# 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*


# Use and Distribution of these Slides

These slides are primarily intended for the students in classes I teach. In some cases, I only make PDF versions publicly available. If you would like to get a copy of the originals (Apple KeyNote or Microsoft PowerPoint), please contact me via email at fkurfess@calpoly.edu. I hereby grant permission to use them in educational settings. If you do so, it would be nice to send me an email about it. If you're considering using them in a commercial environment, please contact me first.

# 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

# 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 Devices

- 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

# 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

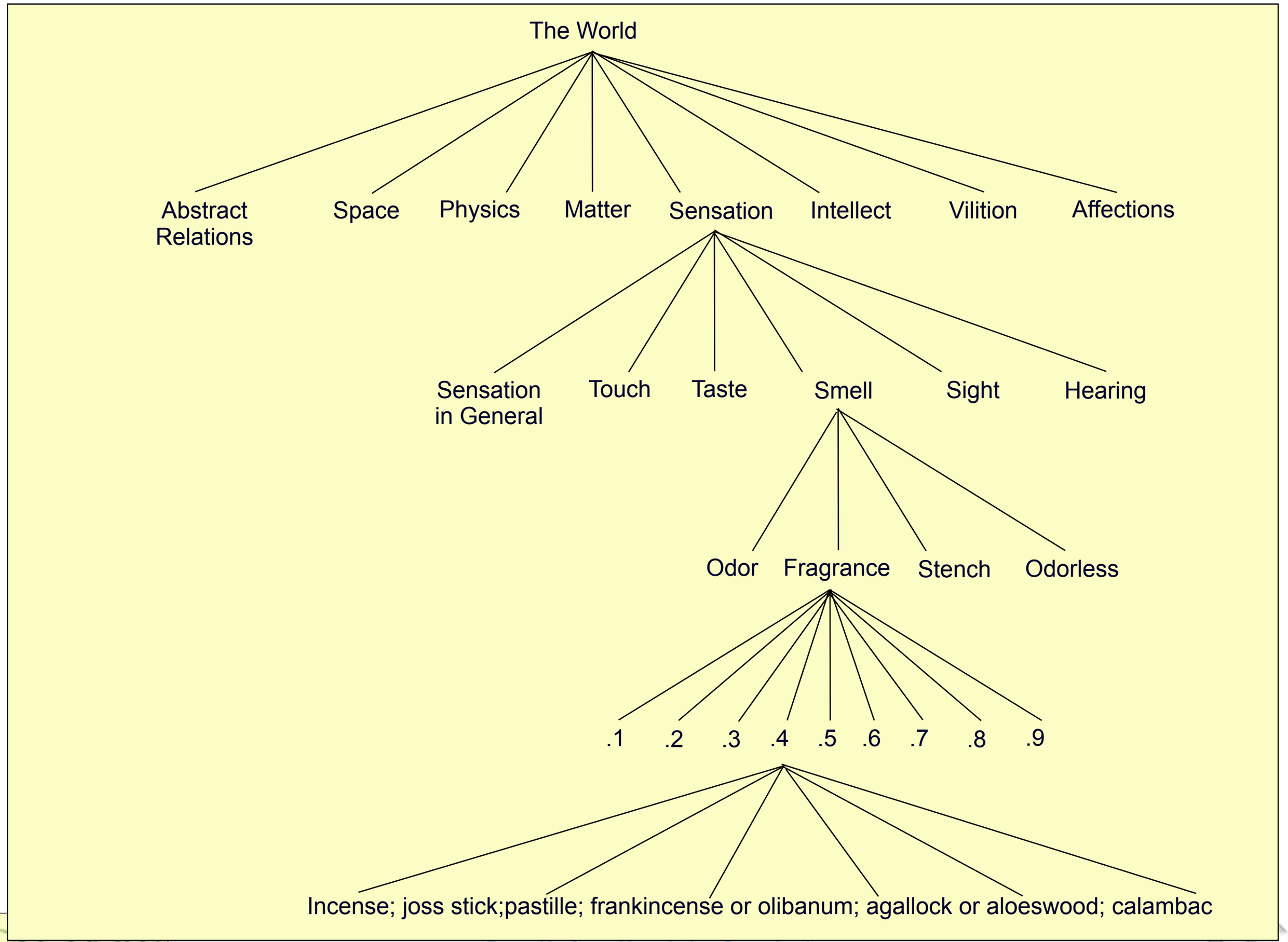
- manually constructed for a specific domain
- intended for human indexers and searchers
- contains
  - synonyms (“use for” UF)
  - more general (“broader term” BT)
  - more specific (“narrower” NT)
  - otherwise associated words (“related term” RT)
- example: “data base management systems”
  - UF data 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]

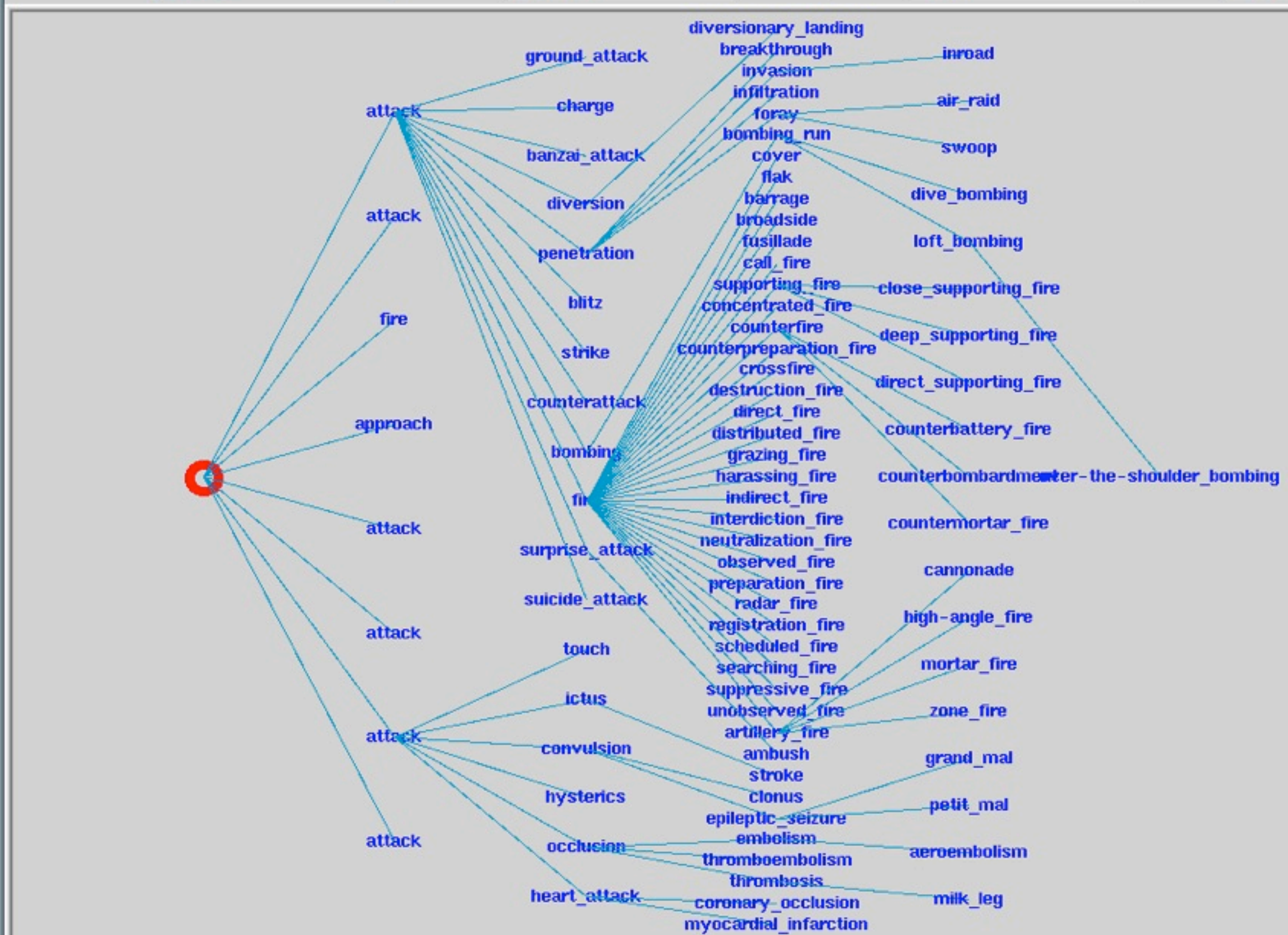

The World
Abstract Relations
Space
Physics
Matter
Sensation
Intellect
Vilition
Affections
Sensation in General
Touch
Taste
Smell
Sight
Hearing
Odor
Fragrance
Stench
Odorless
.1
.2
.3
.4
.5
.6
.7
.8
.9
Incense; joss stick;pastille; frankincense or olibanum; agallock or aloeswood; calambac

[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*
- *statutes*
- *applicability*
- *seeking amnesty*
- *legal status*
- *immigration act*
- *undocumented workers*
- *guest worker*
- *sweeping immigration law*

[Liddy 2000]

# Index

- listing of words that appear in a (set of) documents, together 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

# Natural Language Methods

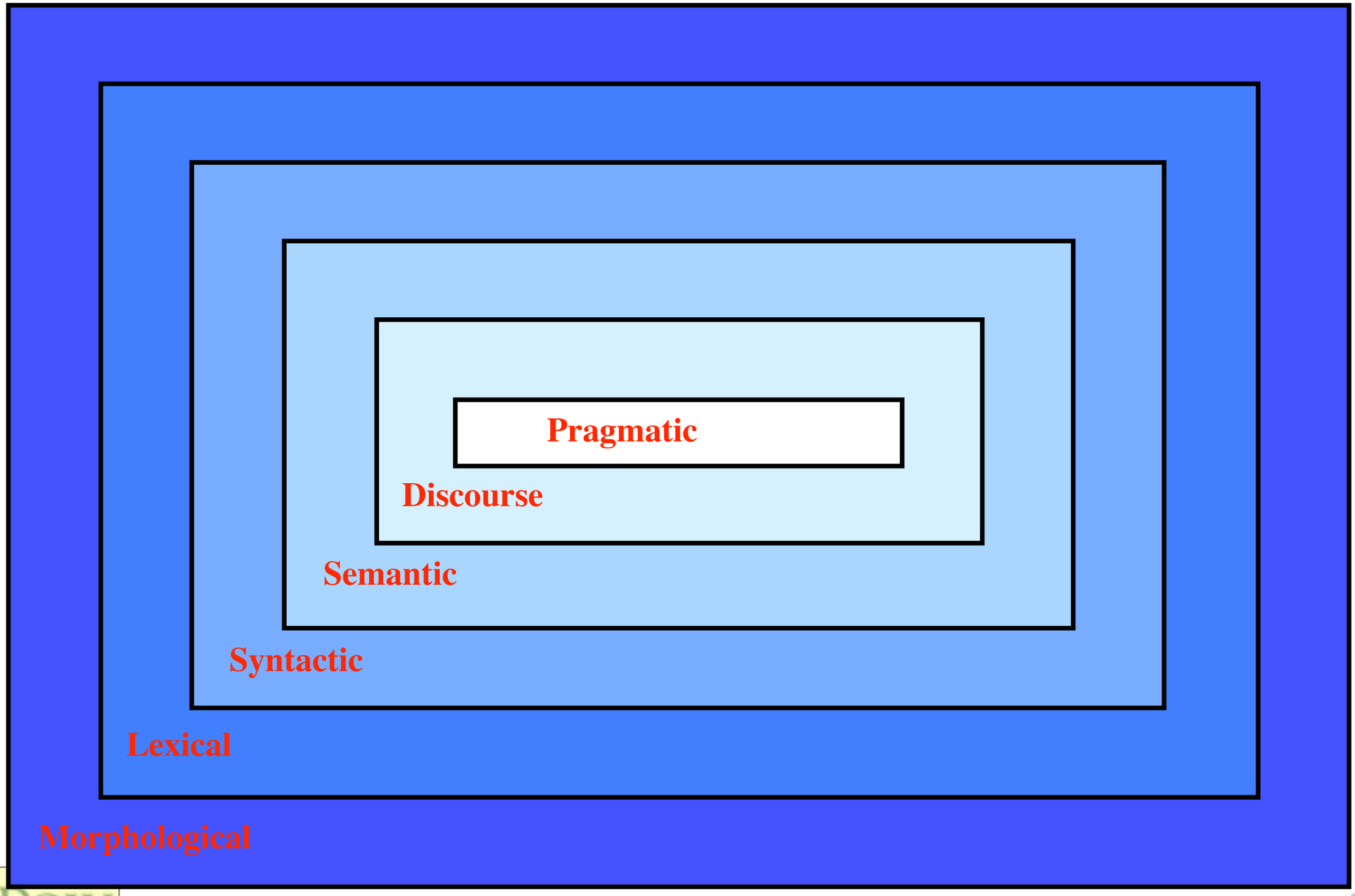
- Natural Language Processing
- Natural Language Understanding
- NLP-based Indexing

# 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]

# Levels of Language Understanding




[Liddy 2000]

# NLP-based Indexing

- the computational process of identifying, selecting, and extracting useful information from massive volumes of textual data
  - for potential review by indexers
  - stand-alone representation of content
  - using Natural Language Processing

[Liddy 2000]

# What can NLP Indexing do?

- phrase recognition
- disambiguation
- concept expansion

# Ontologies

- description
- “representational promiscuity
- ontology types
- usage of ontologies
  - domain standards and vocabularies
- ontology development
  - development proces
  - specification languages

# Categorization

- Feature-based Categorization
- Hierarchical Categorization

# Hierarchical Categorization

- a set of objects is divided into smaller and smaller subset, forming a hierarchical structure (tree) with the elementary objects as leaf nodes
  - typically one feature is used to distinguish one category from another
  - often constitutes a relatively stable “backbone” of a knowledge organization scheme
  - re-organization requires a major effort

# Feature-based Categorization

- objects or documents are assigned to categories according to commonalties in specific features
- can be used to dynamically group objects into categories that are of interest for a particular task or purpose
- re-organization is easy with computer support

