SSSW 2005 Tutorial

Ontology Evaluation and Validation

Aldo Gangemi

Laboratory for Applied Ontology, ISTC-CNR, Roma
aldo.gangemi@istc.cnr.it
http://www.istc.cnr.it/createhtml.php?nbr=71

thanks to:
Carola Catenacci, Massi Ciaramita, Rosa Gil, Nicola Guarino, Jos Lehmann

Diagrams in this presentation

- UML class diagrams with OWL(DL) semantics:

- uml:Class --- owl:Class
- uml:Association --- owl:ObjectProperty
- uml:Generalization --- rdfs:subClassOf
- no cardinality --- ≥0
- 1...* cardinality --- ≥1
- 1 cardinality --- =1
- uml:AssociationClass --- owl:Class, or swrl:Rule

Recap on ontology engineering techniques

- **Creation** of datamodels, e.g. pre-designed patterns for SWS (*push*)
- **Extraction** of patterns by *machine learning* and *NLP* (*pull*)
- **Reuse** of existing/legacy ontologies (*pull/push*)
  - *Simple reuse (pull/push)*
  - *Standards* for definite domains (*push*)
  - *Building blocks* (*pull/push*)
- **Re-engineering** of metadata (*pull*)
- **Emergence from communities of interest** (*pull/push*)

Main needs

- There is a need for - *a priori* or *a posteriori*:
  - **Modularity**: is there a viable design methodology for ontology architectures?
  - **Reusability**: are there viable ontological components to reuse (e.g. as building blocks)?
  - **Quality**: what expertise is addressed, and to what extent?
  - **Selection**: how can ontologies survive and reproduce?

Outline of the tutorial *

- A communication perspective on ontologies
- Structural measures (evaluation)
- Functional measures (evaluation)
- Usability measures (evaluation)
- Principles, parameters, and preferences for quality (validation)

- * Non-addressed in this tutorial:
  - design-time vs reuse time, and related methodologies
  - tools and ontology realization

A communication perspective on ontologies

What is an ontology - socio-cognitive cut

- An ontology is (<u>extension</u> of Gruber 1993, Guarino 1998):
  - A
  - Formal,
  - Partial Specification of the
  - Conceptualization of a world
  - Conceived by some
  - Rational agent for some (good or bad)
  - Reason, and made in order to
  - Negotiate that conceptualization with
  - Someone else, or to
  - Reuse it.

Logic
Representation
Meaning
Cognition
Embodiment
Motivation
Agreement
Society
Culture

- In other words it is:
  - *a descriptive specification of a set of contextual assumptions about a world (≈ a typed logical theory?)*
- While it is not:
  - *a prescriptive specification of the inner structure of 'true reality' (unless the conceptualization aims at this status)*

Ontologies? Resources to be reengineered?

