Ontology-Supported Web Searching

Vicky Dritsou

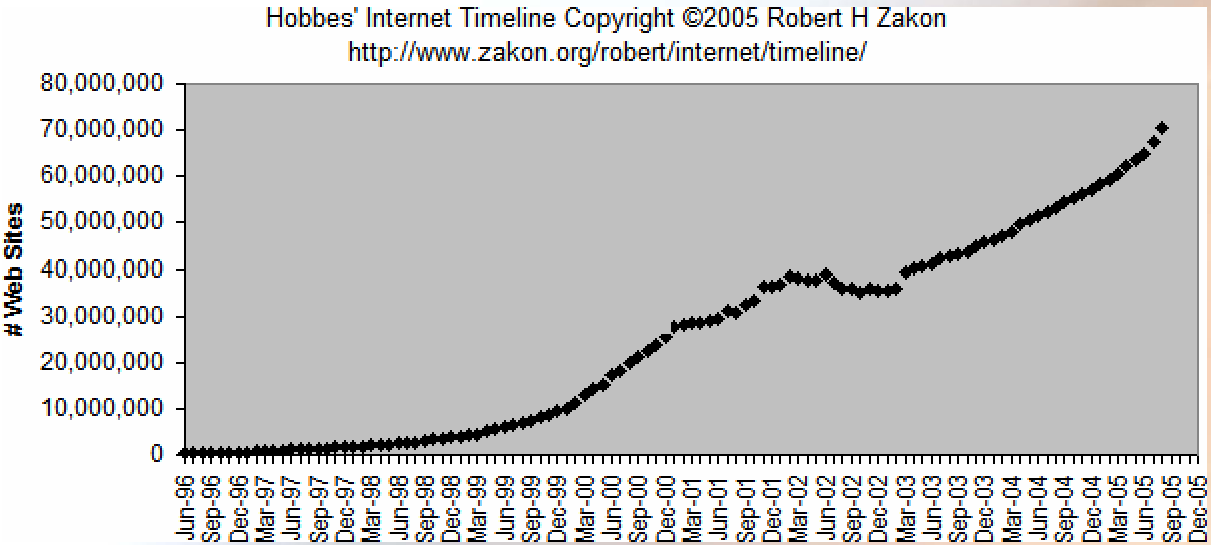
Athens University of Economics and Business
Department of Informatics

Structure of the presentation

- Search engines: current use and problems
- Web catalogues
- Ontologies and classification
- Faceted classification schemes
- Ontology learning
- Ontology mapping
- Logic-based mappings
- Distance-based mappings
- Research plan

Information on the web

- **Est. 695 millions users (comScore Networks Press Release / March 2006[1])**
- **5.48 billions searches during Jan. 2006 only in the US[2]**
- **213 millions searches / day only in the US[3] (March 2006)**
- **Est. 75 millions websites (Nov. 2005 [NIE05])**



[1] http://www.comscore.com/press/release.asp?press=849
[2] http://www.comscore.com/press/release.asp?press=764
[3] http://searchenginewatch.com/reports/article.php/2156461

Problems of search engines

- Users cannot understand the query structure
- Search keywords - very broad terms
  - Both from common users and experts
- Ambiguities of words
- Recall and Precision cannot be calculated, due to web's size
- Users check at most the 20 first results
- No mechanism for narrowing – broadening the search
  - e.g. Google the great!
  - Start a new search: very common
- Refresh rate of indexed sites
  - Pages unavailable or altered
- Cross-lingual retrieval
- No sufficient guidance for the user

Web-catalogues

- Hierarchical classification schemes
  - A priori division of concepts
  - Revision is complex
- Often employ ontologies
- Usually display tree structure
  - Very big size
    - e.g. OpenDirectory: 600.000 categories (terms)
- User's difficulties
  - In understanding the terms – not familiar with the terminology
  - In browsing
- Manual indexing
  - Labour-intensive
  - Results in incomplete – inconsistent indexing
- An example: OpenDirectory [4]
  - 72.500 volunteer editors
  - Appr. 5.500.000 sites indexed

[4] http://www.dmoz.org

Ontologies

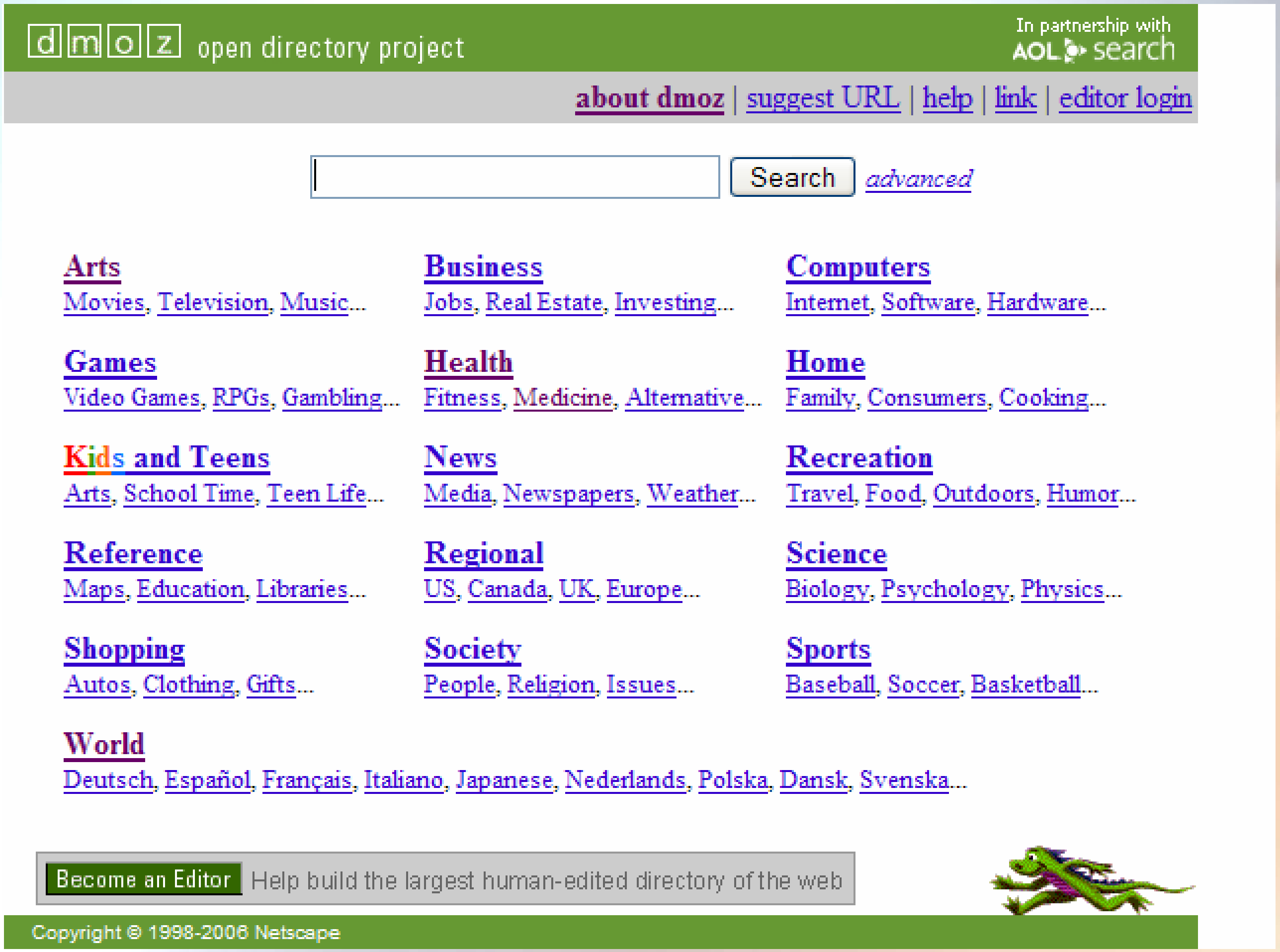
- Ontology is a branch of philosophy
- Ontologies occur in many branches of computer science
- Numerous definitions exist:
  "Ontologies are consensual and formal specifications of a vocabulary used to describe a specific domain." [DEFS99]
  "An ontology is an explicit specification of a conceptualization." [GRU94]
- An ontology consists of:
  - Concepts, relationships, vocabulary
  - Logical propositions
- Usually accepted by a community of experts
- Benefits of using ontologies
  - Give meaning to data
  - Resolve conceptual incompatibilities
  - Deal with synonyms and homonyms
  - Enable communication between different applications