# Ontology

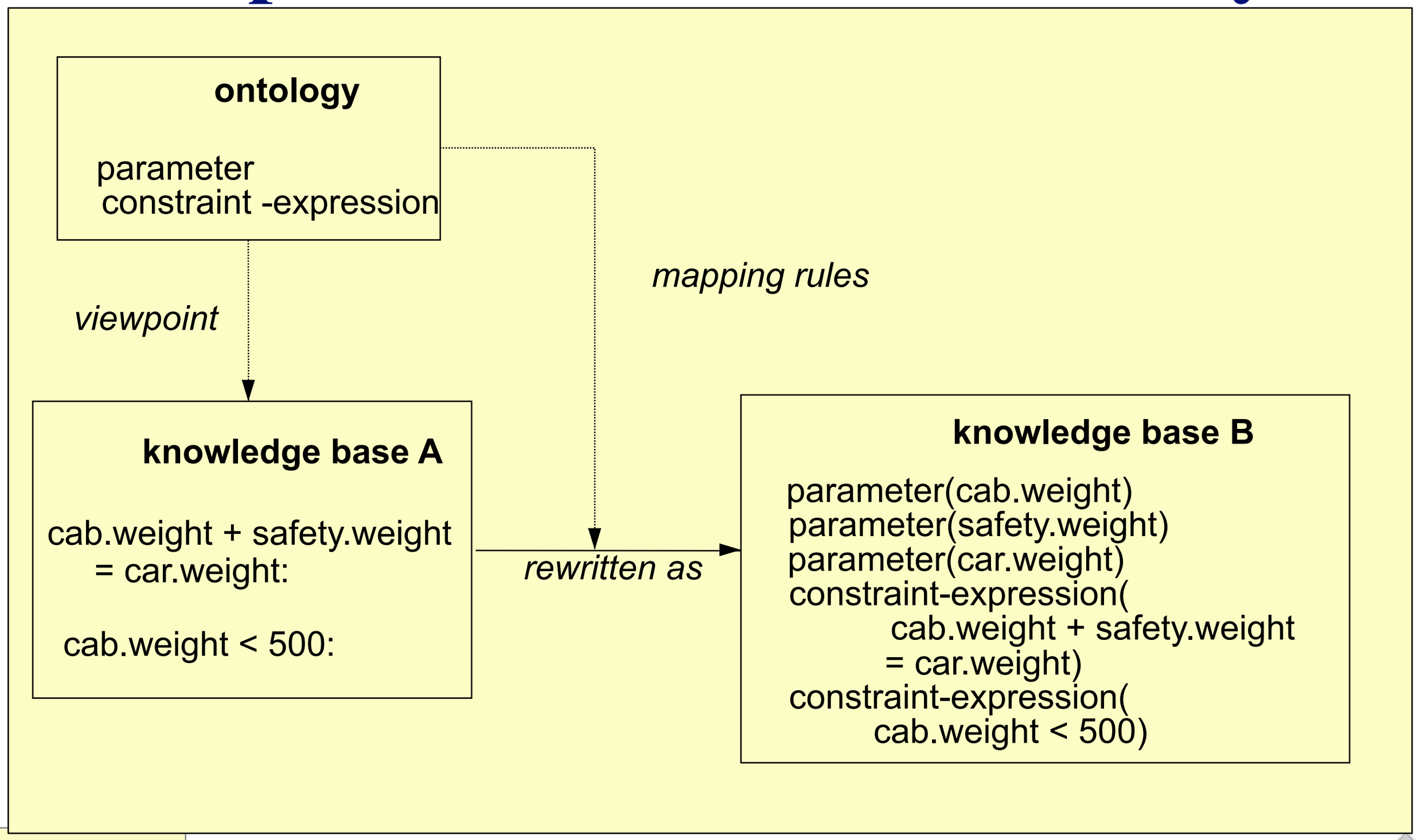
- examines the relationships between words, and the corresponding concepts and objects
  - in practice, it often combines aspects of thesaurus and dictionary
  - frequently uses a graph-based visual representation to indicated relationships between words
- used to identify and specify a vocabulary for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

# 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 Should Allow for "Representational Promiscuity"




[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

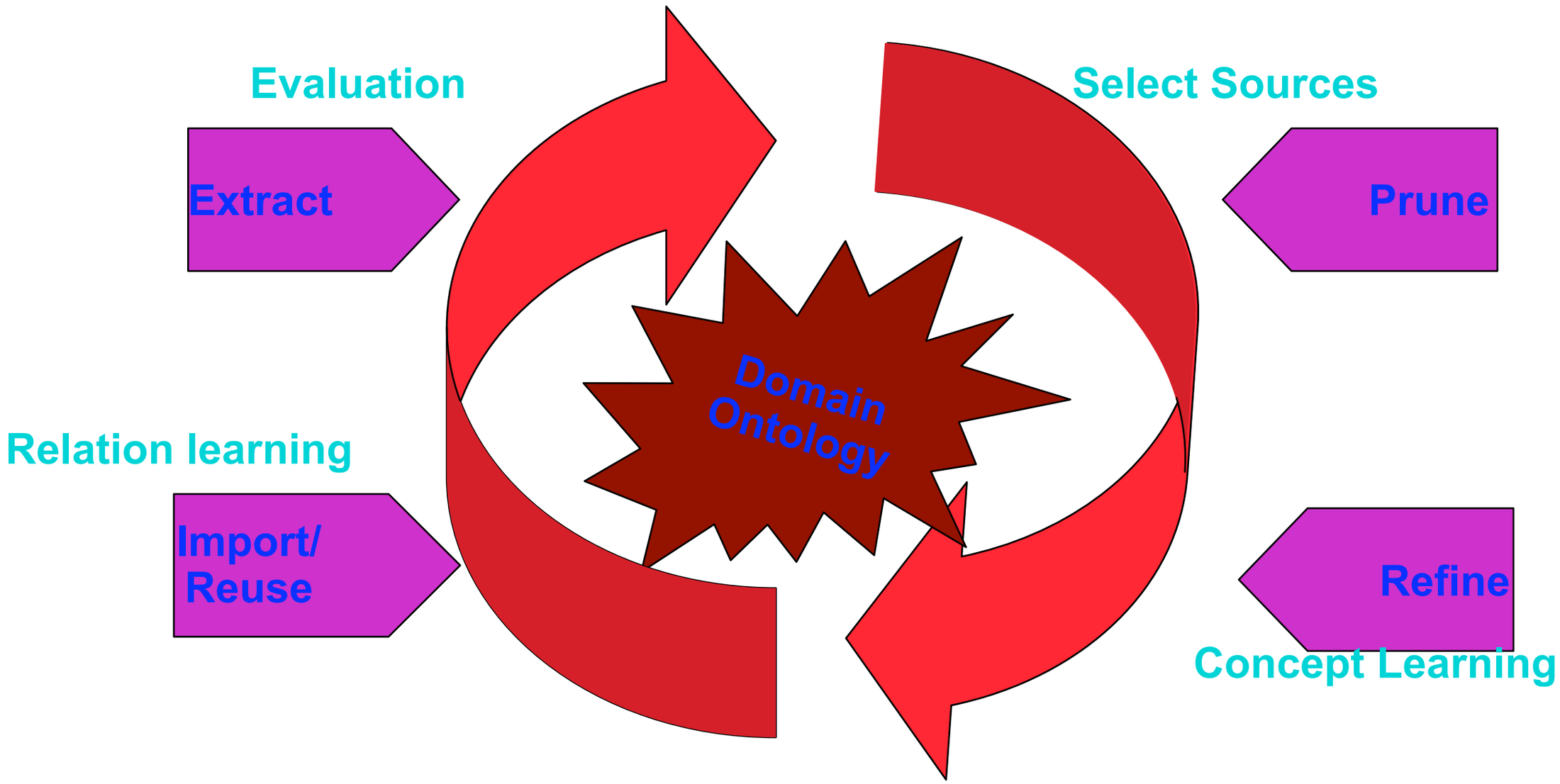

[Schreiber 2000]

# Domain Standards and Vocabularies As Ontologies

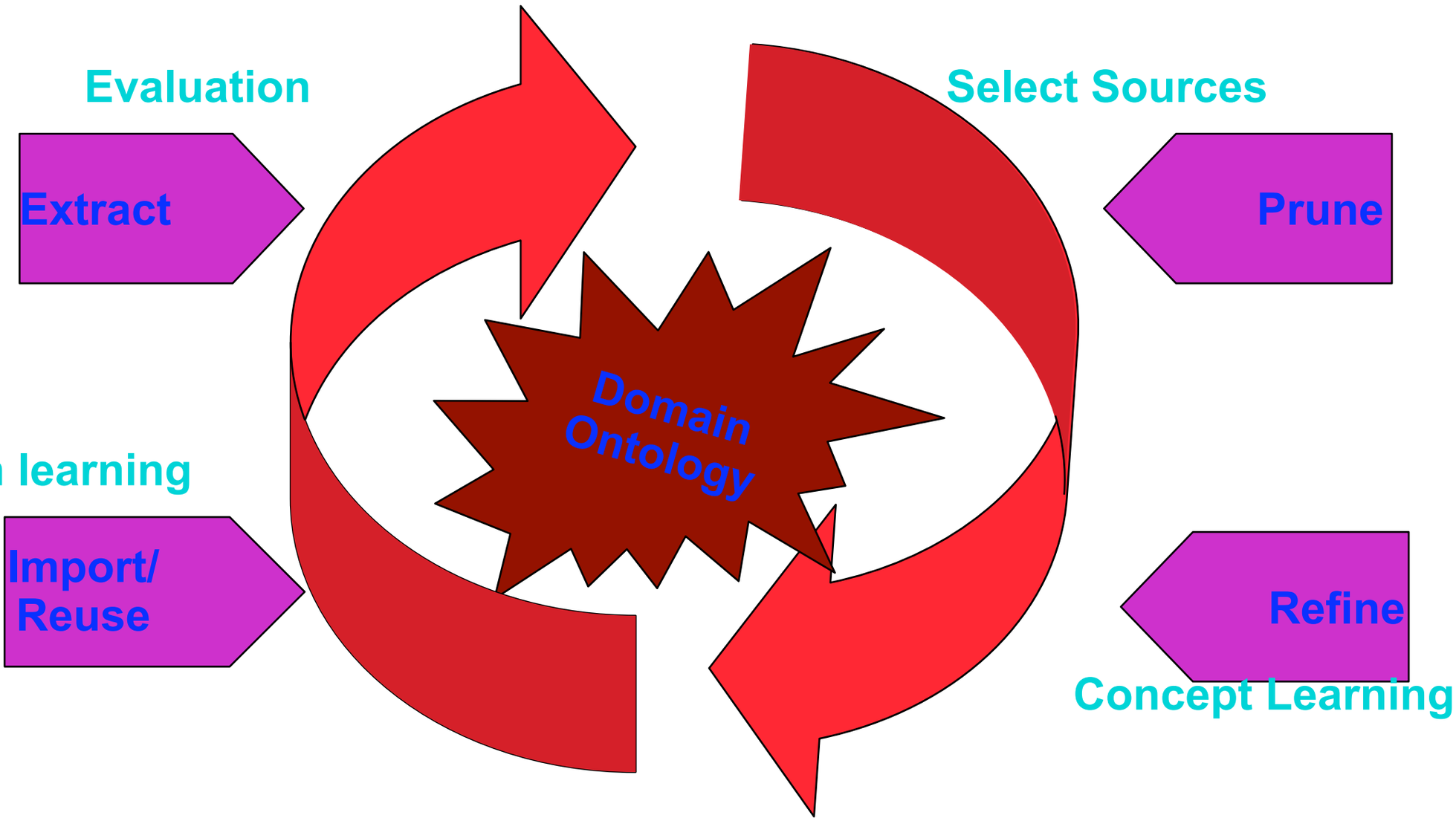
- example: Art and Architecture Thesaurus (AAT)
- contains ontological information
  - AAT: structure of the hierarchy
- structure needs to be “extracted”
  - not explicit
- can be made available as an ontology
  - with help of some mapping formalism
- lists of domain terms are sometimes also called “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 Development



# Ontology Development




Scott Patterson, CS8350

Kietz, Maedche, Voltz; A Method for Semi-Automatic Ontology acquisition from a Corporate Intranet

Maedche & Staab; Ontology Learning for the Semantic Web

# 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)

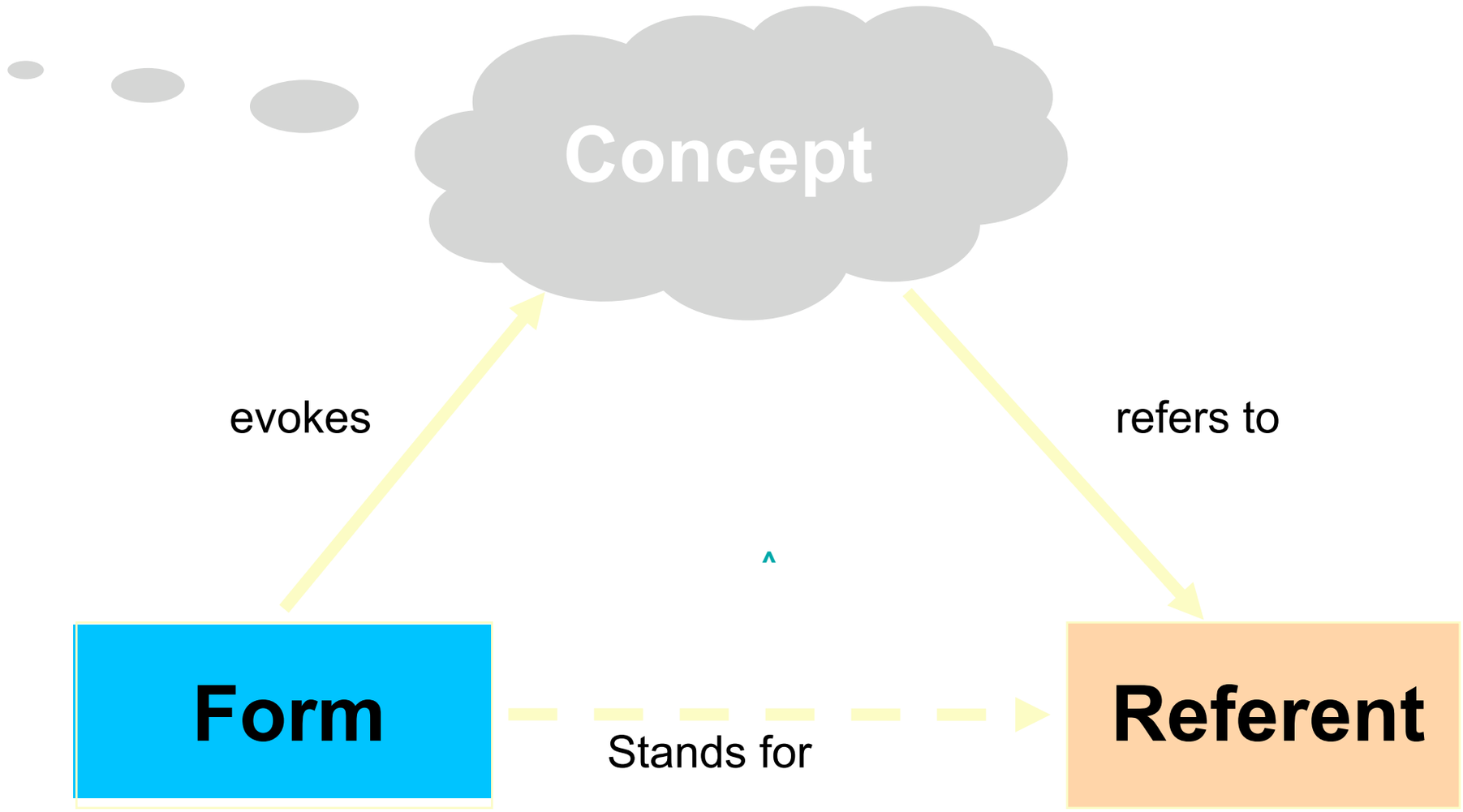
[Schreiber 2000]

