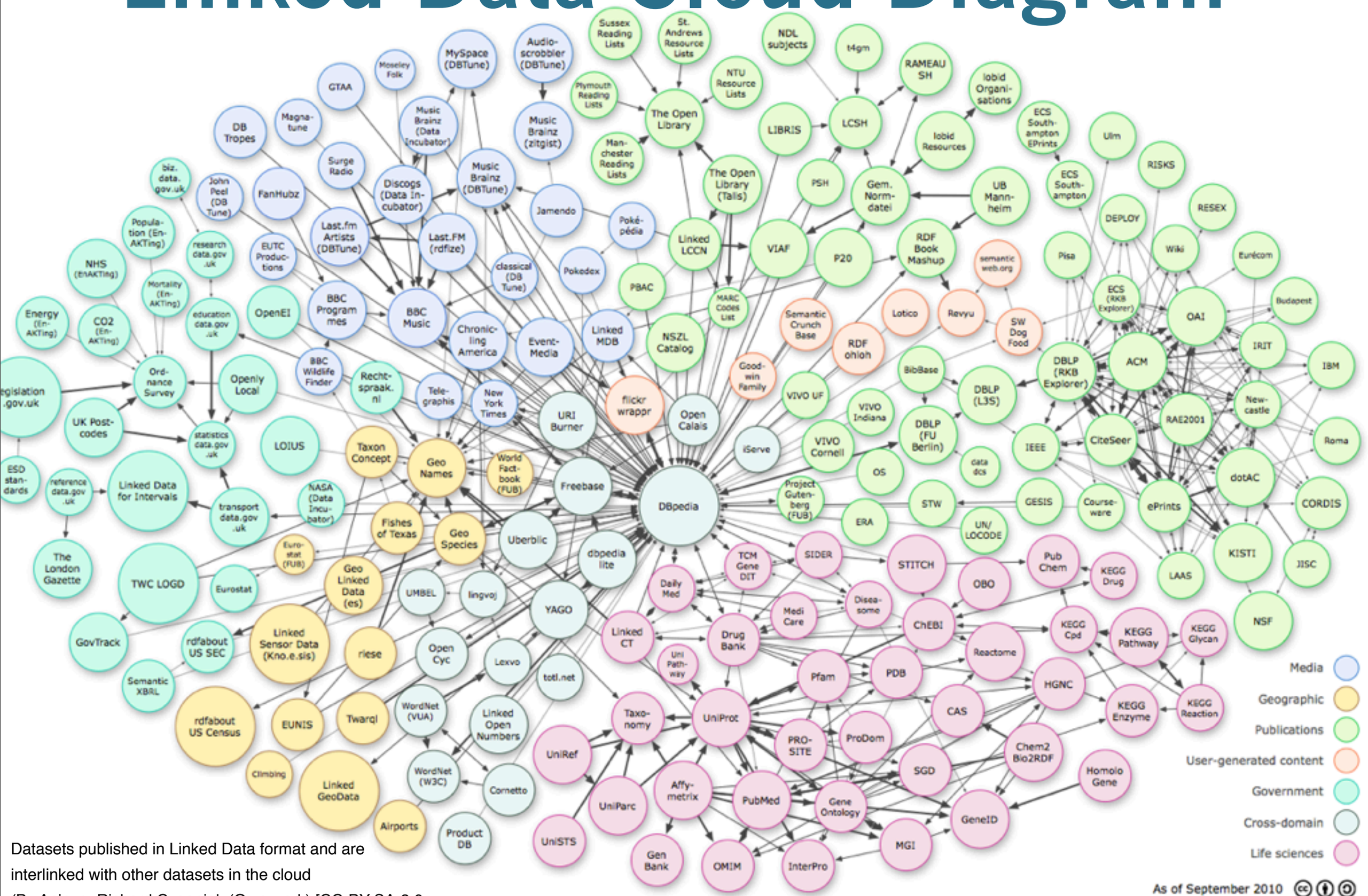
LOD Classes

- **Linking Open Data project**
  - open data sets on the Web
  - RDF triples
  - RDF links



Class diagram for the LOD datasets (http://umbel.org/lod_constellation.html)

Linked Data Cloud Diagram



Datasets published in Linked Data format and are interlinked with other datasets in the cloud
(By Anjeve, Richard Cyganiak (Own work) [CC-BY-SA-3.0 (www.creativecommons.org/licenses/by-sa/3.0) or GFDL (www.gnu.org/copyleft/fdl.html)], via Wikimedia Commons)

http://commons.wikimedia.org/wiki/File:Lod-datasets_2010-09-22_colored.