CYC

theory of flogistum

mereology

conversational maxims

“how to make a coffee” for dummies

WordNet

DOLCE

Allen’s event calculus

Italian Constitution

RCC-8

the modal jazz harmony

UMLS Metathesaurus

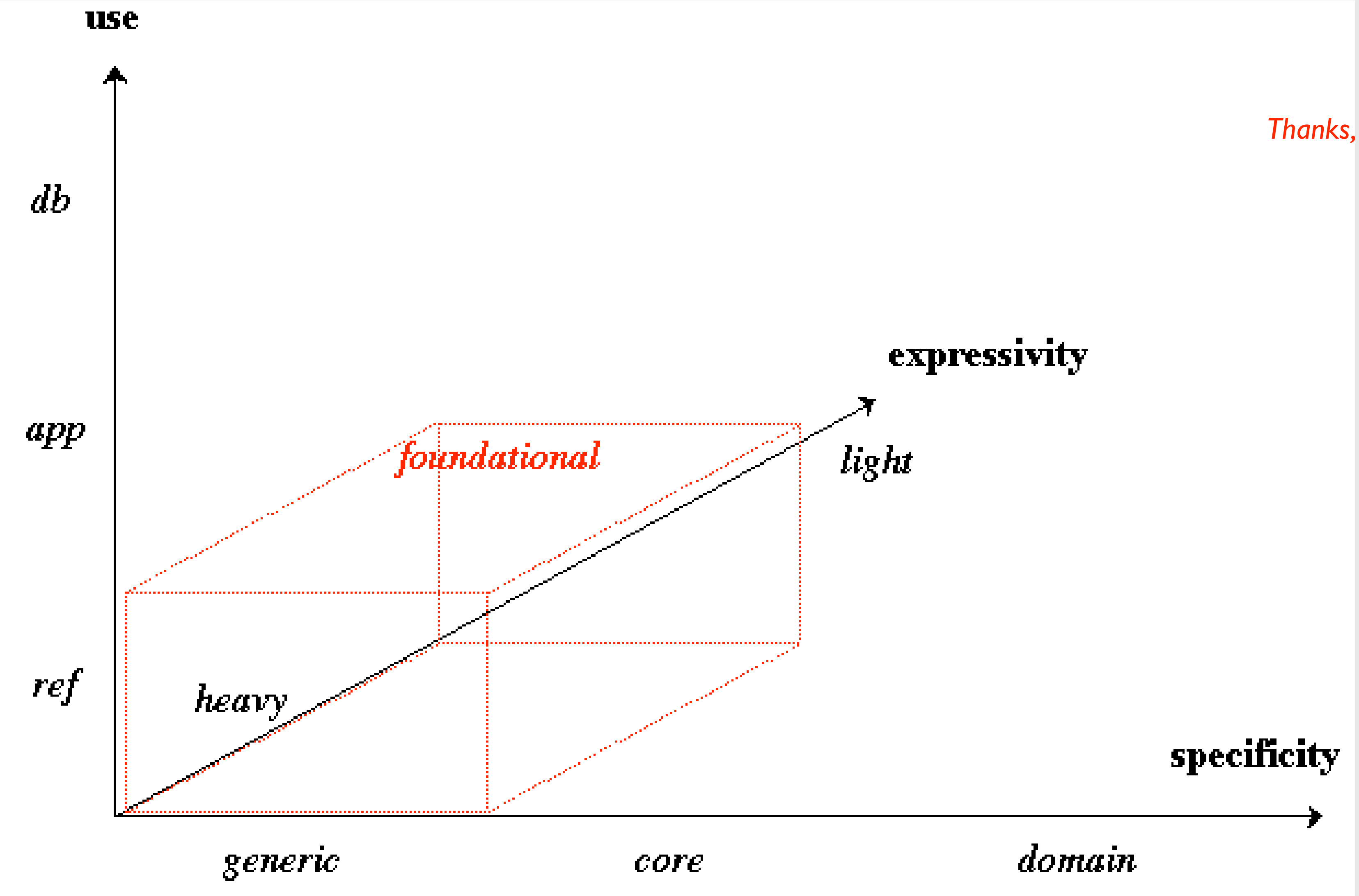
railway timetable

Roget’s thesaurus

AAT thesaurus

the glossary of nurses

An early model of ontology dimensions



Thanks, Daniel!