# 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

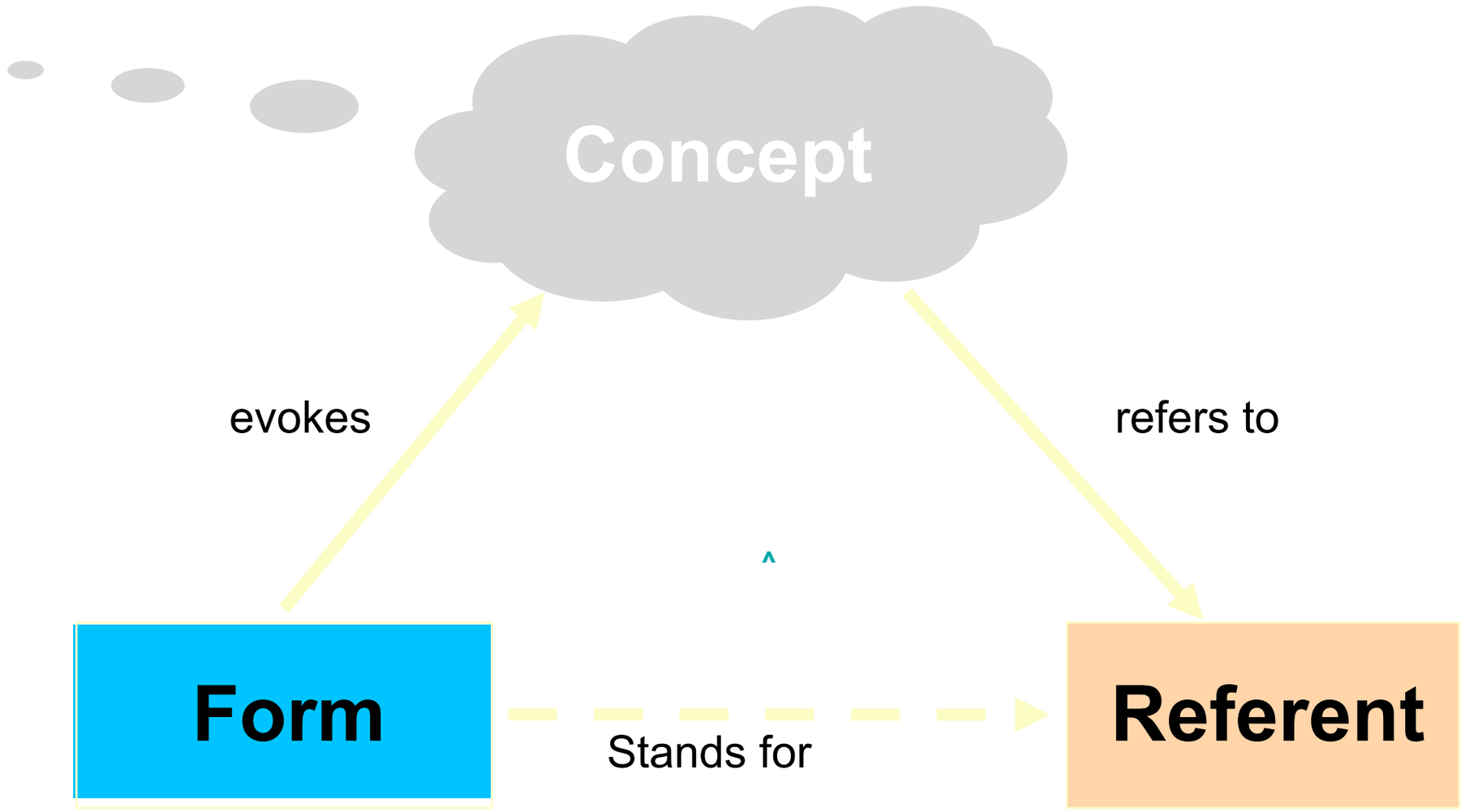
# Communication Principle




**[Odwen, Richards, 1923]**

[Hotho, Sure, 2003]

# Communication Principle

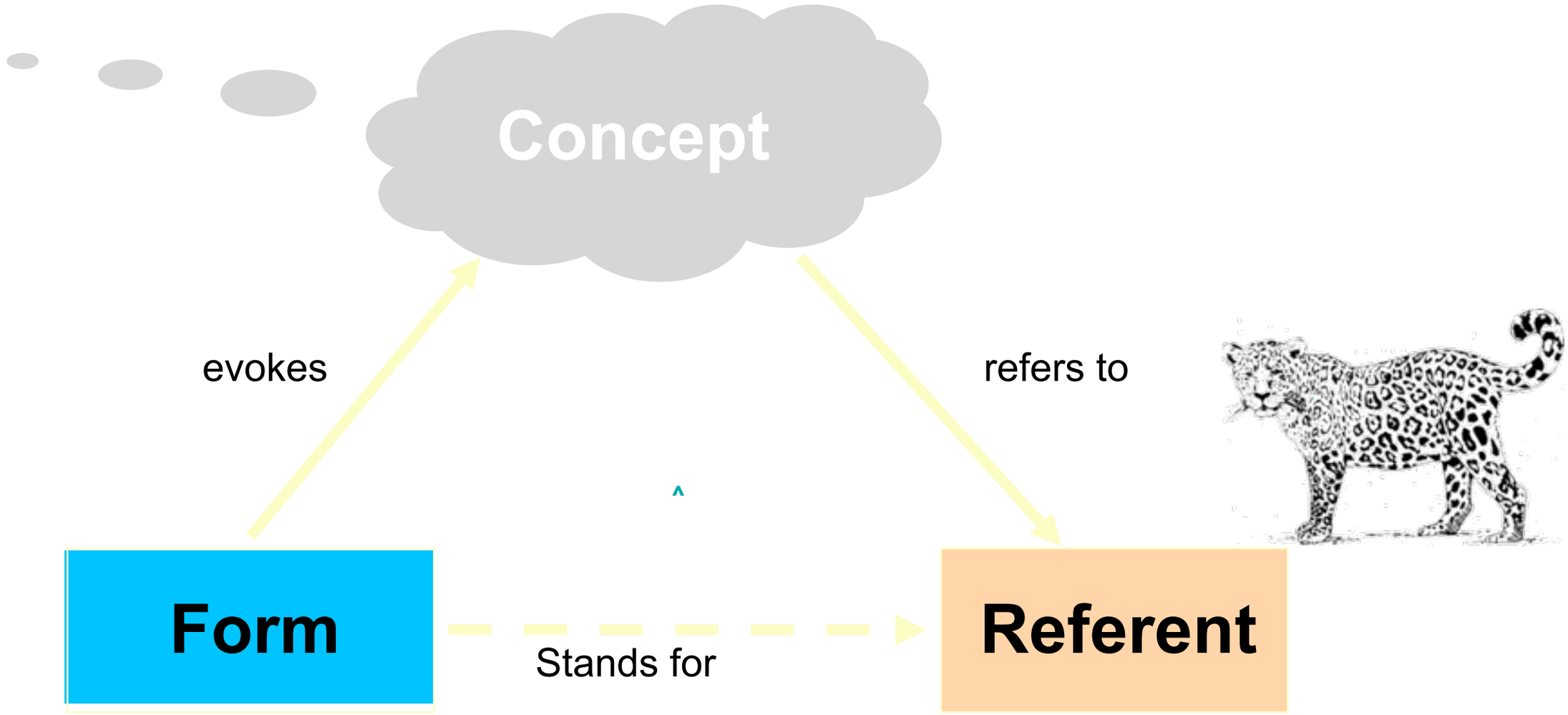



***"Jaguar"***

**[Odwen, Richards, 1923]**

[Hotho, Sure, 2003]

# Communication Principle

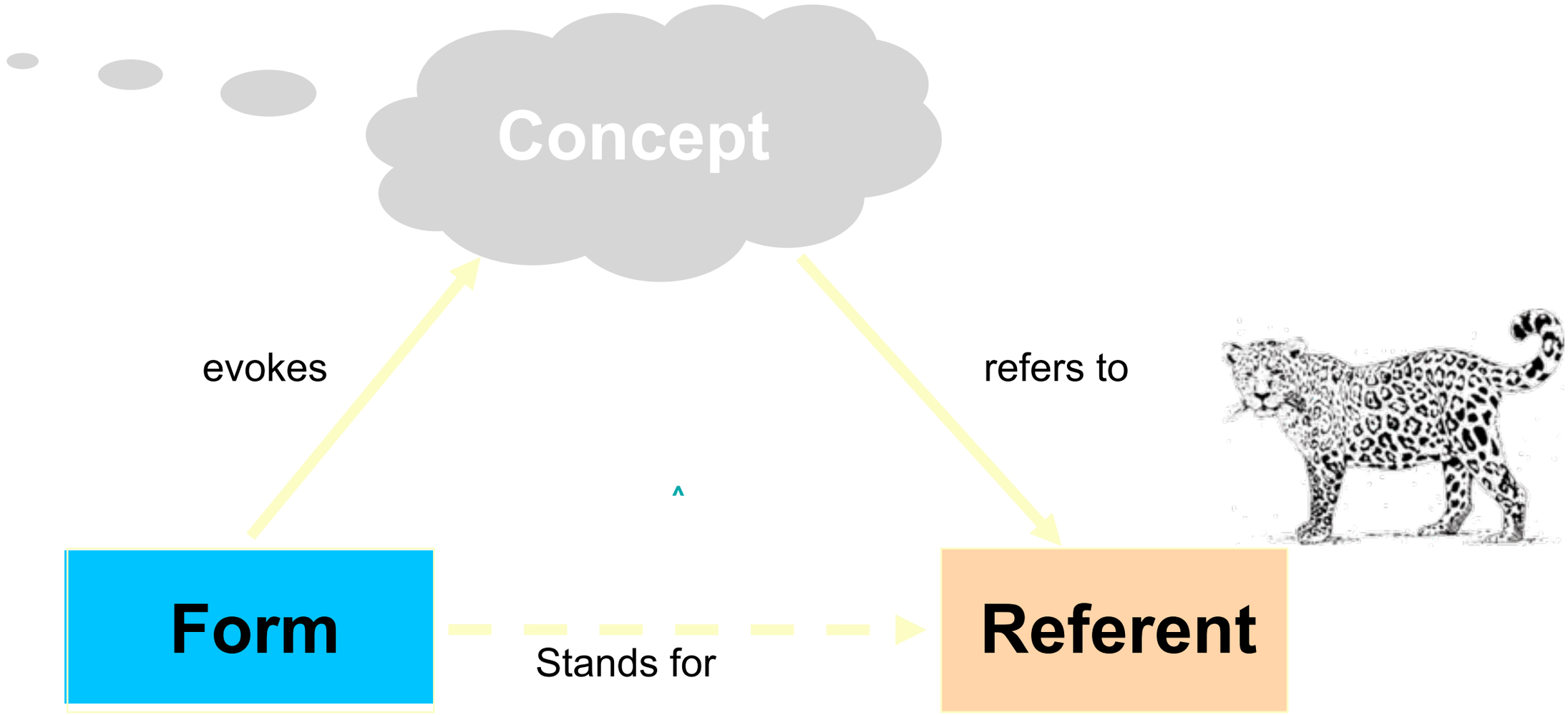



***"Jaguar"***

**[Odwen, Richards, 1923]**

[Hotho, Sure, 2003]

# Communication Principle




**[Odwen, Richards, 1923]**

[Hotho, Sure, 2003]