Classification of ontologies

- By subject [GUA97]
  - Generic: classes of general concepts
    - time, event, object
  - Domain: general classes for a specific domain of discourse
    - account, bank-rate
  - Task: general processes of a specific domain
    - deposit, withdrawal, loan
  - Application: specialization to a specific application or group of users
- By complexity
  - Dictionaries: list of terms + their definitions
  - Taxonomies: terminology + subsumption relations
  - Thesauri: terminology + subsumption, equivalence and ‘related terms’ relations (ISO 2788)
  - Conceptual models: terminology + arbitrary semantic relations

Web catalogues: An example

Open Directory Project



Web catalogues: An example (2)

dmoz open directory project

In partnership with AOL search

about dmoz | suggest URL | update listing | become an editor | report abuse/spam | help

Search | the entire directory

Top: Science: Chemistry: Nobel Prize in Chemistry: Nobel Laureates *(31)*

- **Alder, Kurt** *(1)*
- **Corey, Elias James** *(1)*
- **Curie, Marie** *(6)*
- **Molina, Mario J.** *(1)*
- **Pauling, Linus** *(6)*
- **Rowland, F. Sherwood** *(1)*
- **Rutherford, Lord Ernest** *(3)*
- **Seaborg, Glenn T.** *(9)*

See also:

- **Science: Chemistry: History** *(63)*

- Nobel Prize in Chemistry - 1904 - Sir William Ramsay.
- Richard R. Ernst - Nobel Prize in Chemistry 1991
- Roderick MacKinnon - 2003 Nobel Prize in Chemistry awarded for work explaining how a class of proteins helps to generate nerve impulses

---

- "**Nobel Laureates**" search on: AltaVista - A9 - AOL - Clusty - Gigablast - Google - Lycos - MSN - Teoma - WiseNut - Yahoo

---

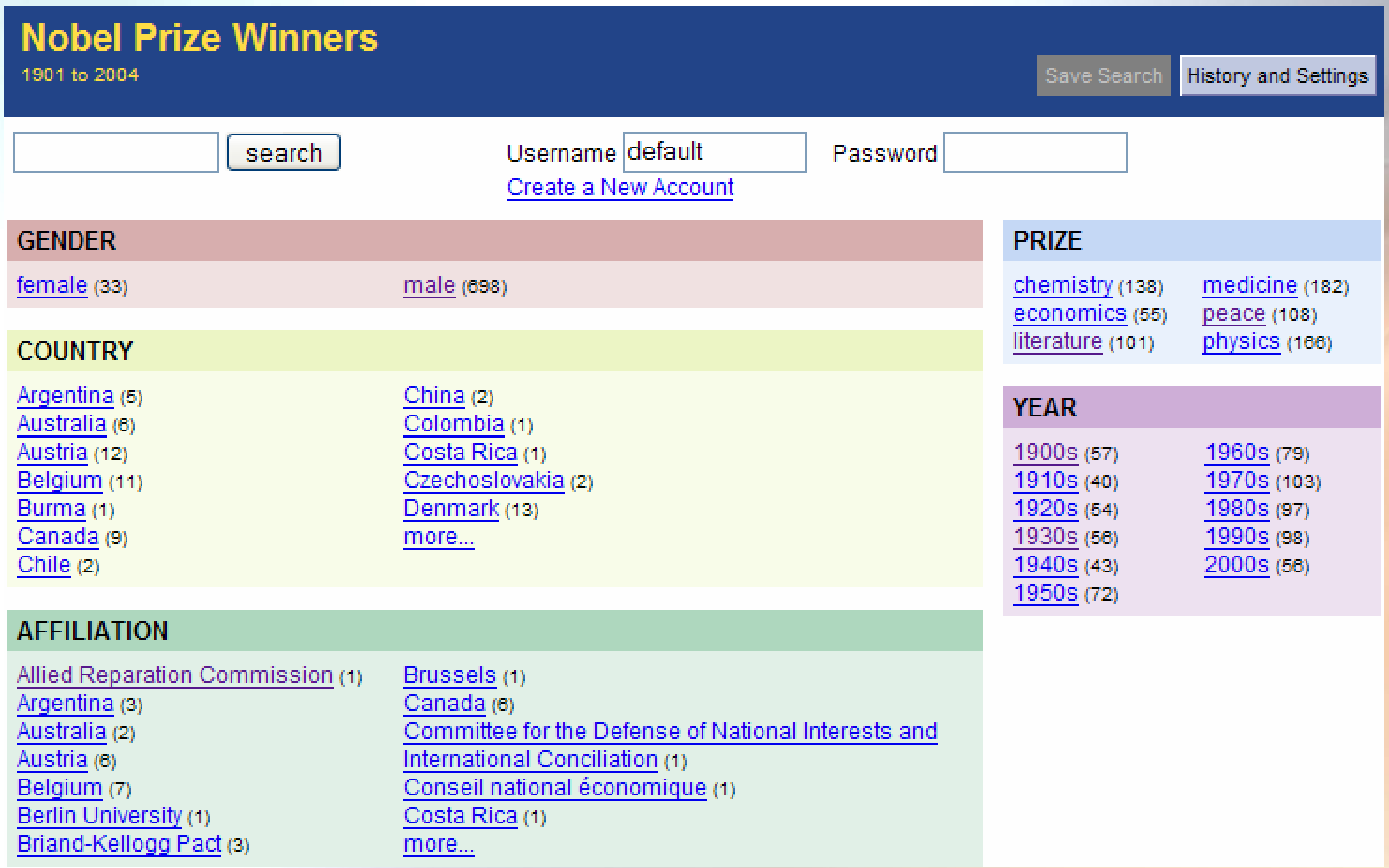
Volunteer to edit this category.