Linked Open Data Visualization

- **Web app allowing interactive exploration of the LOD data set**

http://www.webknox.com/blog/2010/05/linked-open-data-on-the-web-visualization/

DBpedia

- **knowledge base derived from Wikipedia**
  - wiki.dbpedia.org
  - conversion of Wikipedia contents into structured data organized around an ontology
- **nucleus for the W3C Linking Open Data (LOD) effort**
  - W3C Linking Open Data (LOD) community effort

Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, Sebastian Hellmann: DBpedia – A Crystallization Point for the Web of Data. Journal of Web Semantics: Science, Services and Agents on the World Wide Web, Issue 7, Pages 154–165, 2009.

DBpedia Contents

- **DBpedia 3.8 release, based on Wikipedia dumps dating from May/June 2012**
  - wiki.dbpedia.org : About

1. the new release is based on updated Wikipedia dumps dating from late May / early June 2012.
2. the DBpedia ontology is enlarged and the number of infobox to ontology mappings has risen.
3. the DBpedia internationalization has progressed and we now provide localized versions of DBpedia in even more languages.

The English version of the DBpedia knowledge base currently describes 3.77 million things, out of which 2.35 million are classified in a consistent Ontology, including 764,000 persons, 573,000 places (including 387,000 populated places), 333,000 creative works (including 112,000 music albums, 72,000 films and 18,000 video games), 192,000 organizations (including 45,000 companies and 42,000 educational institutions), 202,000 species and 5,500 diseases.