# Views on Ontologies

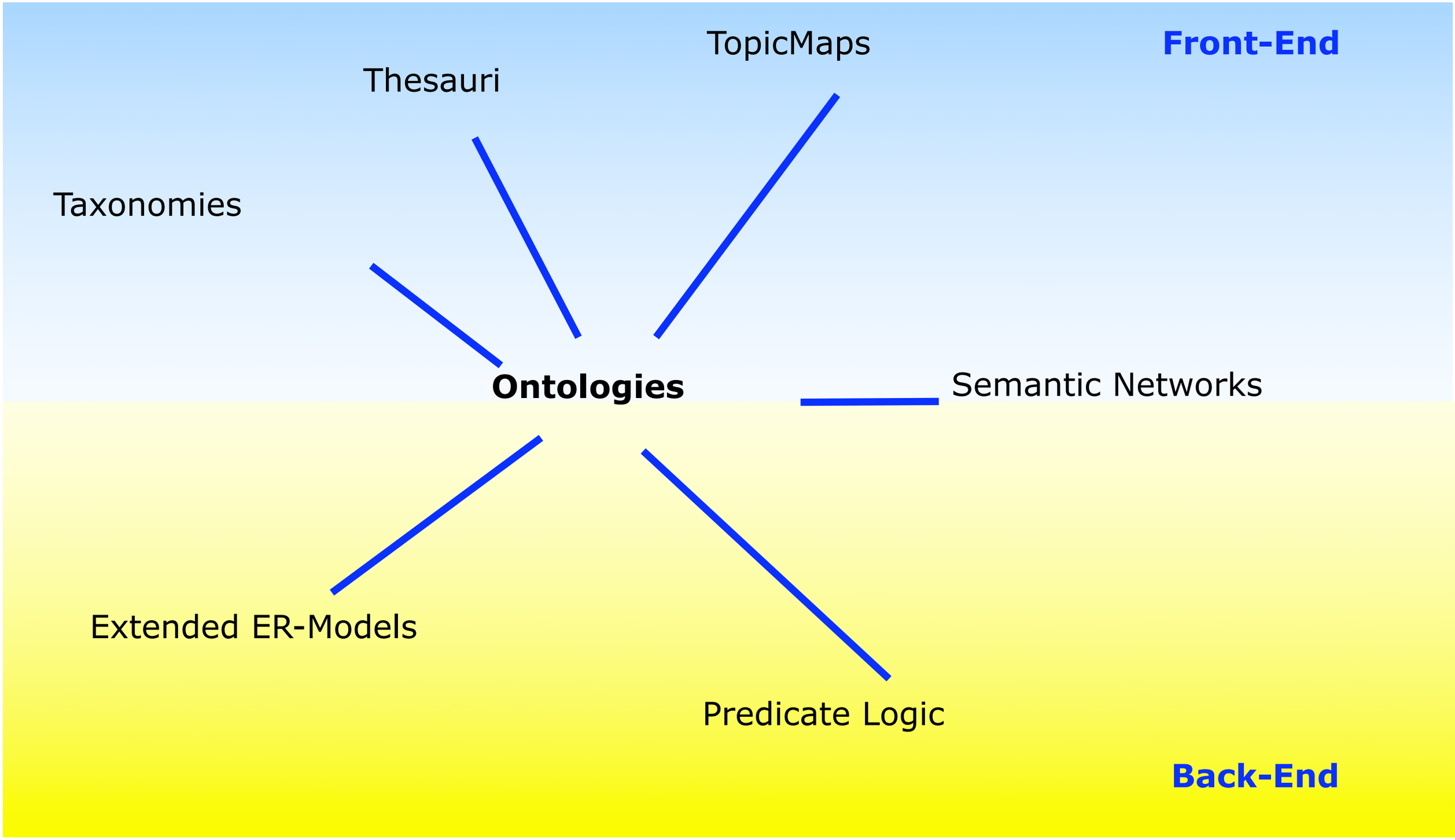
**Front-End**

**Ontologies**

**Back-End**

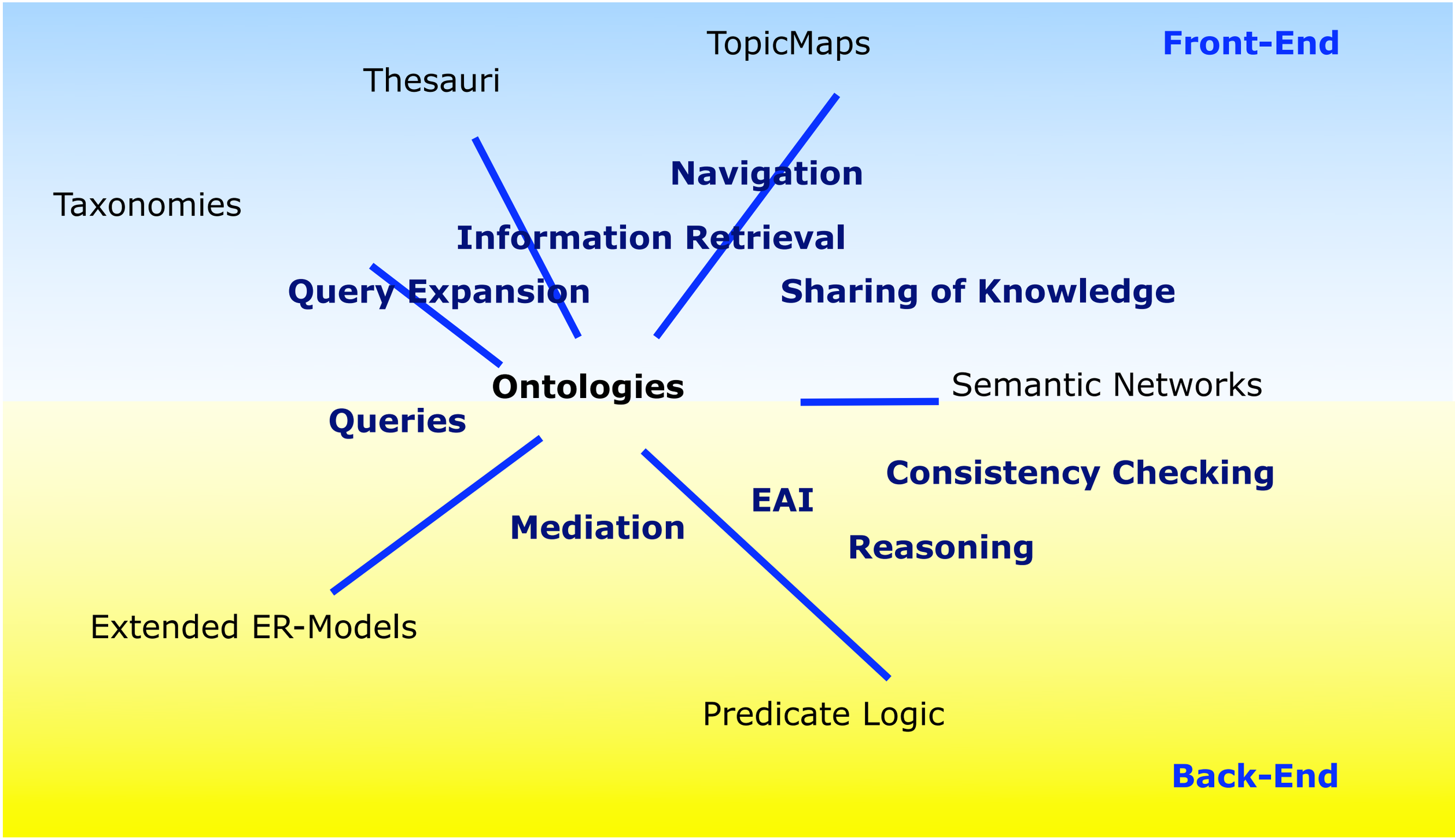
[Hotho, Sure, 2003]

# Views on Ontologies



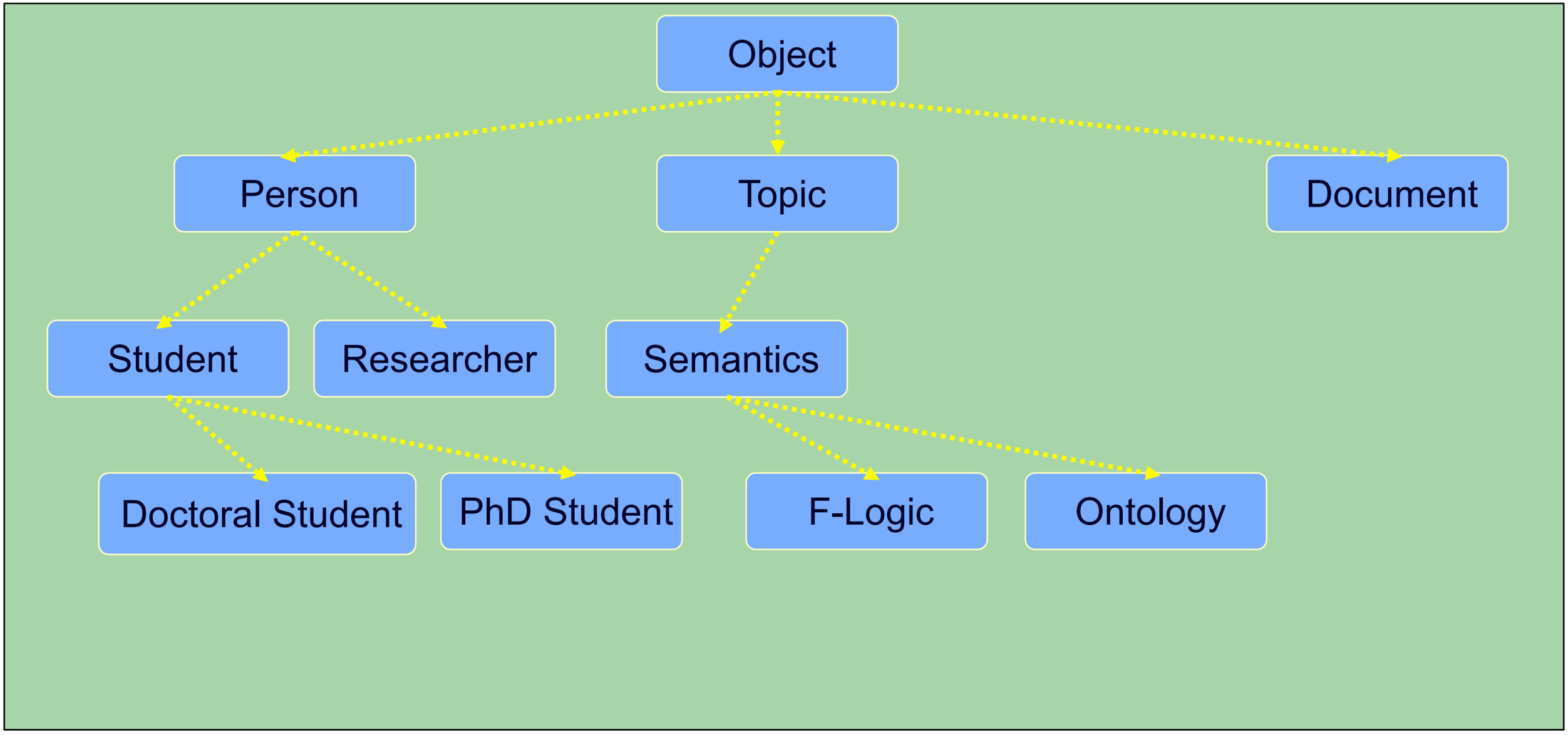

[Hotho, Sure, 2003]

# Views on Ontologies




[Hotho, Sure, 2003]

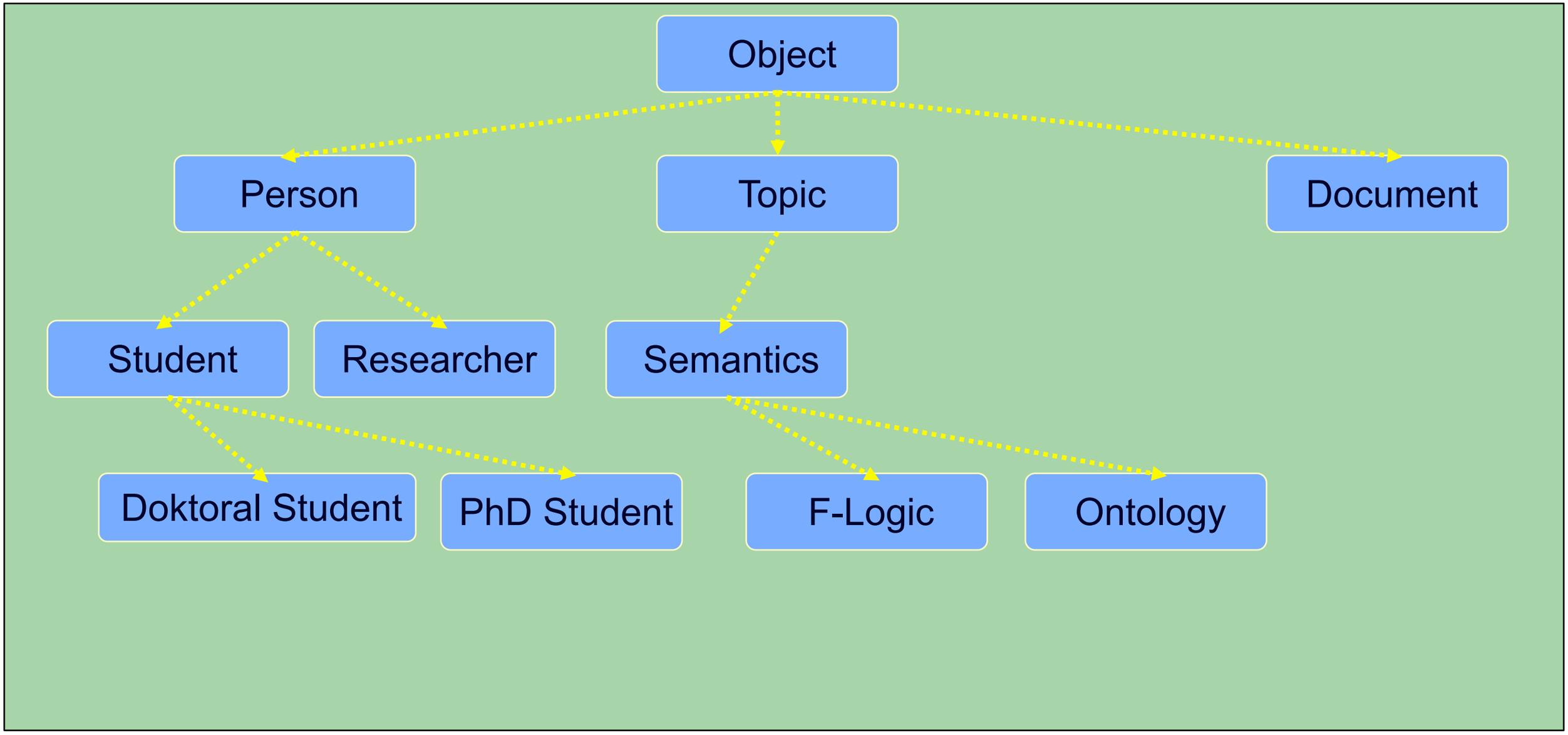
# Taxonomy




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

[Hotho, Sure, 2003]

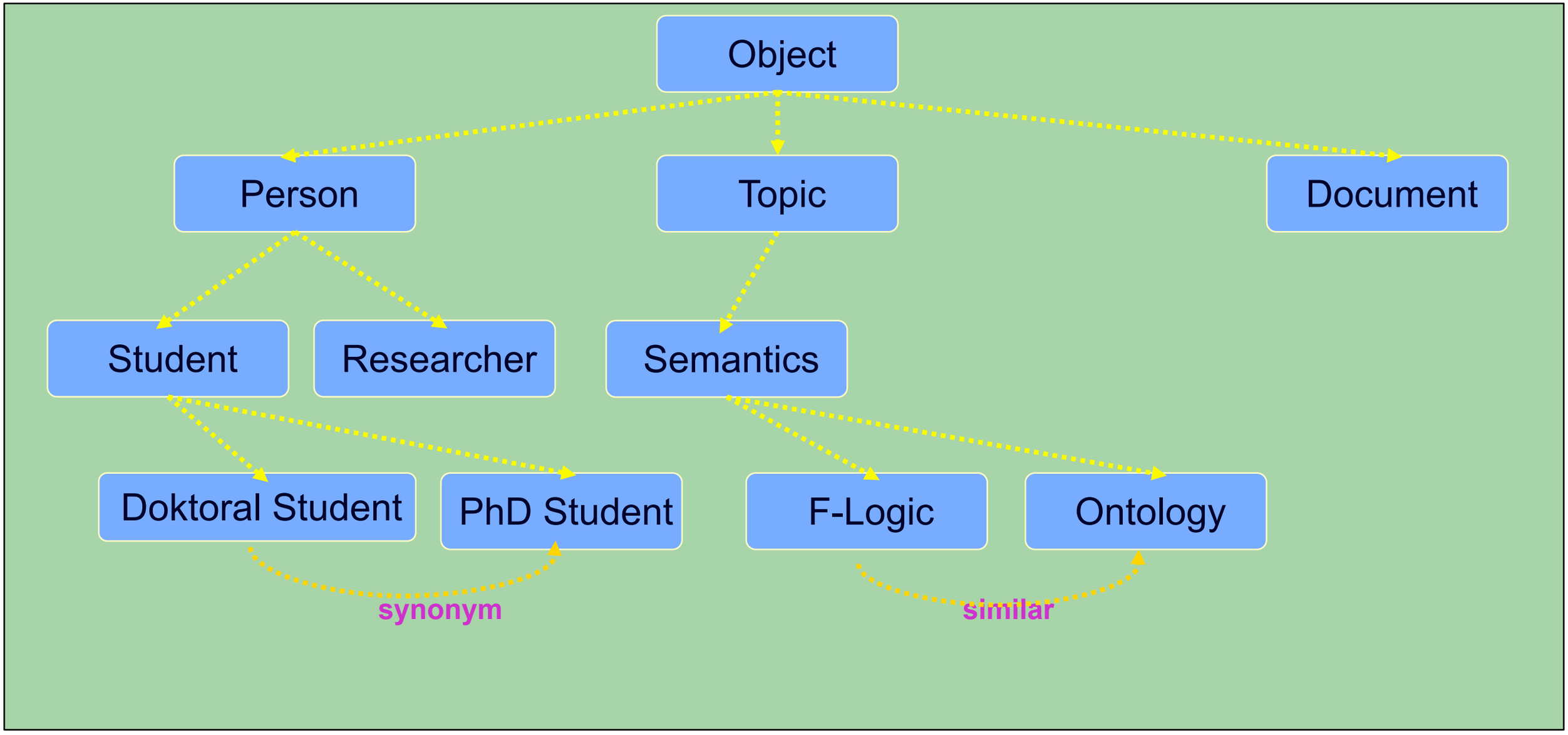
# Thesaurus




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

[Hotho, Sure, 2003]