Faceted classification schemes

- Facets: Different aspects of describing concepts
  - Each concept is associated with several terms drawn from different facets
- Structure
  - Compactness
  - Clarity
  - Scalability
  - Independent maintenance of each concept category
  - Combines several classification criteria
- Dynamic creation of compound terms
  - Not all term combinations are valid
    - Combinations that cannot be applied to any of the objects of the domain
    - Algebra for defining the set of valid and non-valid terms [TASC04]
    - Effectiveness not evaluated
- Query reformulation
  - Broader term (query relaxation)
  - Narrower term (query expansion)

Faceted taxonomies: An example

The Flamenco Search Interface Project [5]



[5] http://orange.sims.berkeley.edu/cgi-bin/flamenco.cgi/nobel/Flamenco

Faceted taxonomies: An example (2)

The Flamenco Search Interface Project

search
all items | in current results

Refine your search within these categories:

GENDER (group results)
male (1)

COUNTRY (group results)
Germany (1)

AFFILIATION: all > Germany > Heidelberg > I.G. Farbenindustrie A.G.

PRIZE: all > chemistry

YEAR: all > 1930s > 1931

Recently Viewed Items
Go to Item History

These terms define your current search. Click the × to remove a term.

AFFILIATION: Germany > Heidelberg > I.G. Farbenindustrie A.G. ×

PRIZE: chemistry ×

YEAR: 1930s > 1931 ×

1 result (ungrouped)

Carl Bosch
1874-1940

Ontology learning

- Enrichment of an existing ontology
  - Addition of new concepts
- Many techniques proposed
- Classification
  - By required human interaction
    - Fully-automated
    - Semi-automated
    - Manual
  - By kind of input [GPMM03]
    - Text: natural language analysis
    - Dictionary: extraction of concepts from machine readable dictionary
    - Knowledge base
    - Semi – structured data
    - Relation schemas
- Challenges:
  - Evaluation method
  - Accuracy
  - Human interaction required

Heterogeneity and ontologies

- Syntactic heterogeneity
  - Different interfaces
  - Different data and function representations
  - Can be faced through
    - Platforms
    - Operating systems
- Semantic heterogeneity
  - Terms: different terms describe the same objects
  - Concepts: different conceptualizations of reality
  - Can be faced through
    - Knowledge Representation
    - Ontologies

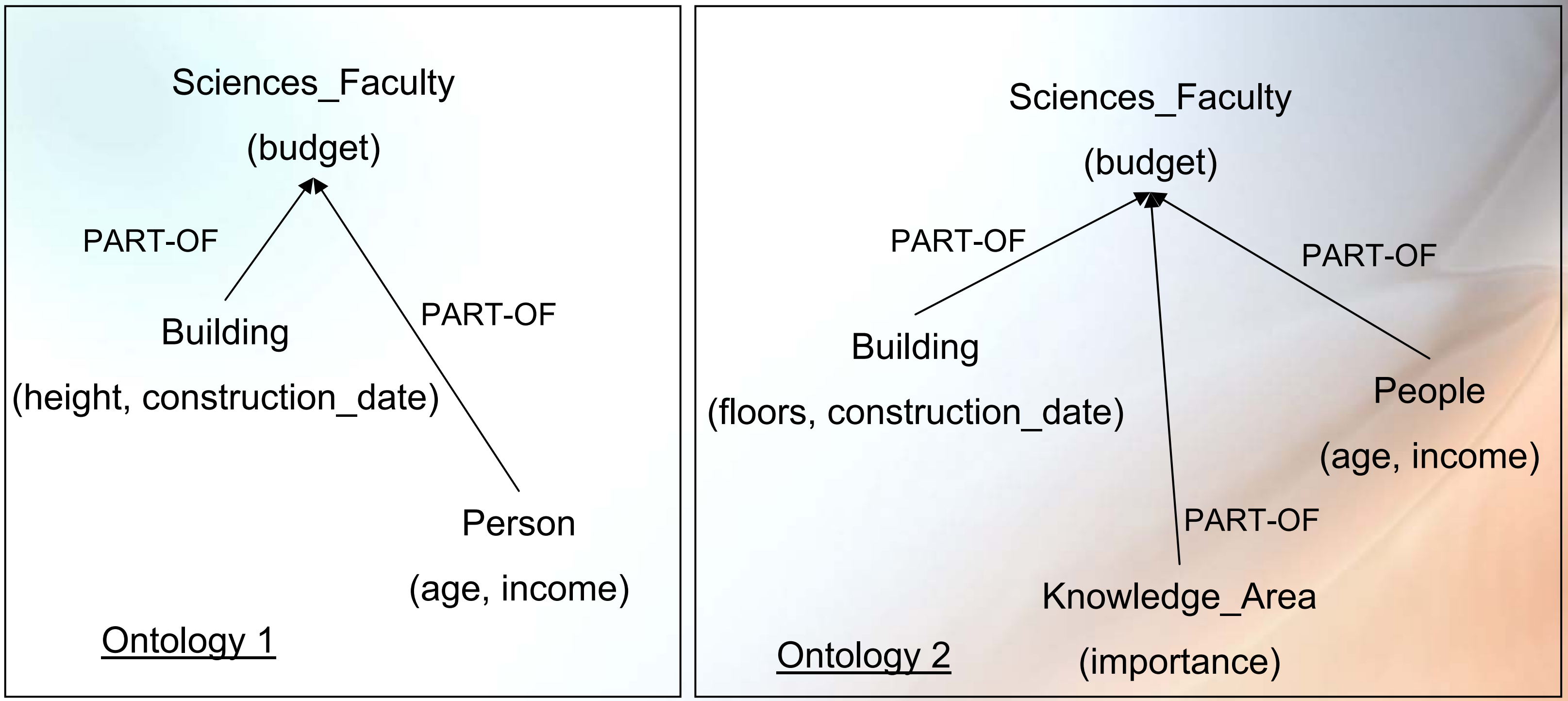
Ontology mapping

- A solution to the heterogeneity of ontologies
- An ontology is considered as a pair of:
  - A signature: the vocabulary
  - A set of axioms (relationships)
- Enables ontologies communication
  - Relates the vocabularies of different ontologies
  - The ontologies describe the same or overlapping domains of discourse
  - All axioms are respected
- Ontology mapping ≠ Ontology translation
  - Mapping involves the definition of collection of functions
  - Translations involves the application of these functions
- Usually not fully automated
  - Human interaction again required
  - Laborious task, when manually constructed

Categorization of logic-based mappings[KASC05]