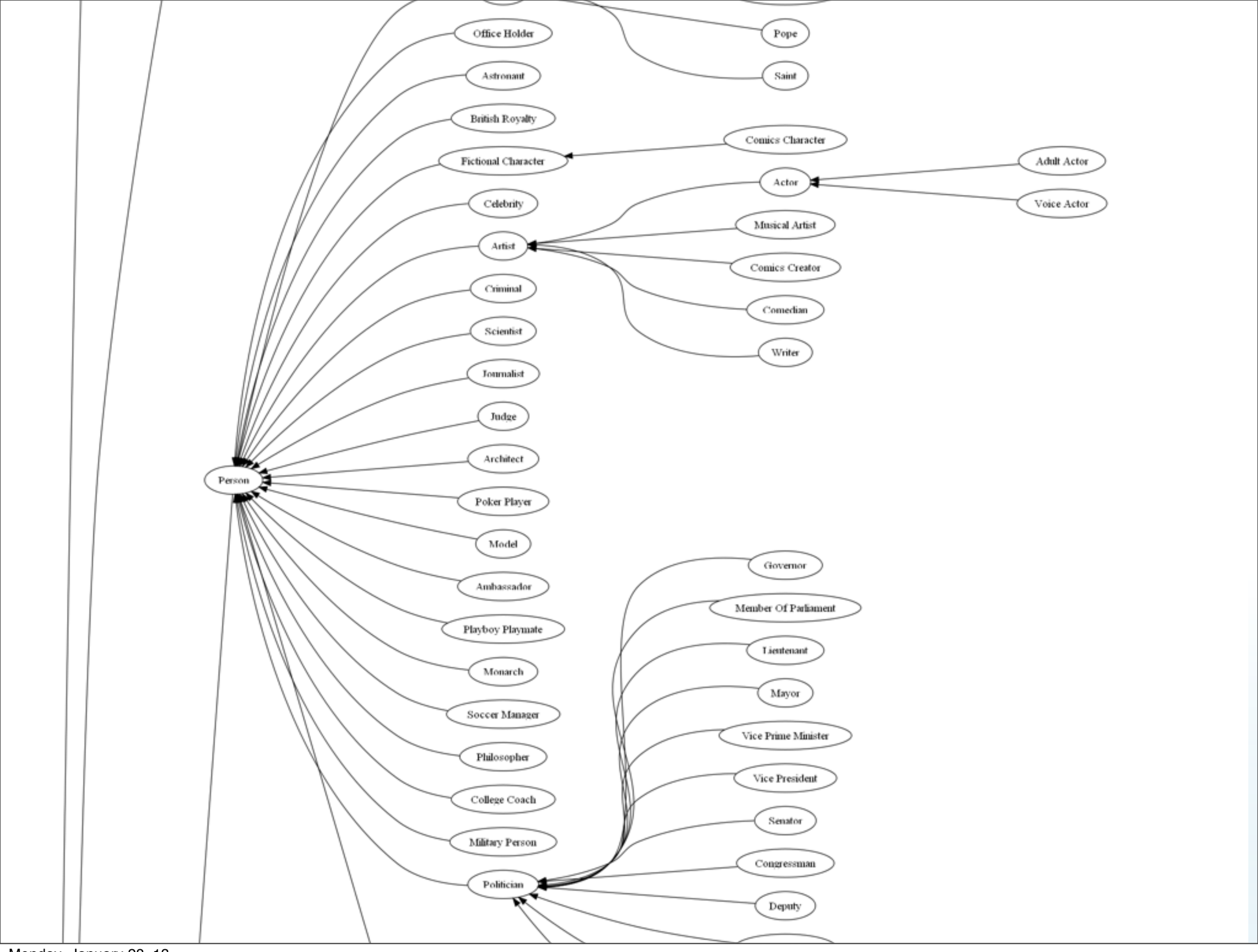
DBpedia Contents

- **DBpedia 3.6 release, based on Wikipedia dumps dating from October/November 2010**
  - wiki.dbpedia.org : About

The DBpedia knowledge base currently describes more than 3.5 million things, out of which 1.67 million are classified in a consistent Ontology, including 364,000 persons, 462,000 places, 99,000 music albums, 54,000 films, 17,000 video games, 148,000 organisations, 169,000 species and 5,200 diseases. The DBpedia data set features labels and abstracts for these 3.5 million things in up to 97 different languages; 1,850,000 links to images and 5,900,000 links to external web pages; 6,500,000 external links into other RDF datasets, 633,000 Wikipedia categories, and 2,900,000 YAGO categories. The DBpedia knowledge base altogether consists of over 672 million pieces of information (RDF triples) out of which 286 million were extracted from the English edition of Wikipedia and 386 million were extracted from other language editions.

DBpedia Ontology

- **manually derived from Wikipedia**
  - based on the most commonly used infoboxes
  - combined with an infobox extraction method
- **shallow**
  - 272 classes arranged in a subsumption hierarchy
    - whittled down from 1124 Wikipedia templates
  - 1300 properties
    - reduced from 3690 Wikipedia template properties
- **cross-domain**
- **multiple access methods**
  - browsers, SPARQL end points


Office Holder
Astronaut
British Royalty
Fictional Character
Celebrity
Artist
Criminal
Scientist
Journalist
Judge
Architect
Person
Poker Player
Model
Ambassador
Playboy Playmate
Monarch
Soccer Manager
Philosopher
College Coach
Military Person
Politician
Pope
Saint
Comics Character
Actor
Musical Artist
Comics Creator
Comedian
Writer
Adult Actor
Voice Actor
Governor
Member Of Parliament
Lieutenant
Mayor
Vice Prime Minister
Vice President
Senator
Congressman
Deputy