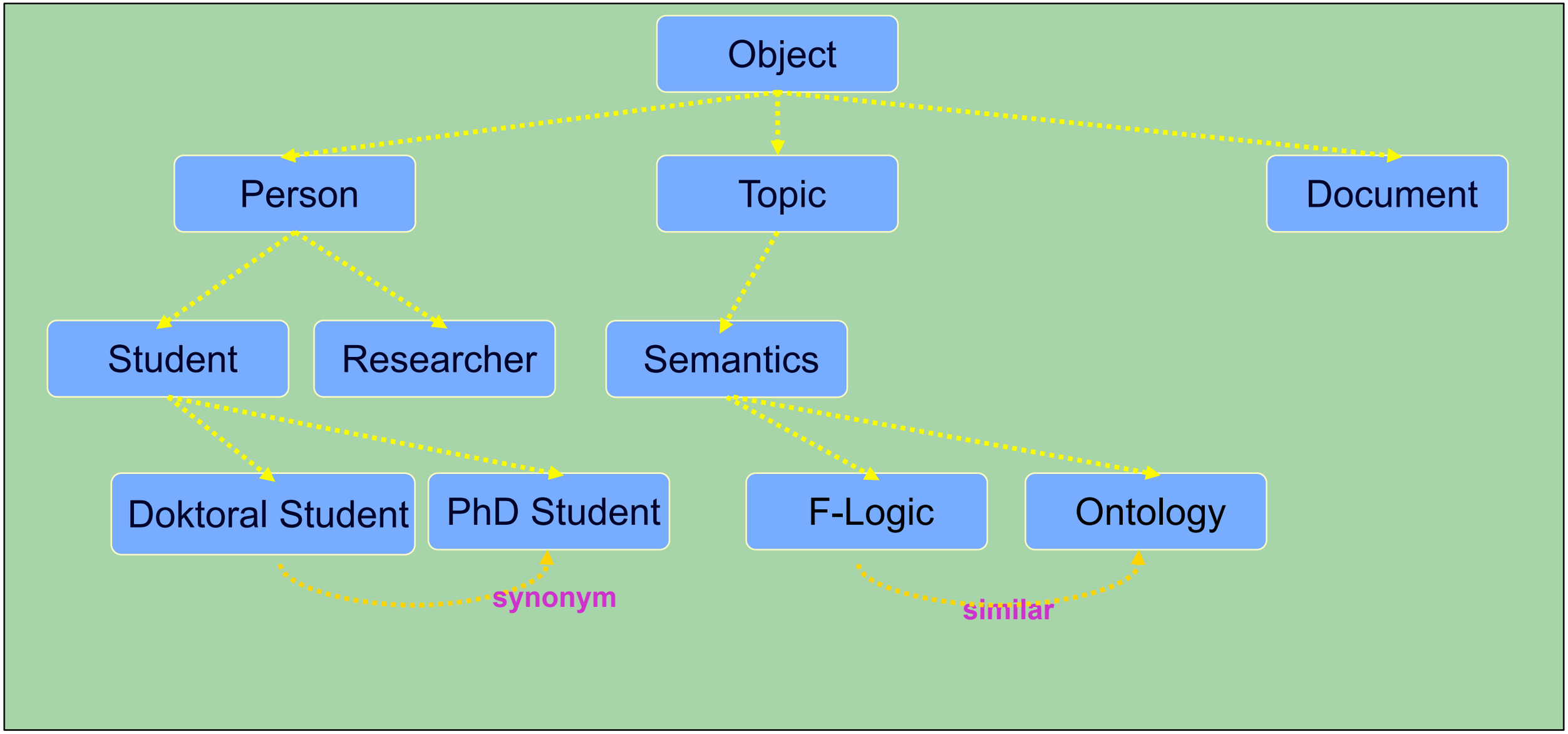
# Thesaurus




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

[Hotho, Sure, 2003]

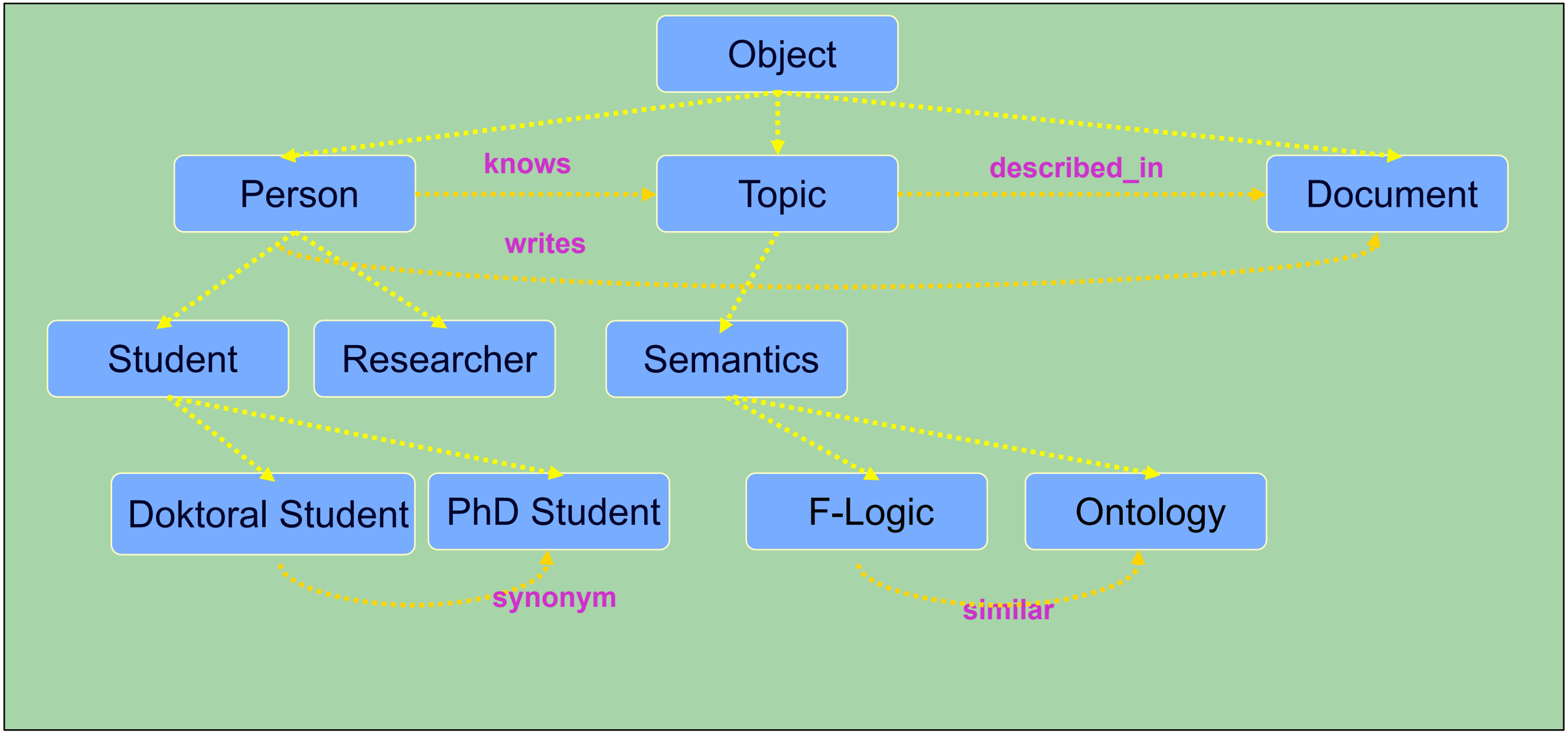
# Topic Map




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

[Hotho, Sure, 2003]

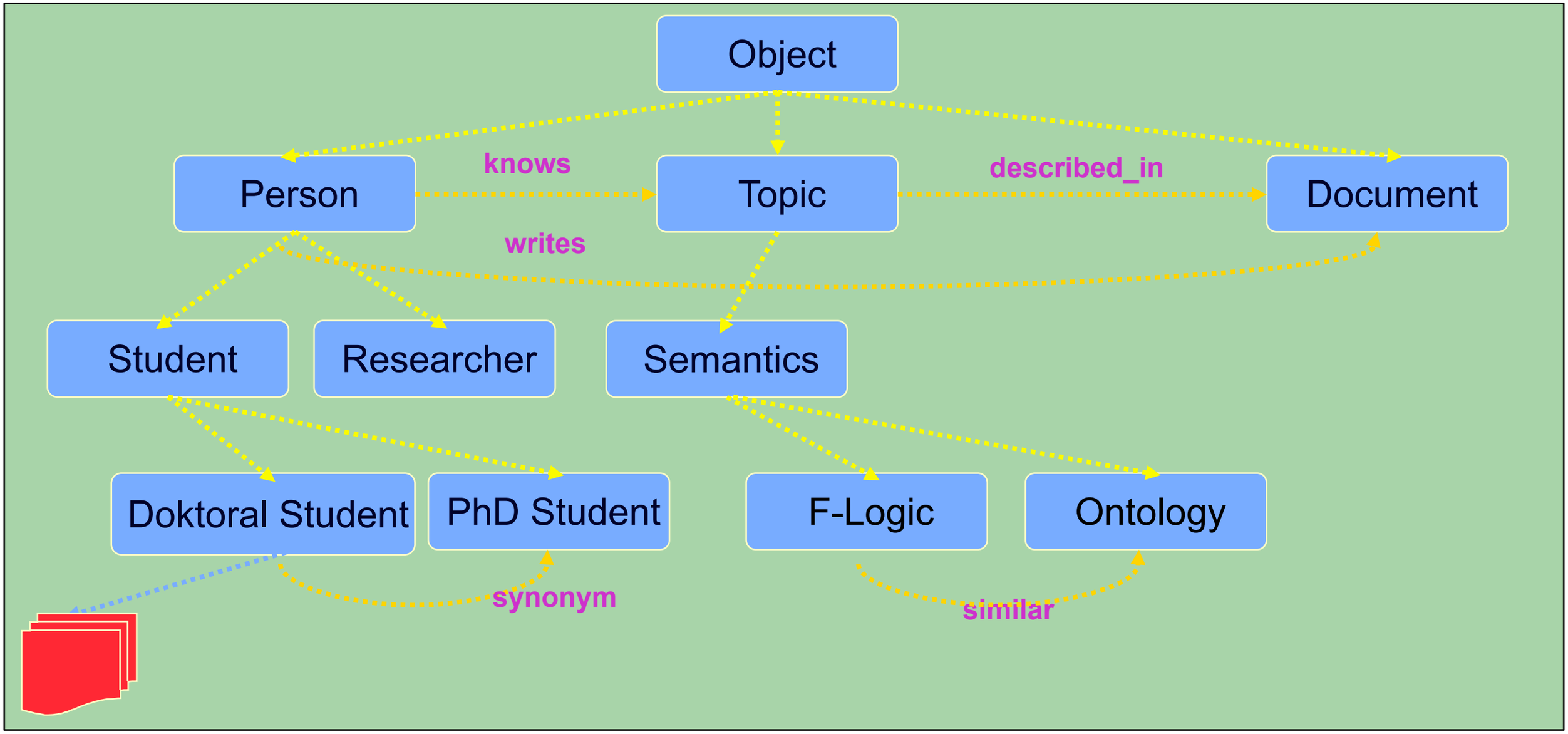
# Topic Map




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

[Hotho, Sure, 2003]

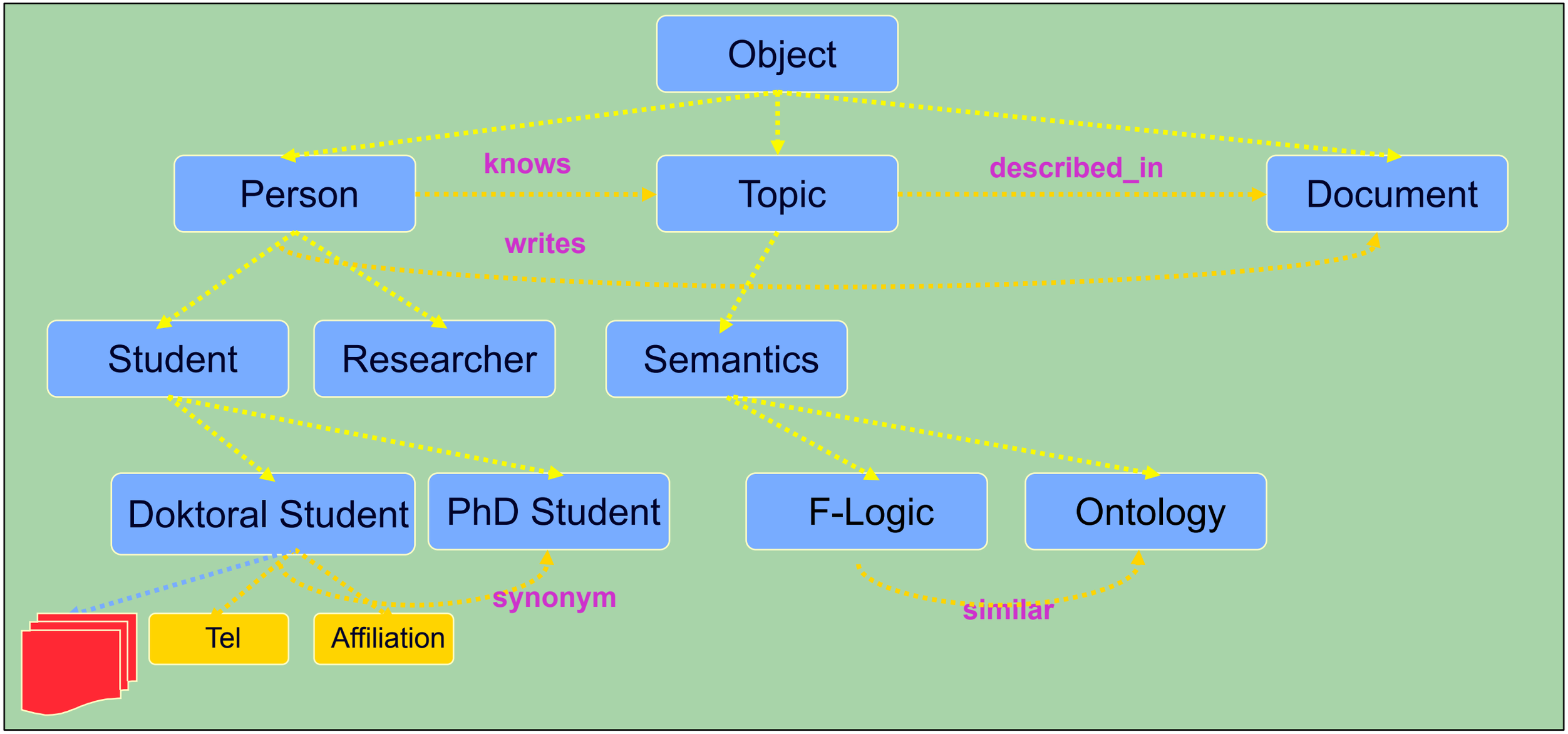
# Topic Map




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

[Hotho, Sure, 2003]