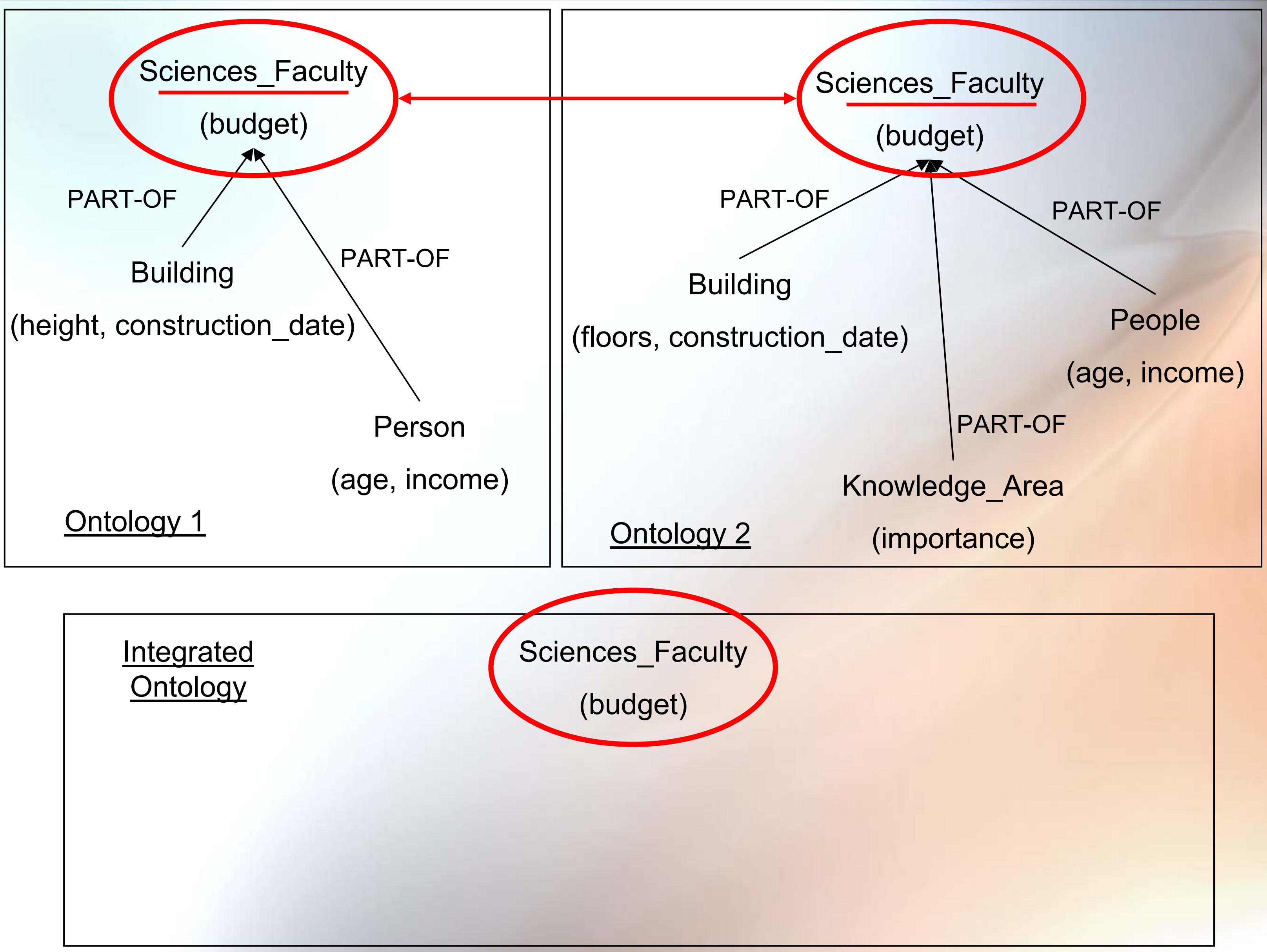
- Ontology signature morphisms
  - Mappings between mathematical structures (functions)
  - Only signatures are respected
    - Loss of information from axioms
- Ontology morphisms
  - Axioms are respected as well
    - More precise and practical
- Ontology alignment
  - Ontologies are mapped with relations, not with functions
    - Collection of binary relations between two vocabularies
- Ontology articulation
  - An intermediate ontology is constructed
  - Each ontology is mapped to the intermediate
- Ontology merging
  - Construction of the union of vocabularies and axioms
  - The union is represented in a new ontology

An example[FBMB02]

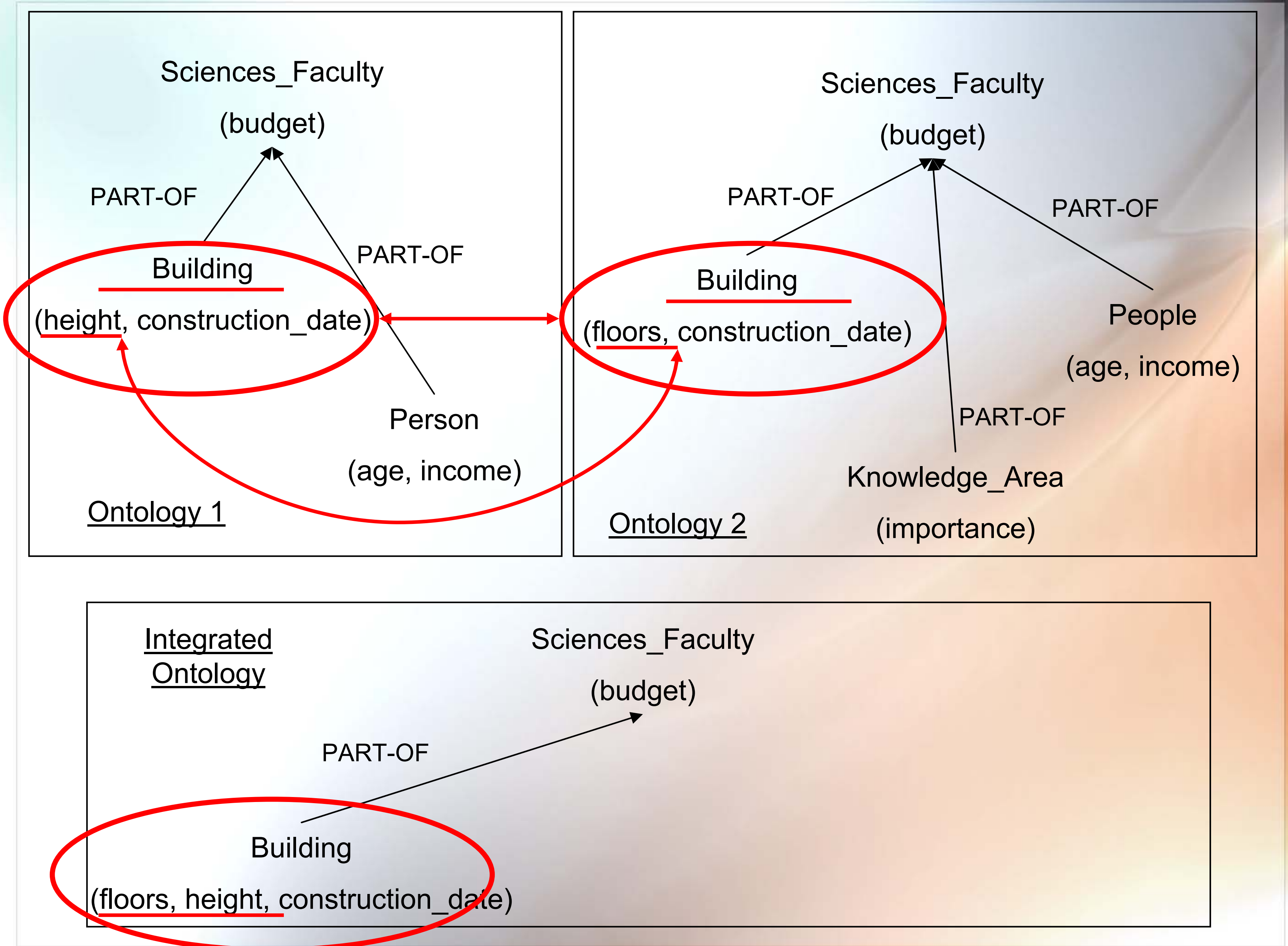


- **Detection of synonymous concepts**
- **Detection of same attributes**
- **Typology of equality criteria for concepts**
  - **For the name-based equality criterion to stand, both concepts must have the same attributes**