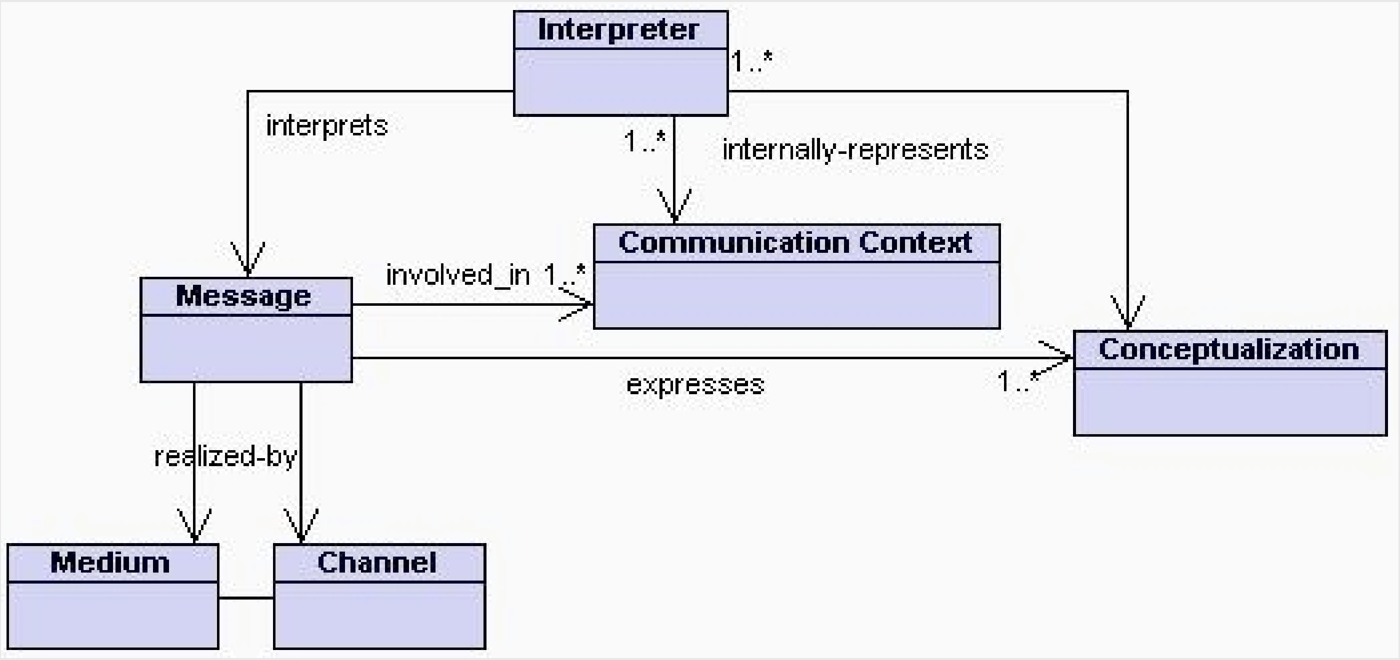
In the real world

- Real developers and (re)users of ontologies result to have many more dimensions to consider: graph properties, logical validity, accuracy, recognition, usability, economical efficiency, modularization, ...
- Which theory can consider them all? and interrelate them all?
- An abstraction step is needed