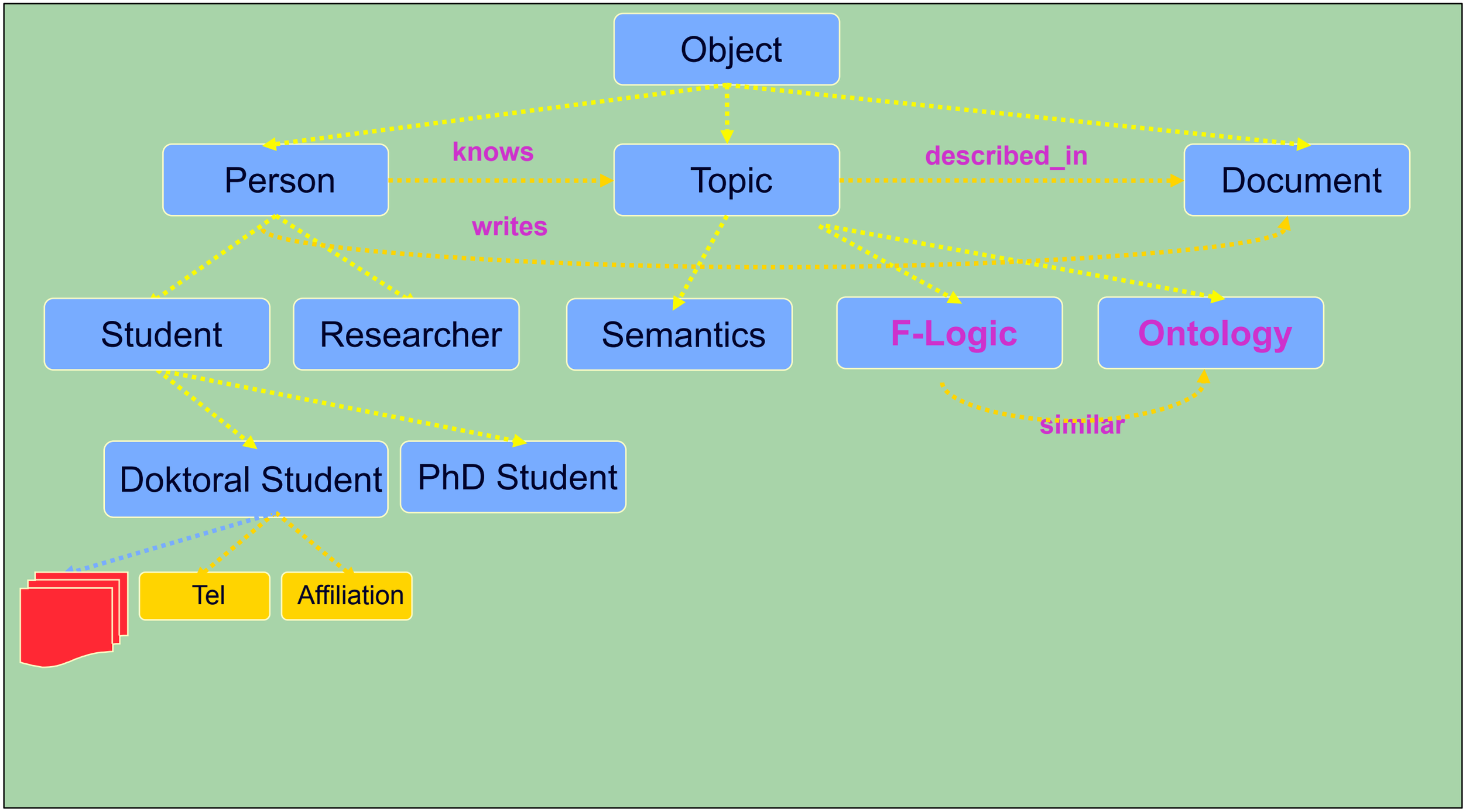
# Topic Map




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

[Hotho, Sure, 2003]