An example

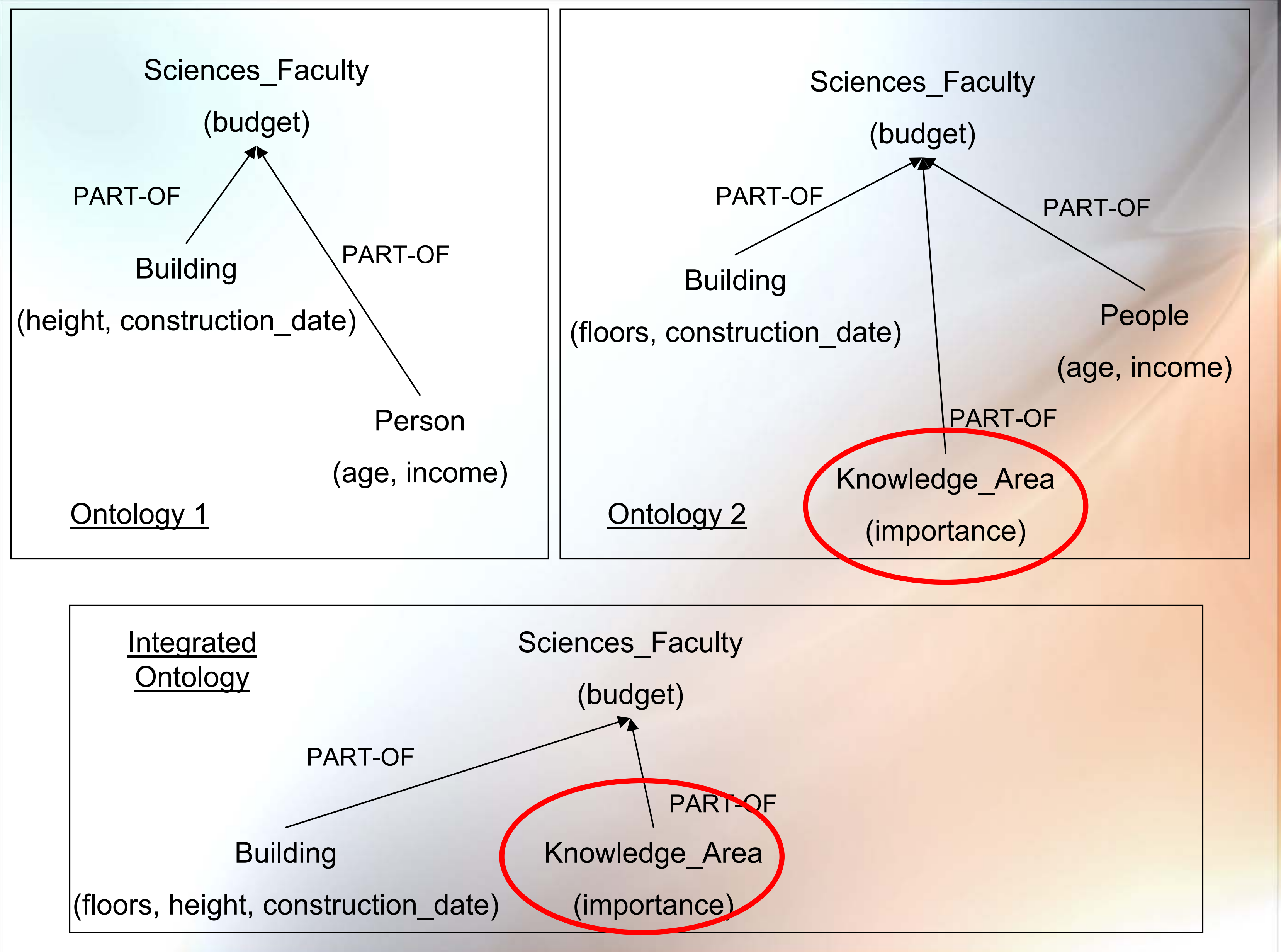

Sciences_Faculty
(budget)
PART-OF
Building
(height, construction_date)
PART-OF
Person
(age, income)
Ontology 1
Sciences_Faculty
(budget)
PART-OF
Building
(floors, construction_date)
PART-OF
People
(age, income)
PART-OF
Knowledge_Area
(importance)
Ontology 2
Integrated Ontology
Sciences_Faculty
(budget)