A different perspective: the communication roles of ontologies

Users

Information

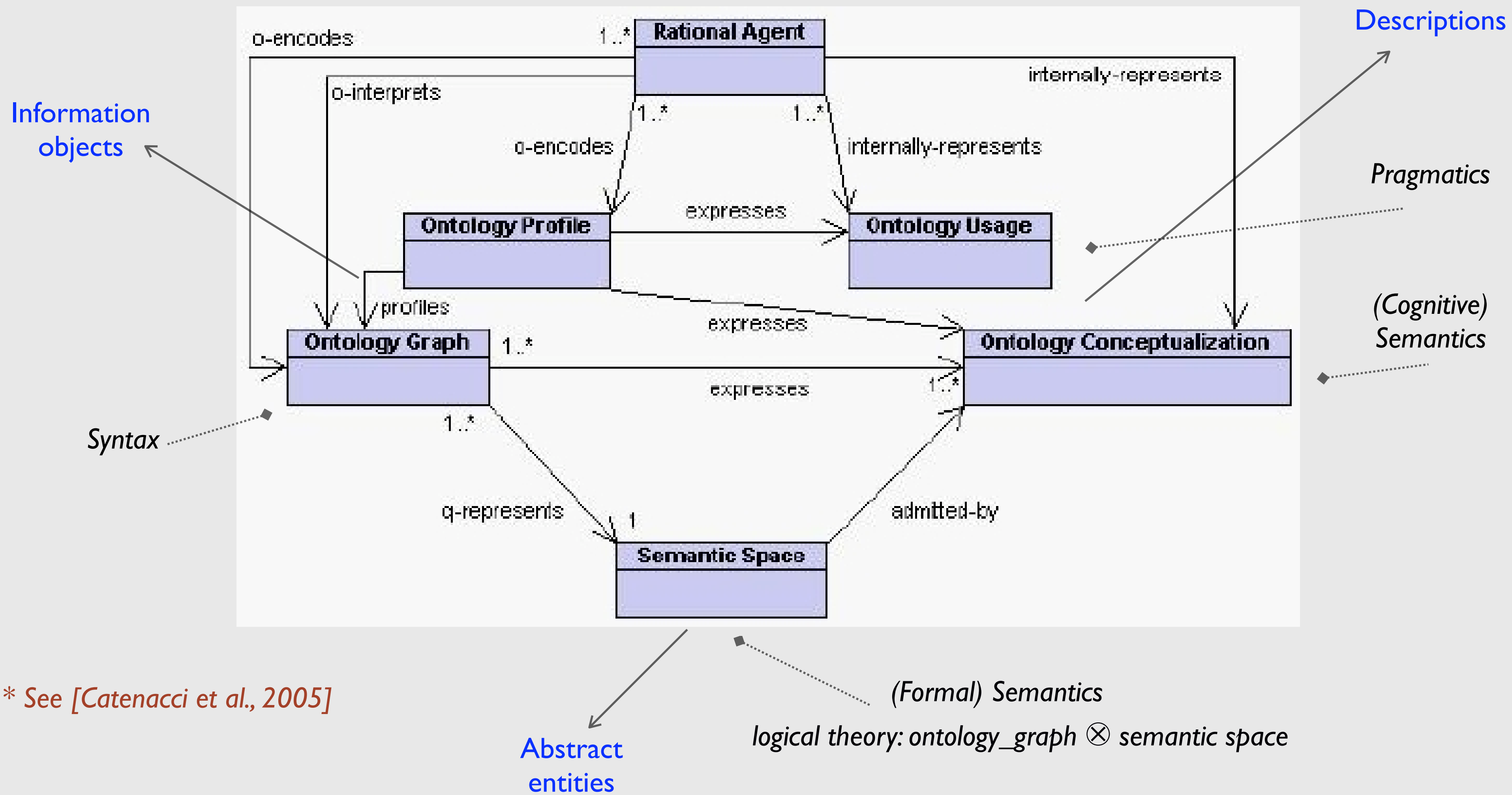


Expertise

Tools

Cf. [Jakobson 1960], [Gangemi et al. 2004]

Ontologies as semiotic objects: the O^2 pattern



* See [Catenacci et al., 2005]

The three semiotic realms of evaluation

- Syntax and formal semantics (*structural evaluation*): graph- and logic-based measures
- Syntax-semantics matching (*functional evaluation*): precision/recall-based measures
- Pragmatics (*usability evaluation*): recognition, (economic) efficiency, and interfacing measures