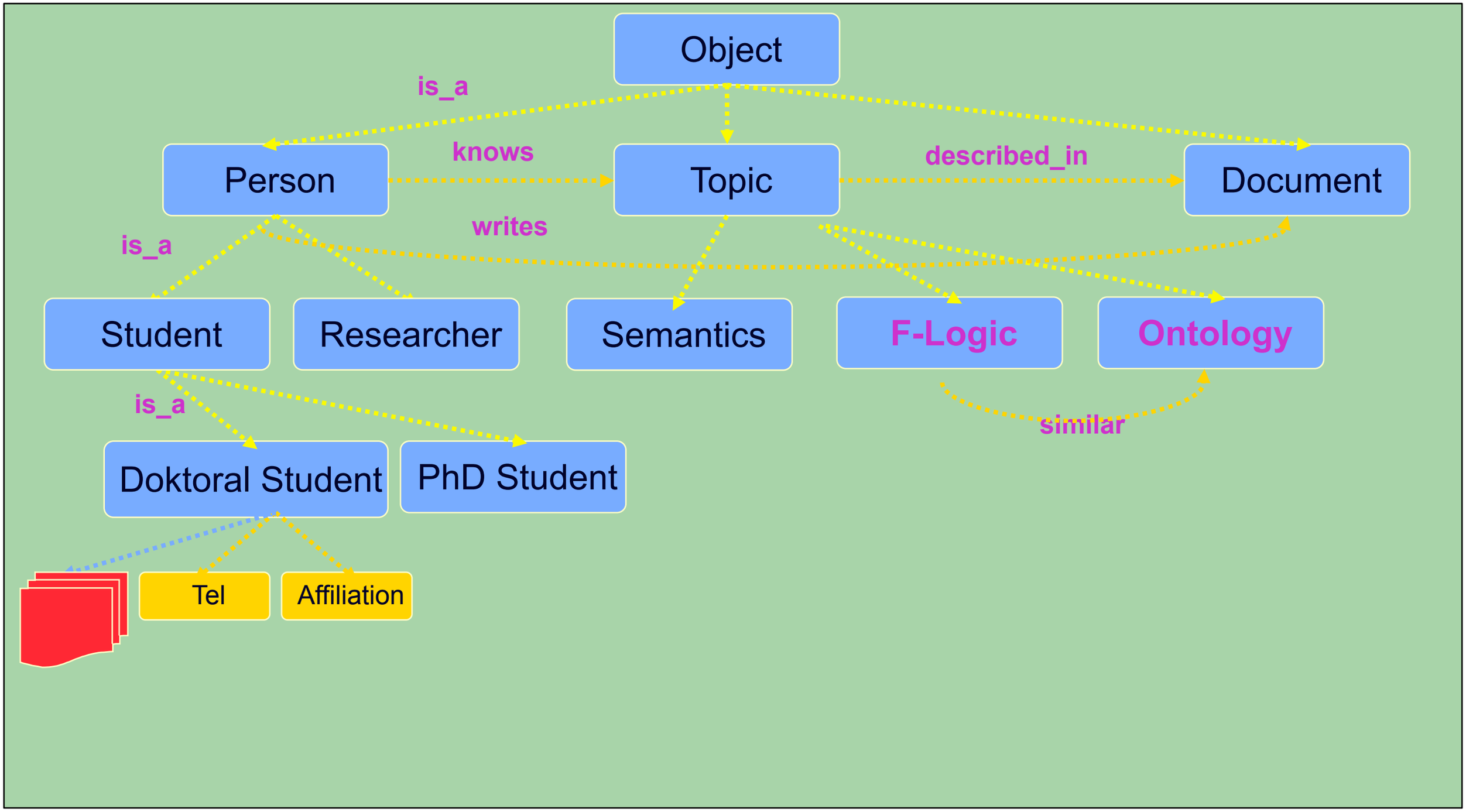
# Ontology




- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

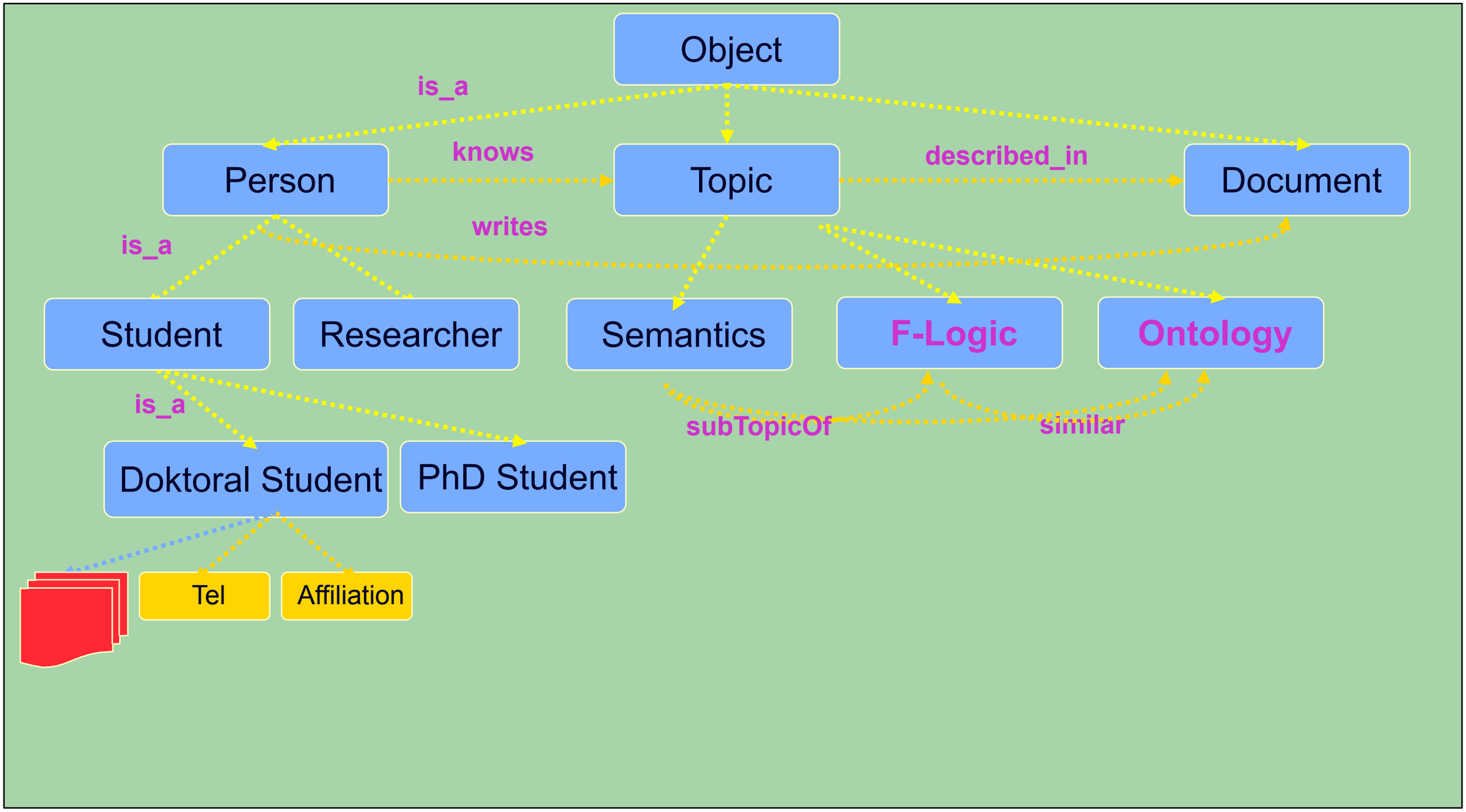
# Ontology




- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

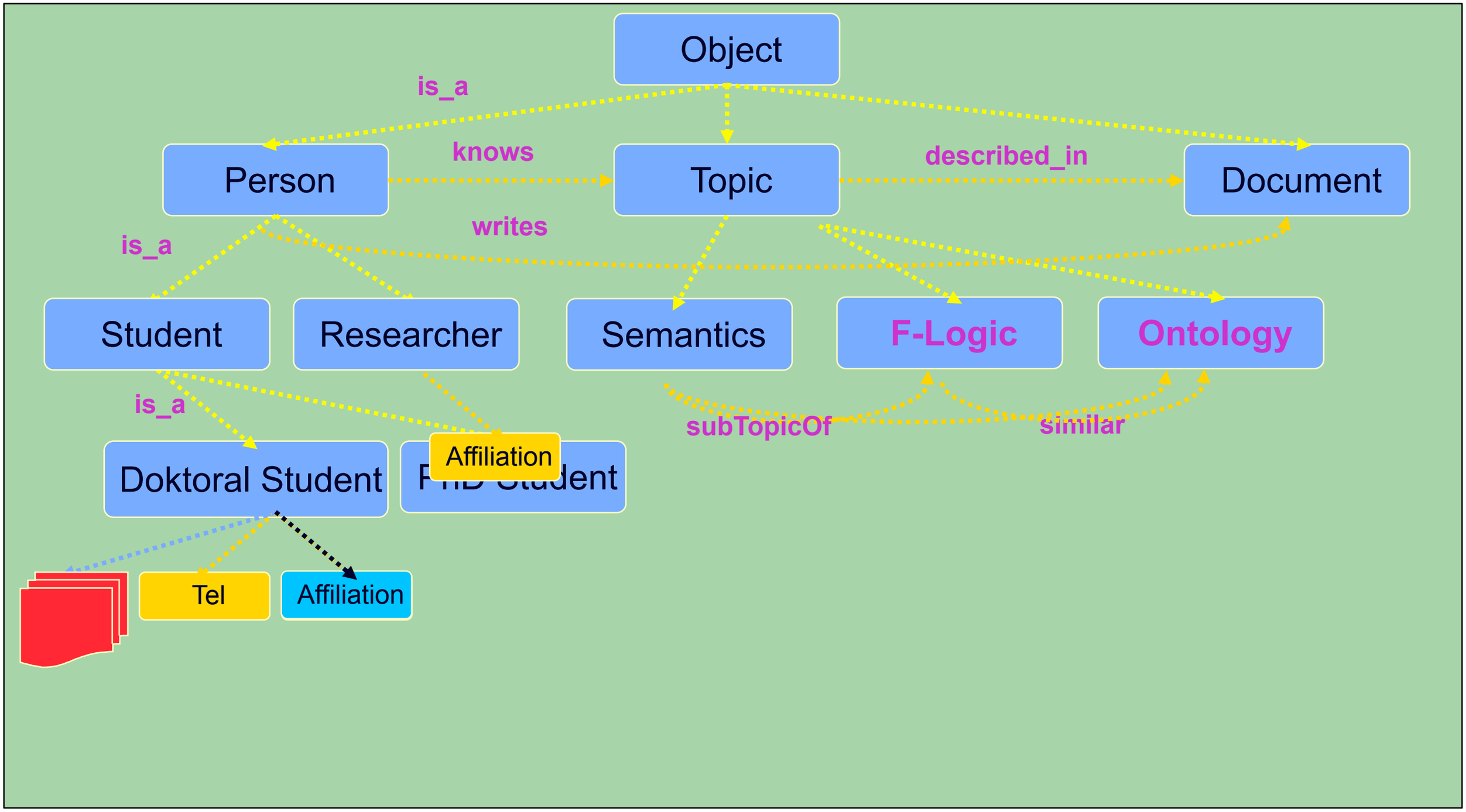
# Ontology




- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

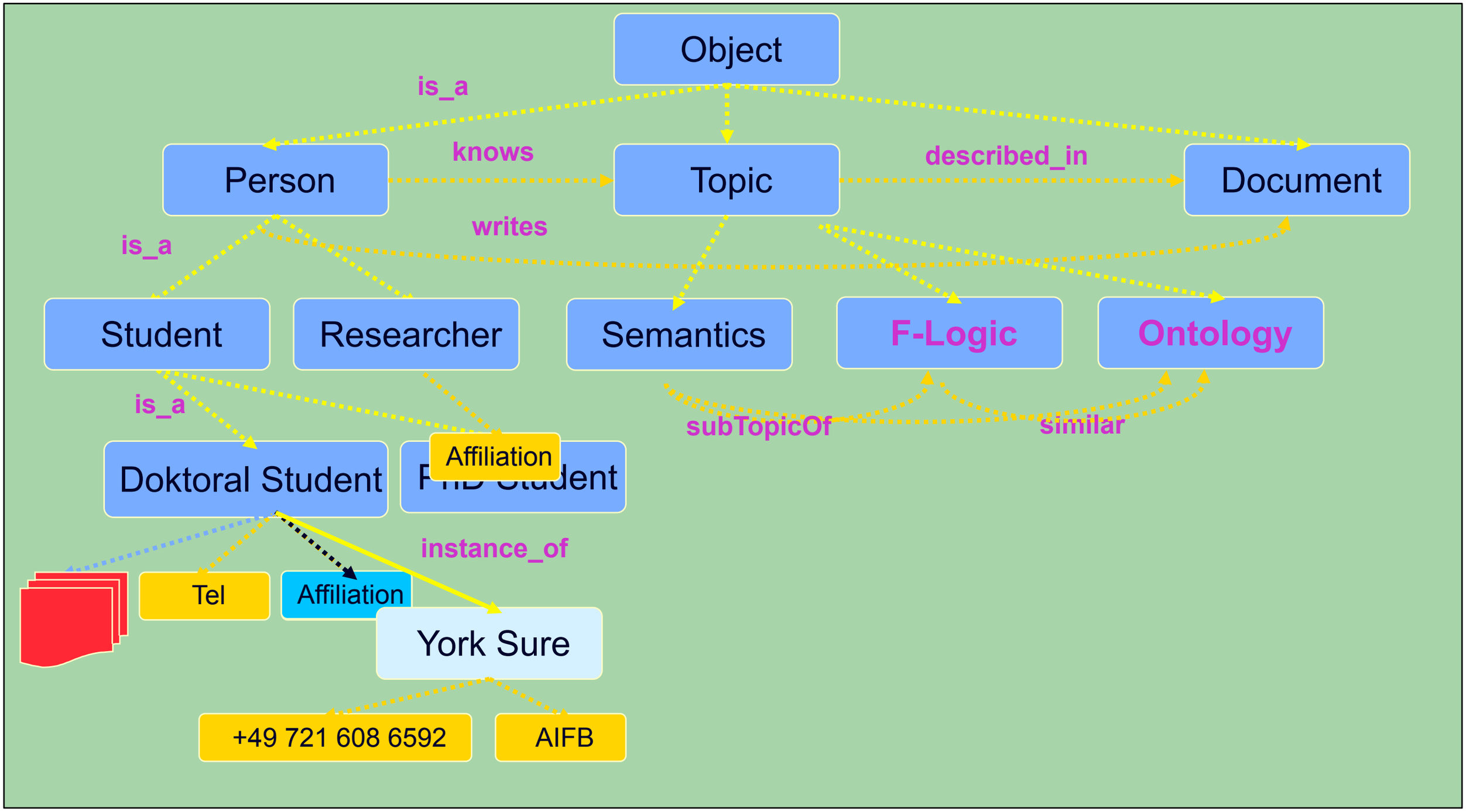
# Ontology




- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

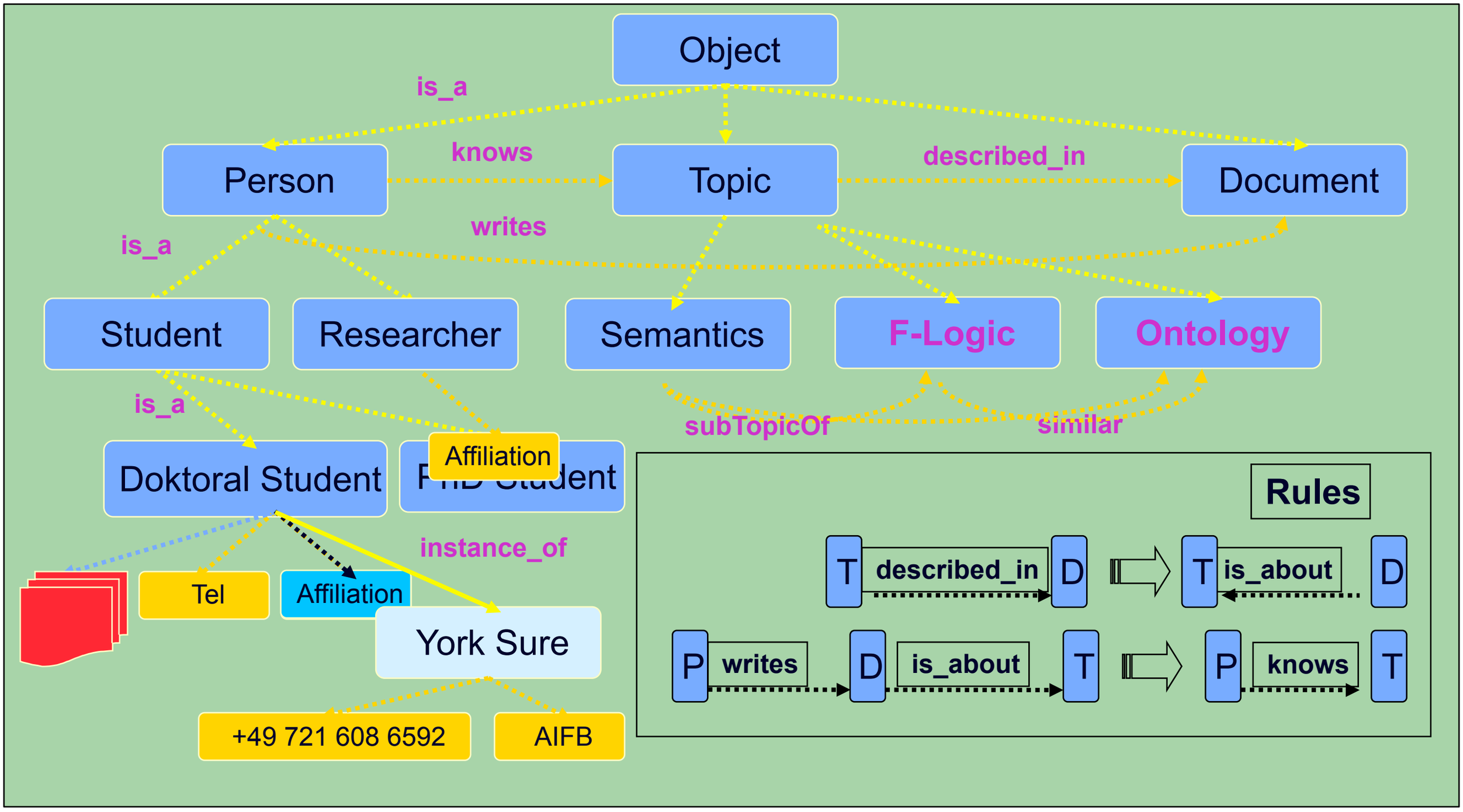
# Ontology




- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

# Ontology

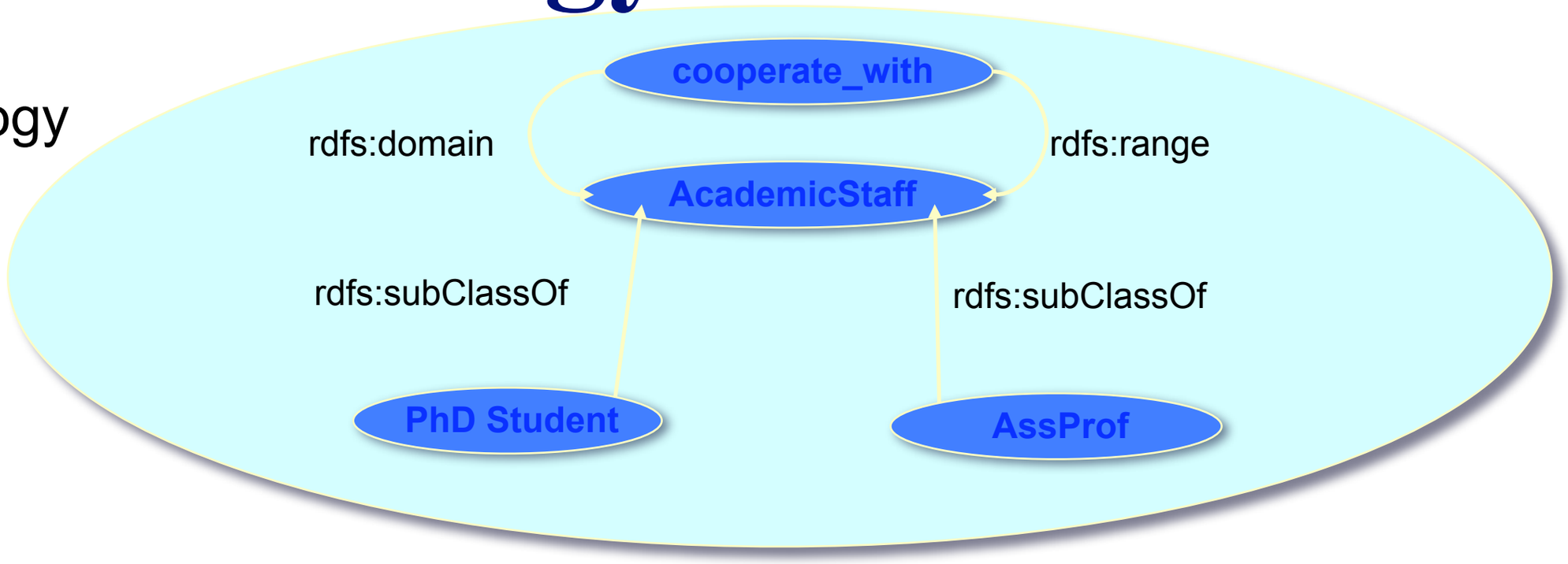



- **Representation Language: Predicate Logic (F-Logic)**
- **Standards: RDF(S); coming up standard: OWL**

[Hotho, Sure, 2003]

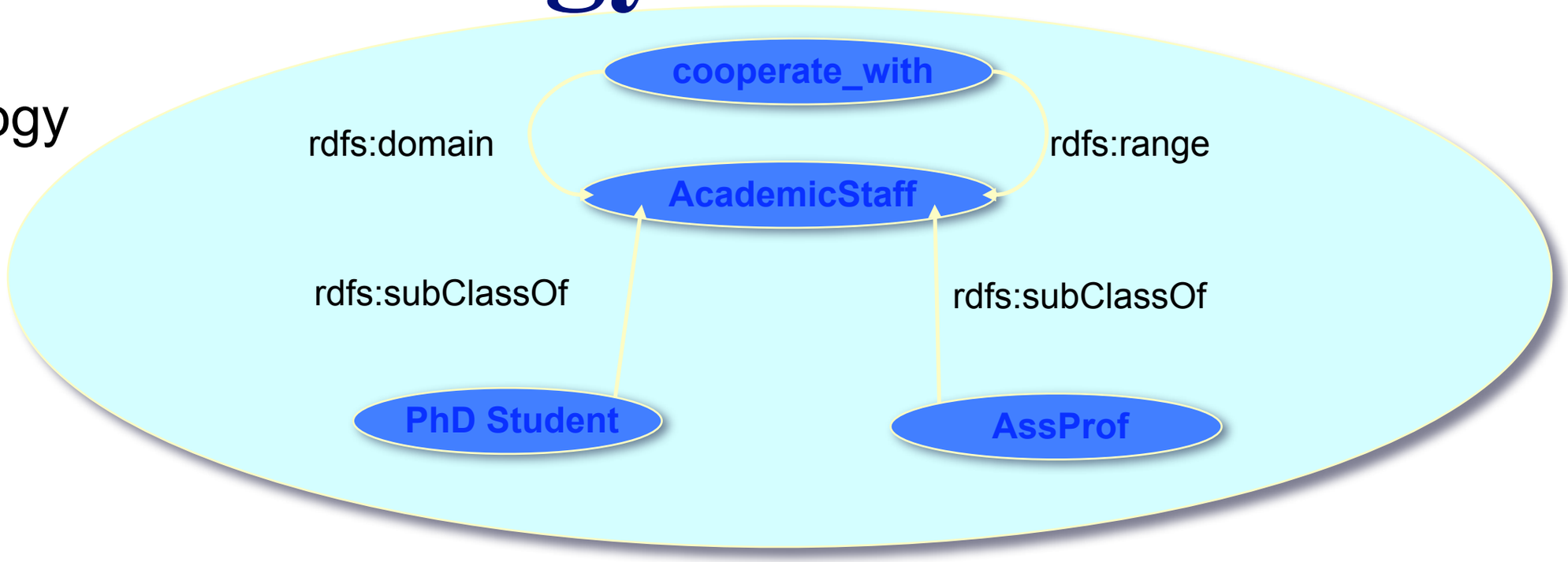
# Ontology & Metadata

Ontology



# Ontology & Metadata

Ontology




Anno-tation

```
<swrc:PhD_Student rdf:ID="sha">
 <swrc:name>Siegfried Handschuh</
swrc:name>

 <swrc:cooperate_with rdf:resource =
   "http://www.aifb.uni-
 karlsruhe.de/WBS/sst#sst"/>
...
</swrc:PhD_Student>
```

```
<swrc:AssProf rdf:ID="sst">
   <swrc:name>Steffen Staab
   </swrc:name>
...
</swrc:AssProf>
```

Web Page



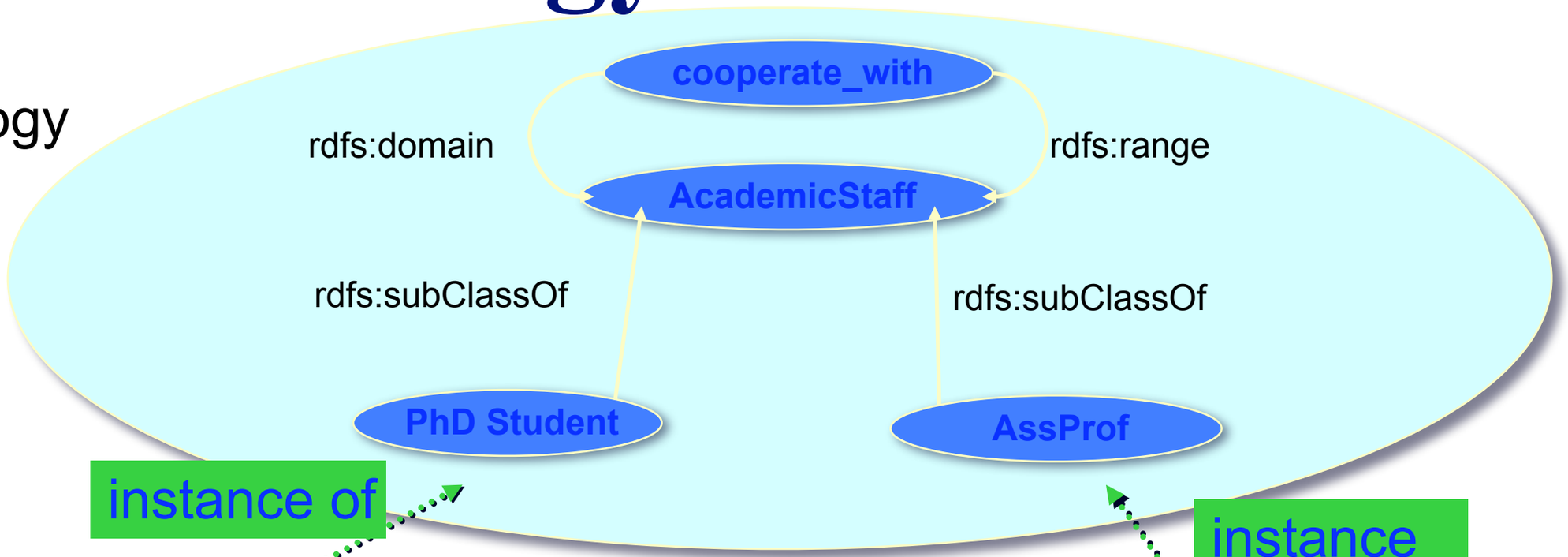
URL

http://www.aifb.uni-karlsruhe.de/WBS/sha

http://www.aifb.uni-karlsruhe.de/WBS/sst

# Ontology & Metadata

Ontology




instance of

instance of

Anno-tation

```
<swrc:PhD_Student rdf:ID="sha">
 <swrc:name>Siegfried Handschuh</
swrc:name>

 <swrc:cooperate_with rdf:resource =
   "http://www.aifb.uni-
 karlsruhe.de/WBS/sst#sst"/>
...
</swrc:PhD_Student>
```

```
<swrc:AssProf rdf:ID="sst">
    <swrc:name>Steffen Staab
    </swrc:name>
...
</swrc:AssProf>
```

Cooperate_with

Links have explicit meanings!