An example

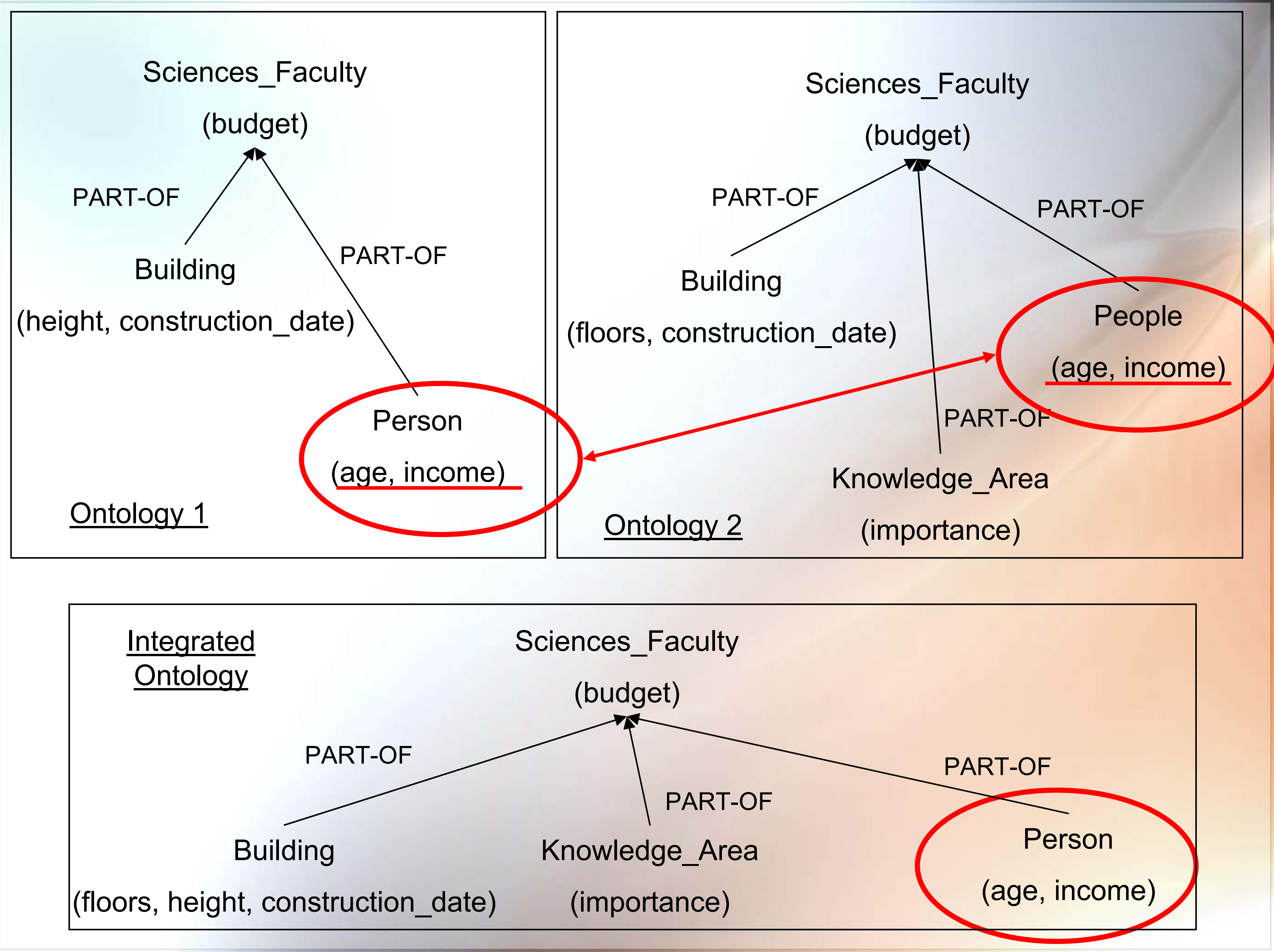


Sciences_Faculty
(budget)
PART-OF
PART-OF
Building
(height, construction_date)
Person
(age, income)
Ontology 1
Sciences_Faculty
(budget)
PART-OF
PART-OF
Building
(floors, construction_date)
People
(age, income)
PART-OF
Knowledge_Area
(importance)
Ontology 2
Integrated Ontology
Sciences_Faculty
(budget)
PART-OF
Building
(floors, height, construction_date)

An example


Sciences_Faculty
(budget)
PART-OF
Building
(height, construction_date)
PART-OF
Person
(age, income)
Ontology 1
Sciences_Faculty
(budget)
PART-OF
Building
(floors, construction_date)
PART-OF
People
(age, income)
PART-OF
Knowledge_Area
(importance)
Ontology 2
Integrated Ontology
Sciences_Faculty
(budget)
PART-OF
Building
(floors, height, construction_date)
PART-OF
Knowledge_Area
(importance)

An example


Sciences_Faculty
(budget)
PART-OF
PART-OF
Building
(height, construction_date)
Person
(age, income)
Ontology 1
Sciences_Faculty
(budget)
PART-OF
PART-OF
Building
(floors, construction_date)
People
(age, income)
PART-OF
Knowledge_Area
(importance)
Ontology 2
Integrated Ontology
Sciences_Faculty
(budget)
PART-OF
PART-OF
PART-OF
Building
(floors, height, construction_date)
Knowledge_Area
(importance)
Person
(age, income)

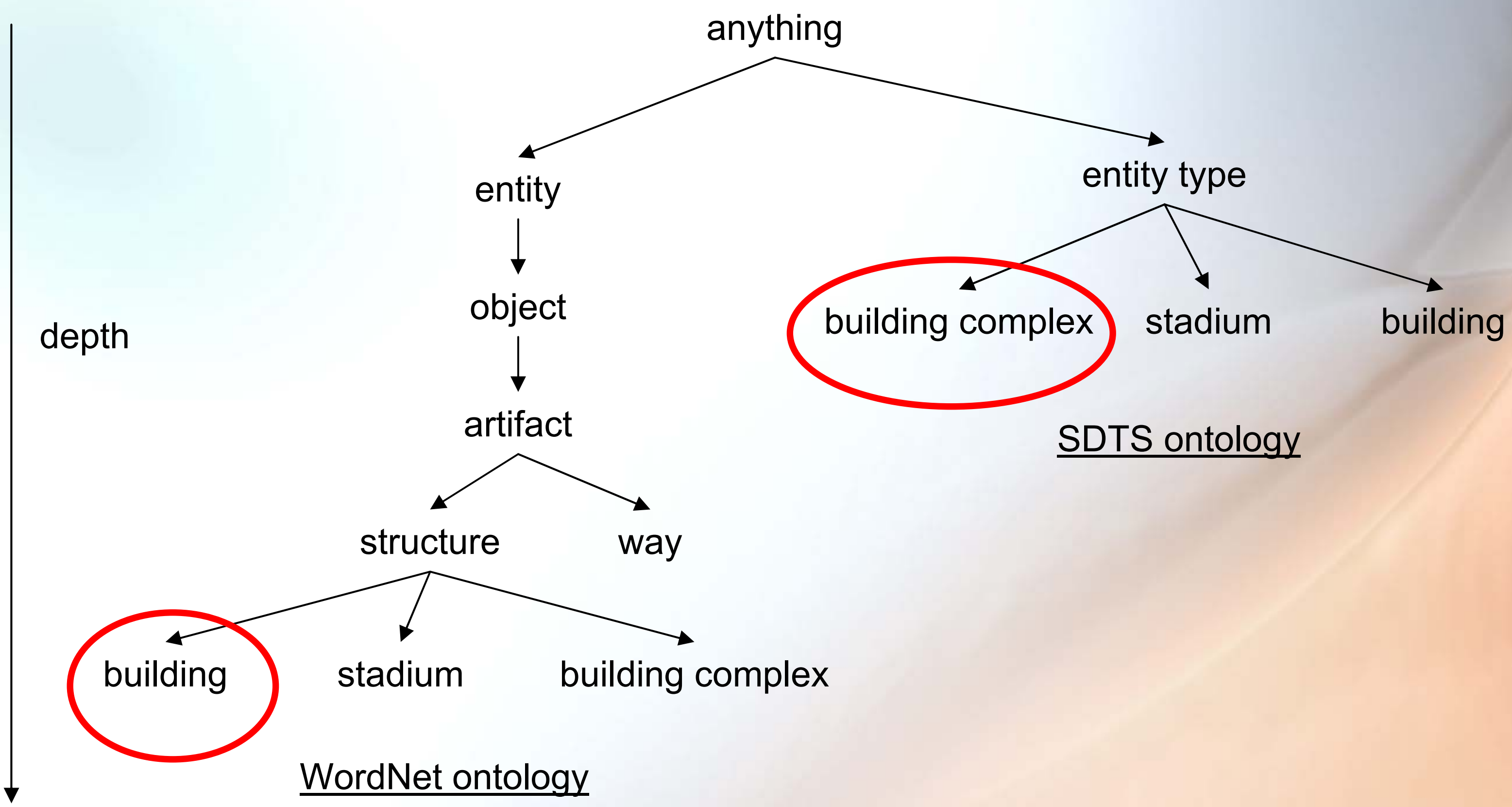
Distance-based mappings

- They are based on conceptual distances
  - Use of computational model
  - Utilization of metrics
  - Detection of analogies between concepts
  - Usually defined as the shortest path through a common ancestor
  - Applied even to ontologies from different domains, as long as they share a common higher level domain
- Defining conceptual distances between objects[SC96]:
  - Based on the names of concepts
  - Based on non-common classes
  - Based on non-common super-classes
  - Based on non-common attributes
- They depend on the design of the ontology
- Semantic links are not always of equal weights
- Evaluation methods need to be defined