SPARQL Explorer for http://dbpedia.org/sparql

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
```

```
SELECT ?property ?hasValue ?isValueOf
WHERE {
 { <http://dbpedia.org/resource/University_of_Ulm> ?property ?hasValue }
 UNION
 { ?isValueOf ?property <http://dbpedia.org/resource/University_of_Ulm> }
}
```

DBPedia Sample Query: “University of Ulm”

Results: Browse Go! Reset

Description of http://dbpedia.org/resource/University_of_Ulm:

property	hasValue	
owl:sameAs	-	<h
dbpedia:ontology/wikiPageRedirects	-	:U
dbpedia:ontology/wikiPageRedirects	-	:U
dbpedia:ontology/wikiPageRedirects	-	:U
dbpedia:ontology/wikiPageRedirects	-	:U
dbpedia:ontology/almaMater	-	:D
dbpedia:ontology/wikiPageDisambiguates	-	:U
foaf:primaryTopic	-	<h
rdf:type	owl:Thing	-
rdf:type	dbpedia:ontology/EducationalInstitution	-
rdf:type	dbpedia:class/yago/UniversitiesInGermany	-
rdf:type	dbpedia:ontology/Organisation	-
rdf:type	dbpedia:ontology/University	-
rdf:type	dbpedia:class/yago/EducationalInstitutionsEstablishedIn1967	-
rdf:type	<http://www.opengis.net/gml/_Feature>	-
owl:sameAs	<http://rdf.freebase.com/ns/m/0dcwl3>	-
rdfs:label	"ウルム大学"@ja	-
rdfs:label	"Universidade de Ulm"@pt	-
rdfs:label	"University of Ulm"@en	-
rdfs:label	"乌尔姆大学"@zh	-
	"Die Universität Ulm wurde 1967 als „Medizinisch-Naturwissenschaftliche Hochschule Ulm“ gegründet und ist somit die jüngste Universität in Baden-Württemberg. Die Universität Ulm hat zur Zeit (Wintersemester 2009/10) über 7.600 Studierende. Im Zuge der Internationalität bietet sie für all ihre Studierenden eine professionelle Sprachausbildung mit vielen	-

DBPedia Sample Query: “Eiffel Tower Vicinity”

```
PREFIX geo: <http://www.w3.org/2003/01/geo/wgs84_pos#>
SELECT ?subject ?label ?lat ?long WHERE {
<http://dbpedia.org/resource/Eiffel_Tower> geo:lat ?eiffelLat.
<http://dbpedia.org/resource/Eiffel_Tower> geo:long ?eiffelLong.
?subject geo:lat ?lat.
?subject geo:long ?long.
?subject rdfs:label ?label.
FILTER(?lat - ?eiffelLat <= 0.05 && ?eiffelLat - ?lat <= 0.05 &&
?long - ?eiffelLong <= 0.05 && ?eiffelLong - ?long <= 0.05 &&
lang(?label) = "en"
).
} LIMIT 20
```

Results: Browse Go! Reset

SPARQL results:

subject	label	lat	long
:Tour_Europlaza	"Tour Europlaza"@en	48.89166641235352	2.244999885559082
:Tour_Michelet	"Tour Michelet"@en	48.88833236694336	2.245138883590698
:Stade_Roland_Garros	"Stade Roland Garros"@en	48.84722137451172	2.246388912200928
:Tour_CBX	"Tour CBX"@en	48.89110946655273	2.246666669845581
:Tour_Descartes	"Tour Descartes"@en	48.8922233581543	2.246666669845581
:Ch%C3%A2teau_de_Bagatelle	"Château de Bagatelle"@en	48.87166595458984	2.247222185134888
:Tour_Aurore	"Tour Aurore"@en	48.88999938964844	2.247361183166504
:Tour_France	"Tour France"@en	48.883056640625	2.247638940811157
:Gare_de_Courbevoie	"Gare de Courbevoie"@en	48.89833450317383	2.248611211776733
:Tour_Generali	"Tour Generali"@en	48.88944625854492	2.2491667270660 4
:French_Open	"French Open"@en	48.84716415405273	2.249216556549072
:Tour_Gan	"Tour Gan"@en	48.88888931274414	2.249805450439453
:Tenniseum	"Tenniseum"@en	48.84722137451172	2.250277757644653
:Bois_de_Boulogne	"Bois de Boulogne"@en	48.86472320556641	2.25083327293396

Important Concepts and Terms

- category
- cognitive science
- computer science
- concept map
- dictionary
- glossary
- hierarchy
- index
- knowledge representation
- linguistics
- logic
- metadata
- natural language
- ontology
- ontological commitment
- Resource Description Format (RDF)
- surrogate
- taxonomy
- term
- thesaurus
- topic map
- Uniform Resource Identifier (URI)

Summary