Quality assessment

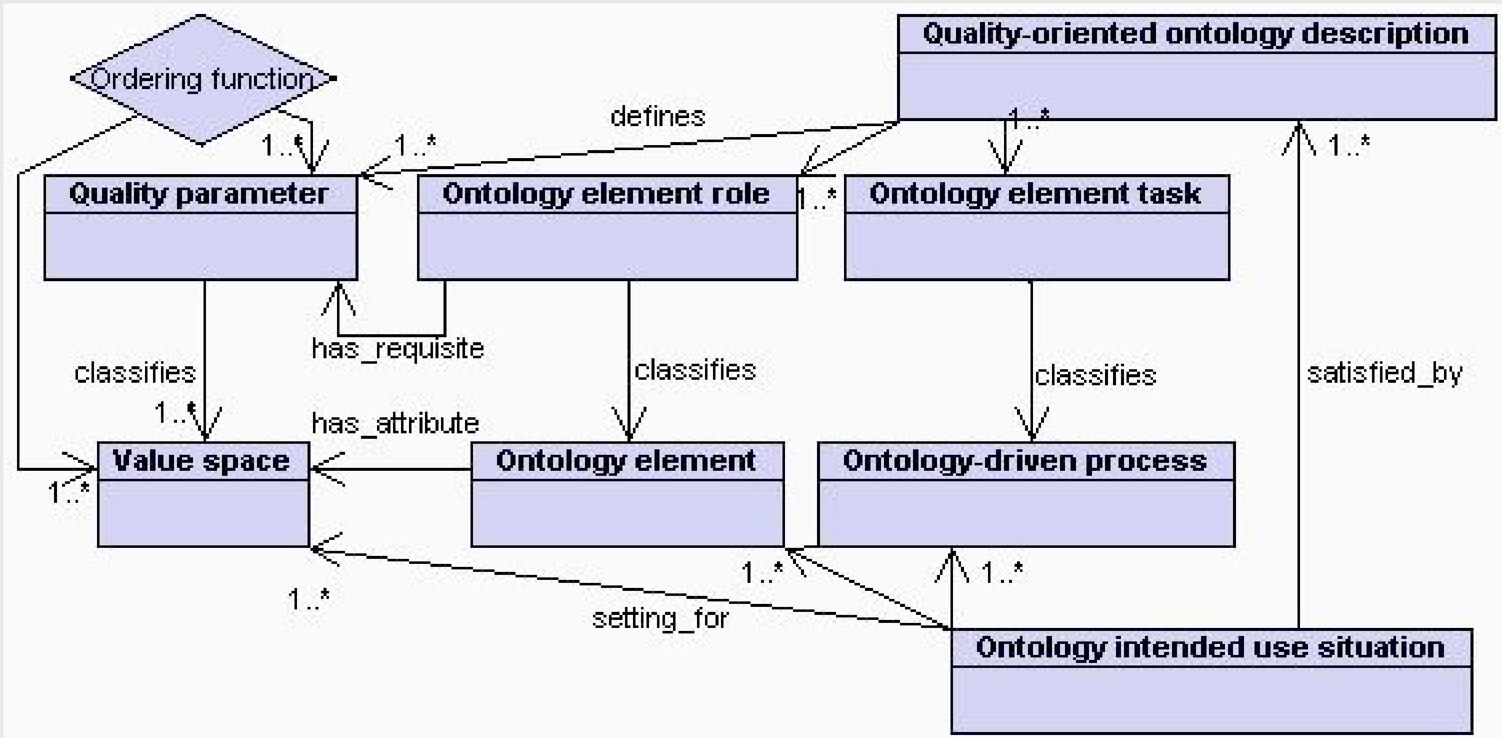
- It is a diagnostic task performed over ontologies
- Needs principles and parameters based on those principles
- Parameters combine in non-trivial ways
- Badness or goodness can change: preferential ordering over parameters (cf. OntoMetric) to represent trade-offs
- oqυαl: a formal characterization of quality assessment (based on the *Description<->Situation* ontology design pattern, cf. [Gangemi 2005])

The oquαl pattern: a formal characterization of quality assessment

E.g. *fitness to competency question*

E.g. *relation*

E.g. *retrieve*



E.g. *project: haemocancer ontology*

E.g. *use case: haemocancer assessment*

An ontology value space is any attribute of an ontology element that has been obtained by means of some measurement procedure
E.g. *the relation fits an expertise' competency question: 'I want to know the family history for condition type c'*