Web Page

**Siegfried Handschuh**

He is working together with Steffen Staab in the Knowledge Management Group

**Research:**

Semantic Web, Knowledge Management, Natural Language,

URL

http://www.aifb.uni-karlsruhe.de/WBS/sha

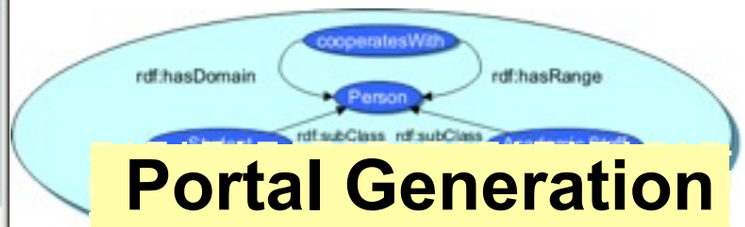
http://www.aifb.uni-karlsruhe.de/WBS/sst

# Knowledge Organization Examples

# OntoWeb.org


[Hotho, Sure, 2003]

# OntoWeb.org

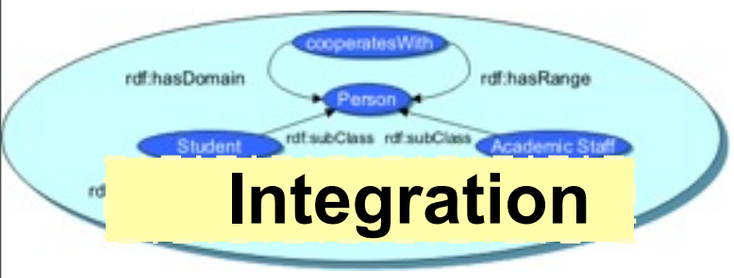

**Portal Generation**

**Navigation**

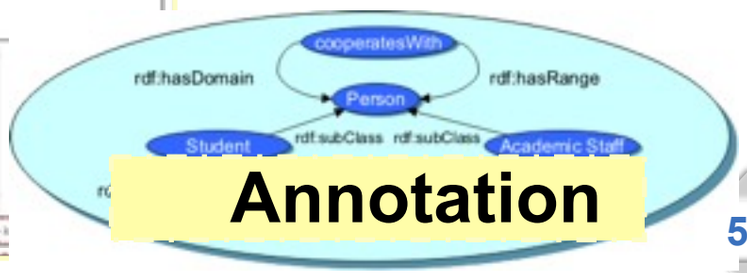
**Query/Serach**

**Content**

**Collect metadata from participating partners**




**Integration**

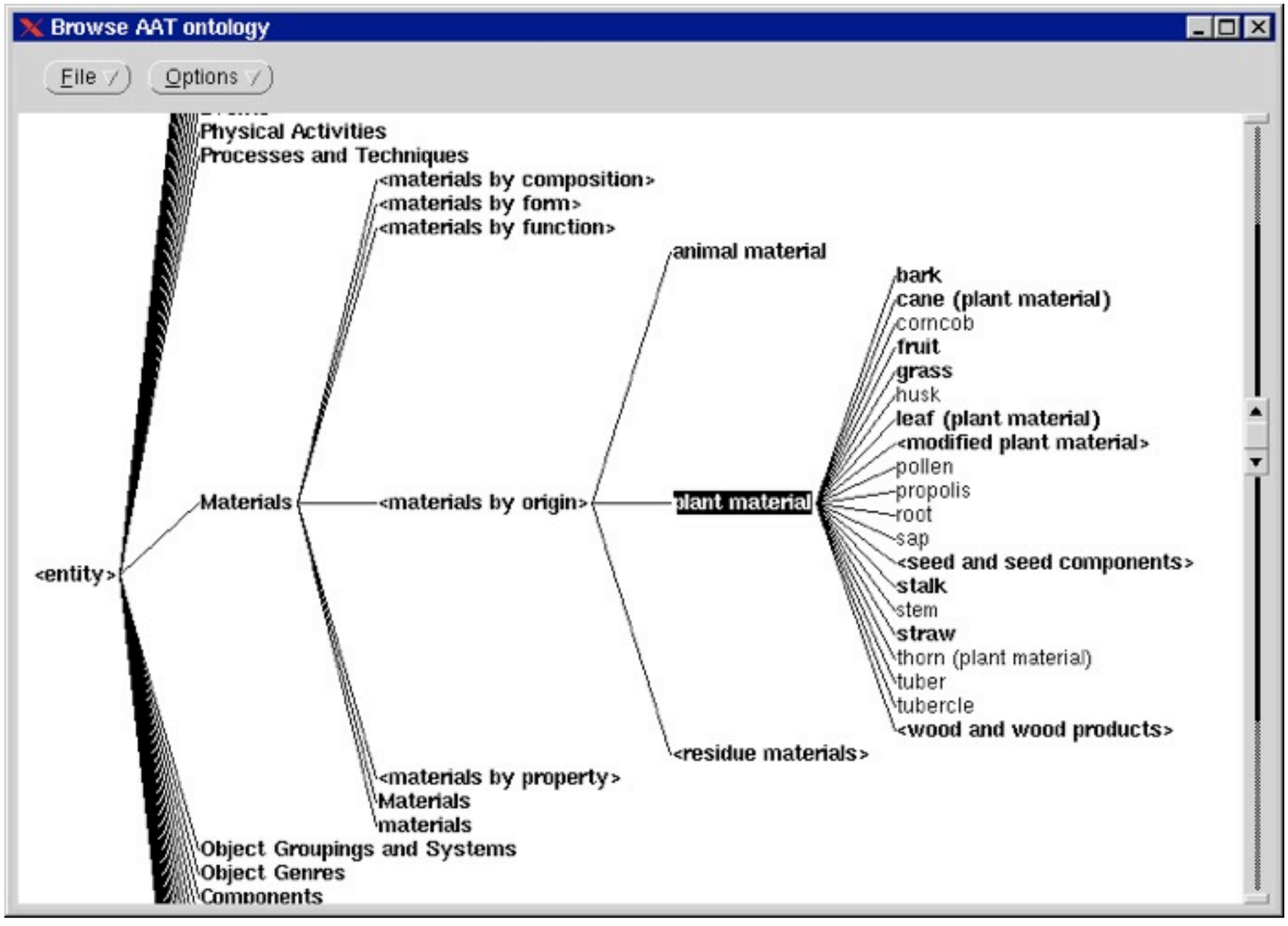




**Annotation**

[Hotho, Sure, 2003]

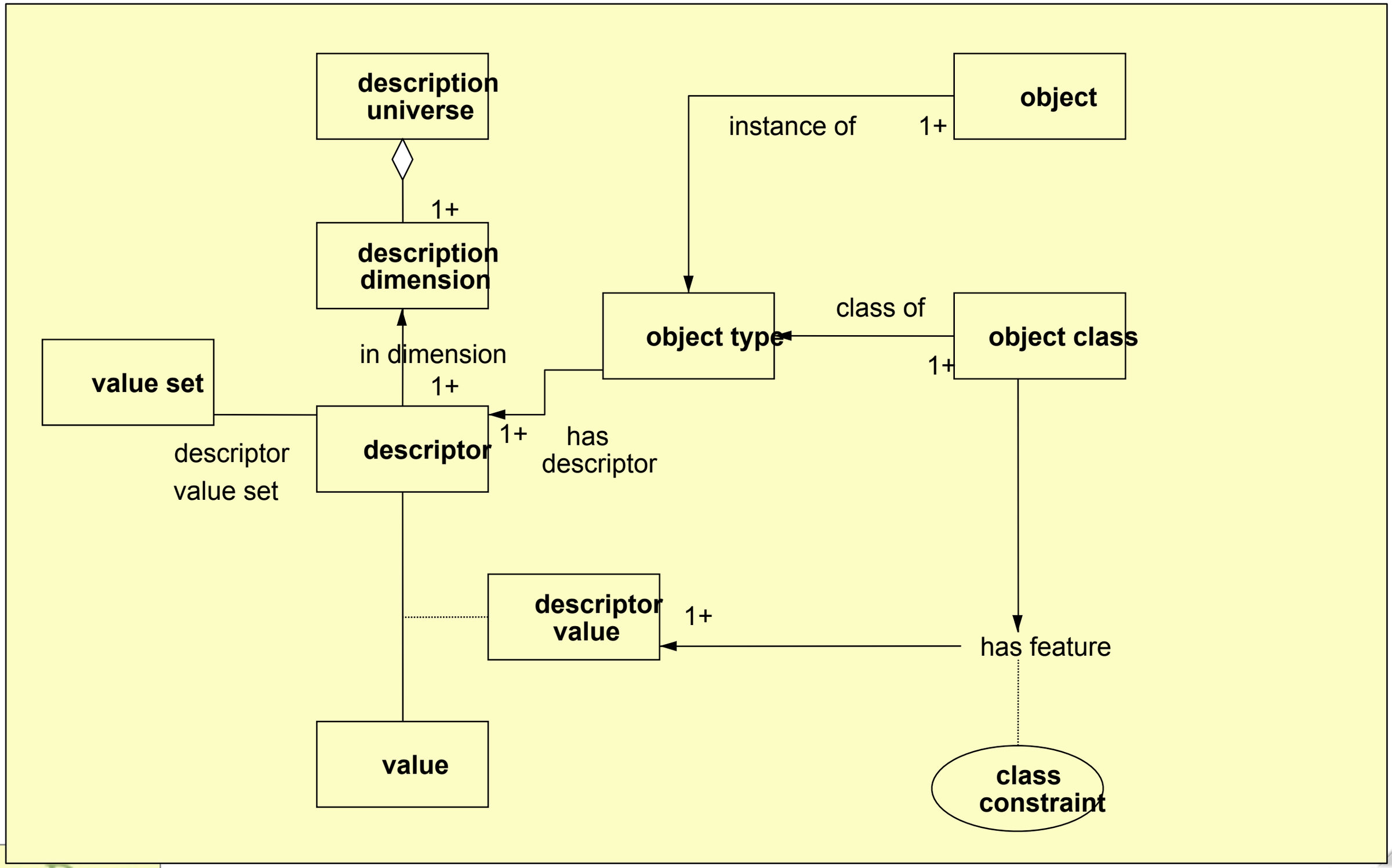
# Art & Architecture Thesaurus

used for indexing stolen art objects in European police databases




[Schreiber 2000]

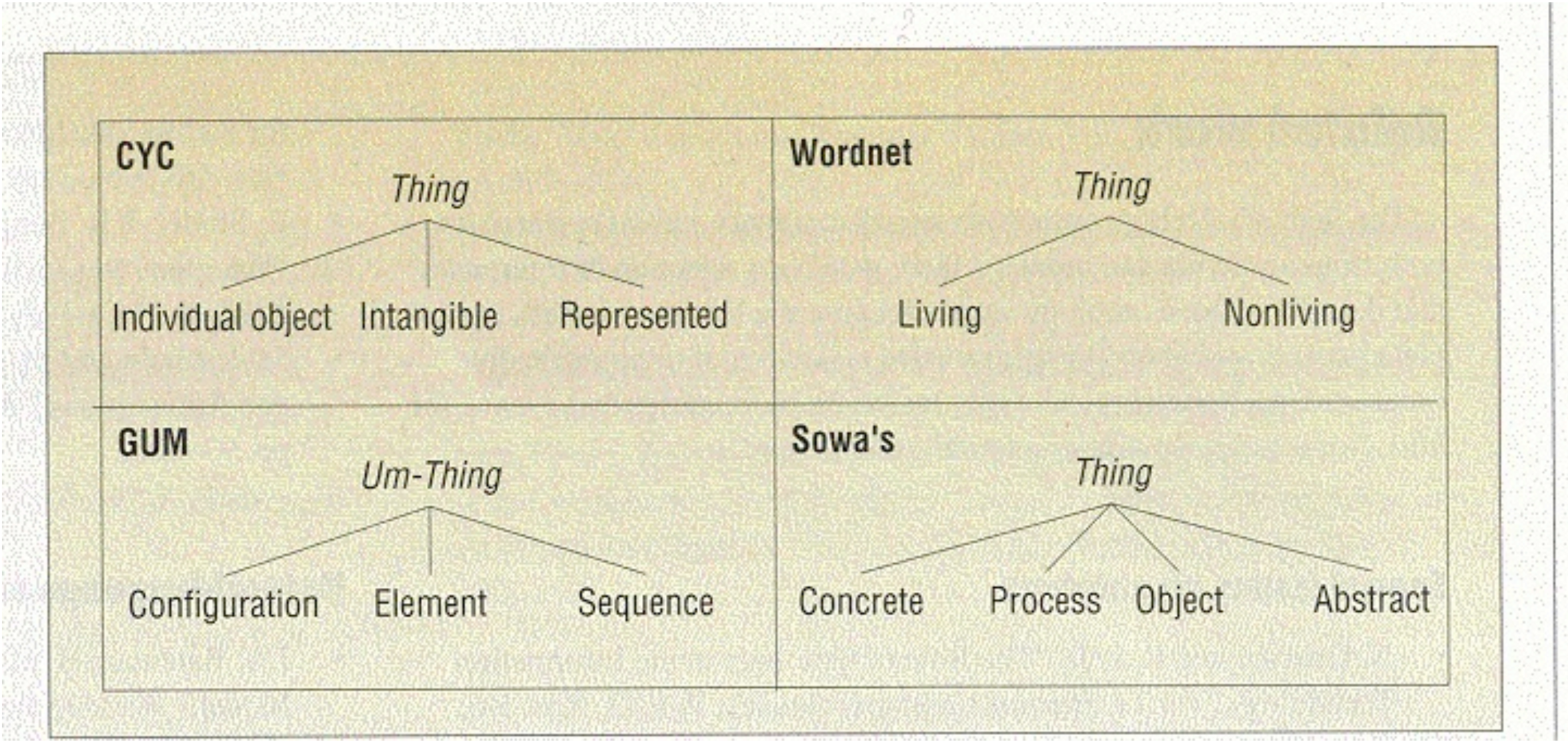
# AAT Ontology






[Schreiber 2000]

# Top-level Categories: Many Different Proposals




Chandrasekaran et al. (1999)

[Schreiber 2000]

# Important Concepts and Terms

automated reasoning
belief network
cognitive science
computer science
deduction
frame
human problem solving
inference
intelligence
knowledge acquisition
knowledge representation
linguistics
logic
machine learning
natural language
ontology
ontological commitment
predicate logic
probabilistic reasoning
propositional logic
psychology
rational agent
rationality
reasoning
rule-based system
semantic network
surrogate
taxonomy
Turing machine

# Summary