An example[RE03]

- An approach applied to two different ontologies
- Counts similarities between all pairs of terms between the two ontologies
- The results are compared, in order to discover the most similar concepts
- Three similarity functions
  - Word matching
  - Feature matching
  - Semantic-neighborhood matching

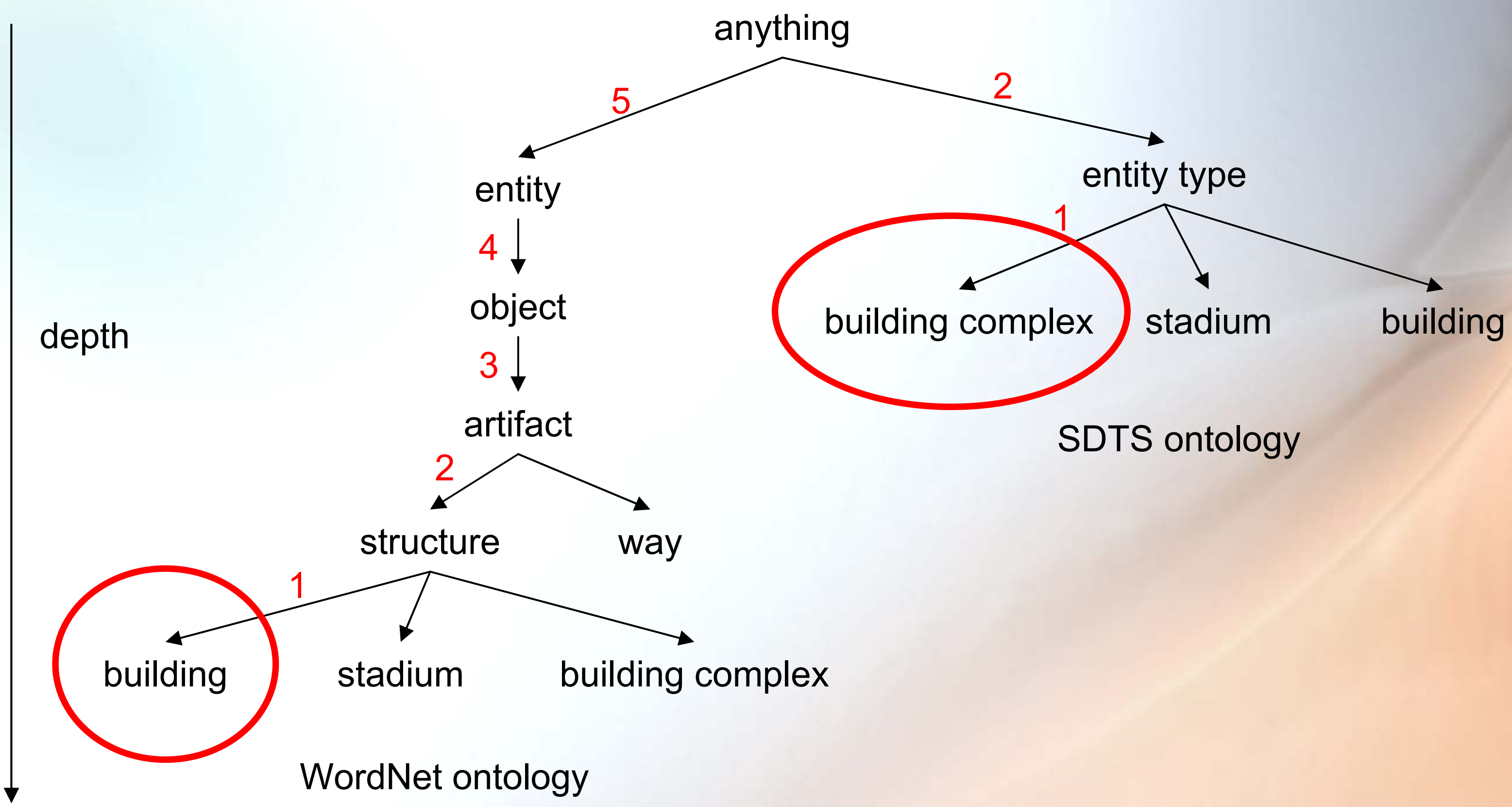
Word Matching



$$\mathbf{a}\,(\text{building, building complex}) = 1 - \frac{\text{depth(building)}}{\text{depth(building)} + \text{depth(building complex)}}$$

$$\mathbf{S}\,(\text{building, building complex}) = \frac{|\{\text{building}\}|}{|\{\text{building}\}| + 0.28|\{\}| + 0.72|\{\text{complex}\}|}$$

Word Matching



$$\mathbf{a}\,(\text{building, building complex}) = 1 - \frac{\text{depth(building)}}{\text{depth(building)} + \text{depth(building complex)}} = 0.28$$

$$\mathbf{S}\,(\text{building, building complex}) = \frac{|\{\text{building}\}|}{|\{\text{building}\}| + 0.28|\{\}| + 0.72|\{\text{complex}\}|} = 0.58$$

Feature Matching

Stadium (SDTS) Definition

entity_class {
name: {stadium, bowl, arena}
is_a: {construction}
part_of: {}
whole_of: {athletic_field}
parts: { {athletic_field, sports_field, playing_field}
{dressing_room}, {foundation},
{midfield}, {spectator_stands, stands},
{ticket_office, box_office, ticket_booth} }
functions: { {play, compete}, {play, practice},
{recreate, play} }
attributes: { {architectural_property},
{covered/uncovered}, {name}
{lighted/unlighted}, {owner_type},
{sports_type}, {user_type} } }

Stadium (WordNet) Definition

entity_class {
name: {stadium, bowl, arena}
is_a: {construction}
part_of: {}
whole_of: {athletic_field, sports_arena}
parts: { {athletic_field, sports_field, playing_field}
{foundation}, {midfield}, {plate},
{sports_arena, field_house}, {stands},
{structural_elements}, {standing_room},
{tiered_seats} }
functions: {}
attributes: {} }

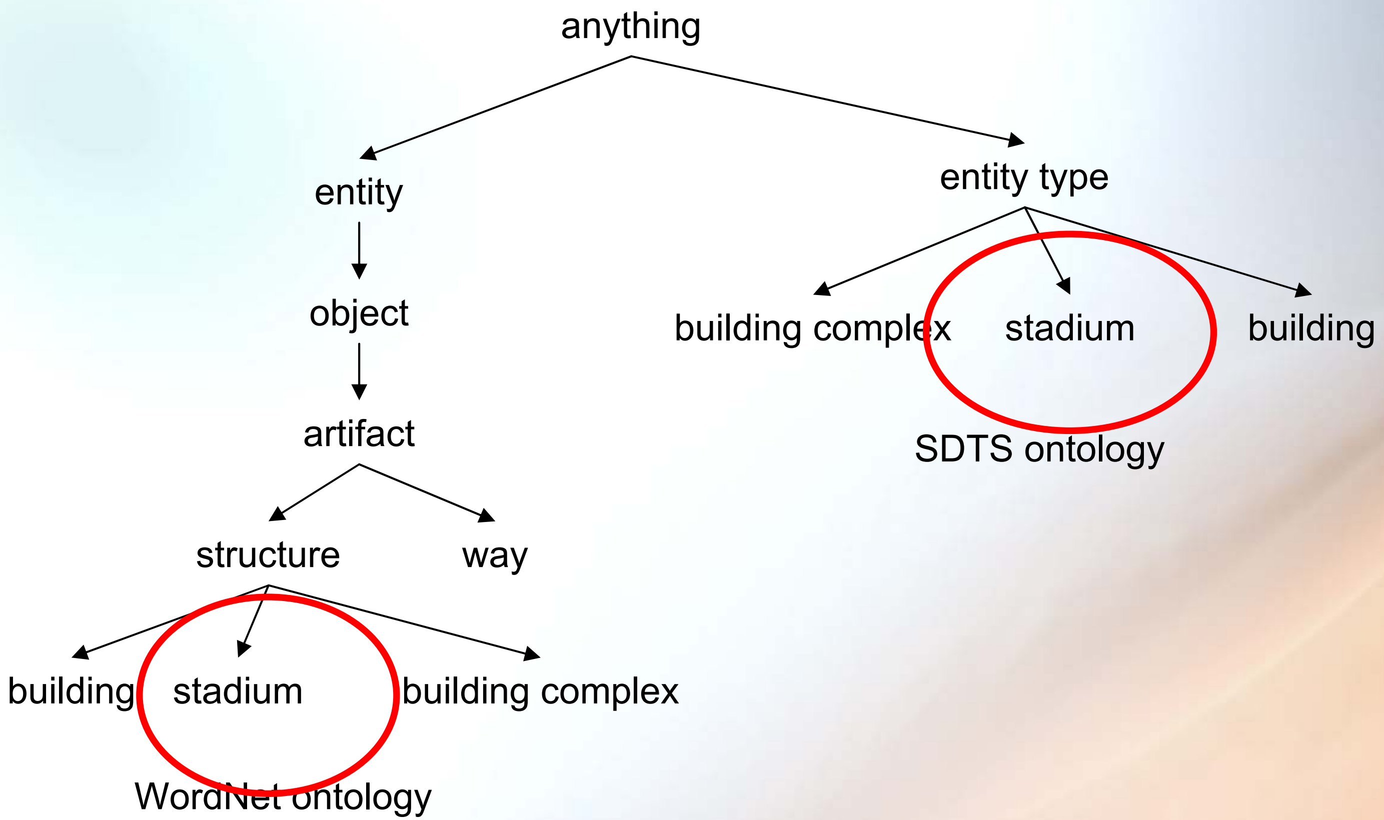
$$X = stadium^{sdts}.parts \cap stadium^{w}.parts$$

$$Y = stadium^{w}.parts - stadium^{sdts}.parts$$

$$Z = stadium^{sdts}.parts - stadium^{w}.parts$$

$$\mathbf{S(stadium^{w}, stadium^{sdts}) = \frac{|X|}{|X| + a\,|Y| + (1-a)|Z|} = 0.54}$$

Semantic-Neighborhood Matching



- **We find the immediate neighbors of the entities**

N $(\text{stadium}^w, 1)$ = {stadium, structure}

N $(\text{stadium}^s, 1)$ = {stadium, entity_type}

- **For each entity in N(stadium^w, 1), we find the most similar one in N(stadium^s, 1)**

Research plan

- Context: faceted taxonomies
- Issues of faceted taxonomies maintenance
  - Especially of the compound terms
    - Effective modification, e.g. term enhancement, deletion
  - Internal reformation of a facet
  - General reformation of facets
- Combined use of distance-based and logic-based approaches of mapping
  - Determination of the combination from which we can benefit most
- Implementation of a system in a search engine
  - Experimental testing of correctness and effectiveness
  - Examination of its influence in information retrieval

Challenges

- Parallelization of the crawling process
  - It should be properly coordinated
  - Facets could be used for determining thematic divisions of the web
    - We could benefit from the terms (meaning) each URL contains – and also from the URLs neighbors
    - Appropriate metadata could be used, that would denote in which facet each site is classified
- The crawler must select the "important" pages to download
  - Textual similarity is mostly used
  - We can define similarity as the distance between the term entered by the user and those of the ontology
- Finding related pages - queries
  - By using synonyms or similar terms
  - When query by example is used, we could define the facet under which it is instantiated

Thank you for your attention!

Vicky Dritsou (vdritsou@aueb.gr)
Information Systems and Databases Laboratory
Department of Informatics
Athens University of Economics and Business, Greece

References

[DEFS99] Decker, S., Erdmann, M., Fensel, D., Studer, R. (1999). Ontobroker: Ontology based Access to Distributed and Semi-Structured Information. In *Semantic Issues in Multimedia Systems.* Kluwer Academic Publisher, 1999.

[FBMB02] Fernández-Breis, J., Martínez-Béjar, R. (2002). A cooperative framework for integrating ontologies. *International Journal of Human – Computer Studies,* 56:665-720, 2002.

[GPMM03] Gómez-Pérez, A., Manzano-Macho, M. (2003). OntoWeb Deliverable 1.5[online]. http://ontoWeb.org/About/Deliverables/deliverable_view [Accessed 18 March 2006].

[GRU94] Gruber, T. (1994). Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In *Formal Ontology in Conceptual Analysis and Knowledge Representation.* Kluwer Academic Publishers, 1994.

[GUA97] Guarino, N. (1997). Understanding, Building, and Using Ontologies: A Commentary to "Using Explicit Ontologies in KBS Development", by van Heijst, Schreiber, and Wielinga. *International Journal of Human and Computer Studies*, 46, 1997, pp. 293-310.

[KASC05] Kalfoglou, Y., Schorlemmer, M. (2005). Ontology mapping: the state of the art. *Dagstuhl Seminar Proceeding*, 04391 IBFI, Schloss Dagstuhl, Germany 2005.

[NIE05] Nielsen, J. (2005). Usability for the Masses. *Journal of Usability Studies* 1, 1 (Nov), 2-3.

[SC97] Spanoudakis, G., Constantopoulos, P. (1996). Elaborating Analogies from Conceptual Models. *International Journal of Intelligent Systems*, Vol. 11, No. 11, pp. 917-974, Wiley, 1996.

[RE03] Rodríguez, A., Egenhofer, M. (2003). Determining Semantic Similarity Among Entity Classes from Different Ontologies. *IEEE Transactions on Knowledge and Data Engineering,* 15 (2):442-456.

[TASC04] Tzitzikas, Y., Analyti, A., Spyratos, N., Constantopoulos, P. (2004). An Algebraic Approach for Specifying Compound Terms in Faceted Taxonomies. *In Proceedings of 13th European-Japanese Conference on Information Modelling and Knowledge Bases, EJC2003,* Kitakyushu, Japan, June 2003, and in book: *Information Modelling and knowledge Bases XV, IOS Press,* 2004, pp. 67-87.