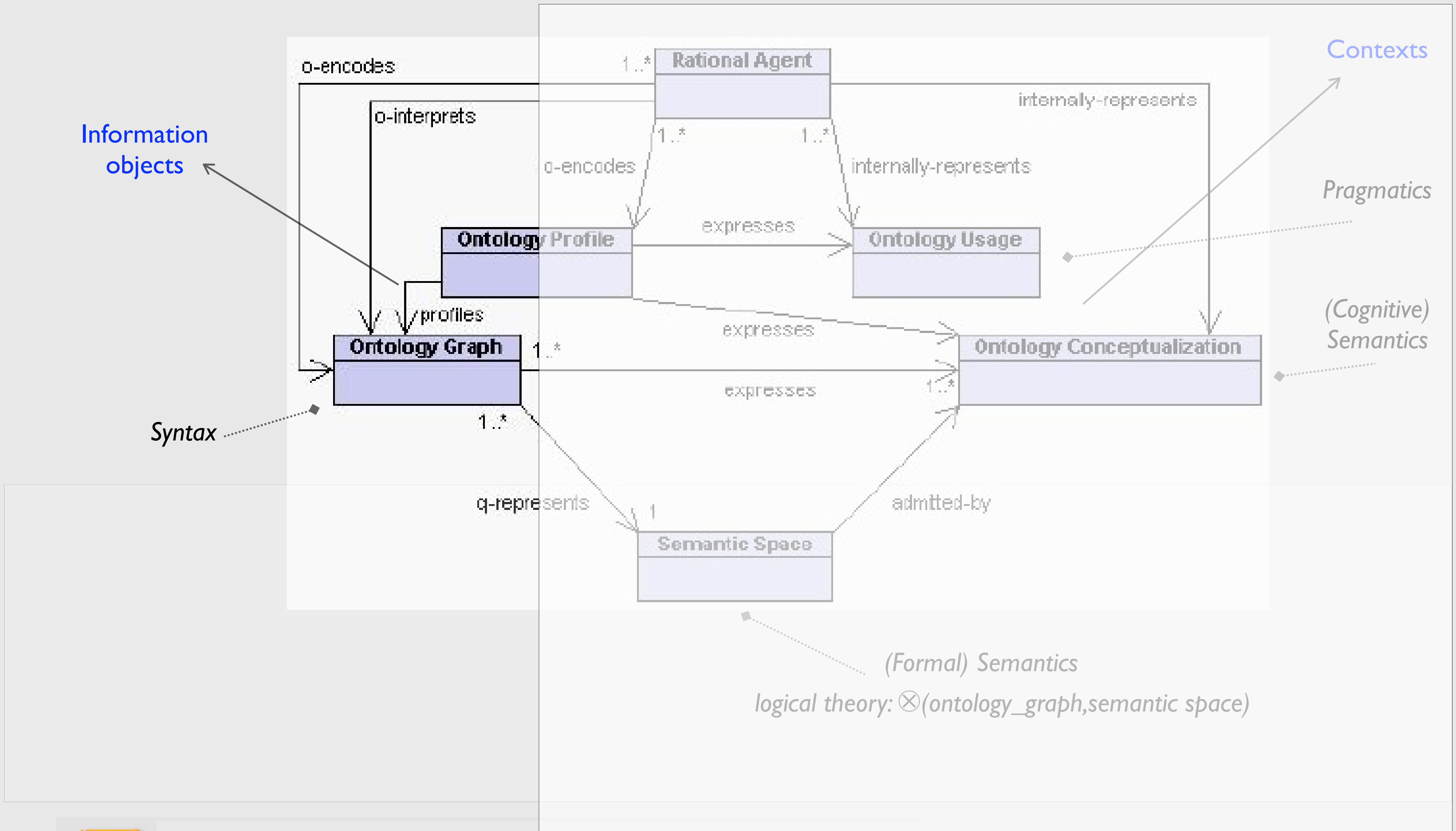
Ontology elements can be: a class, a property, an individual, a module, an annotation, etc.
E.g. *a relation among patient, family, condition, indicator*

Ontology-driven processes make use of an ontology element.
E.g. *retrieve (p,f,c,i)*

See also slide 62

Evaluation - 1
Structural evaluation -
Dimensions, elements, measurement methods

A semiotic perspective: the syntactic aspect



Ontology (structural) measure definition

A structural ontology measure is defined according to the template:

$$M = \langle D, S, mp, c \rangle$$

D=Dimension is the graph property to measure: the intensional counterpart of the metric space

S=Set of graph elements includes the elements of the ontology graph

mp=Measurement procedure is the procedure executed to perform the measurement. The procedure can be a function (as counting) or an algorithm.

c=Coefficient of measurement error adjusts for context-related variations on measurement procedure.

A measure *m* is a real number obtained by applying a measurement procedure mp for a dimension *D* to a set *S* of graph elements, modulo a coefficient *c* (if any):

$$mp_{D,c,S} \xrightarrow[yields]{} m \in \Re$$

Types of structural measures (graph-based)

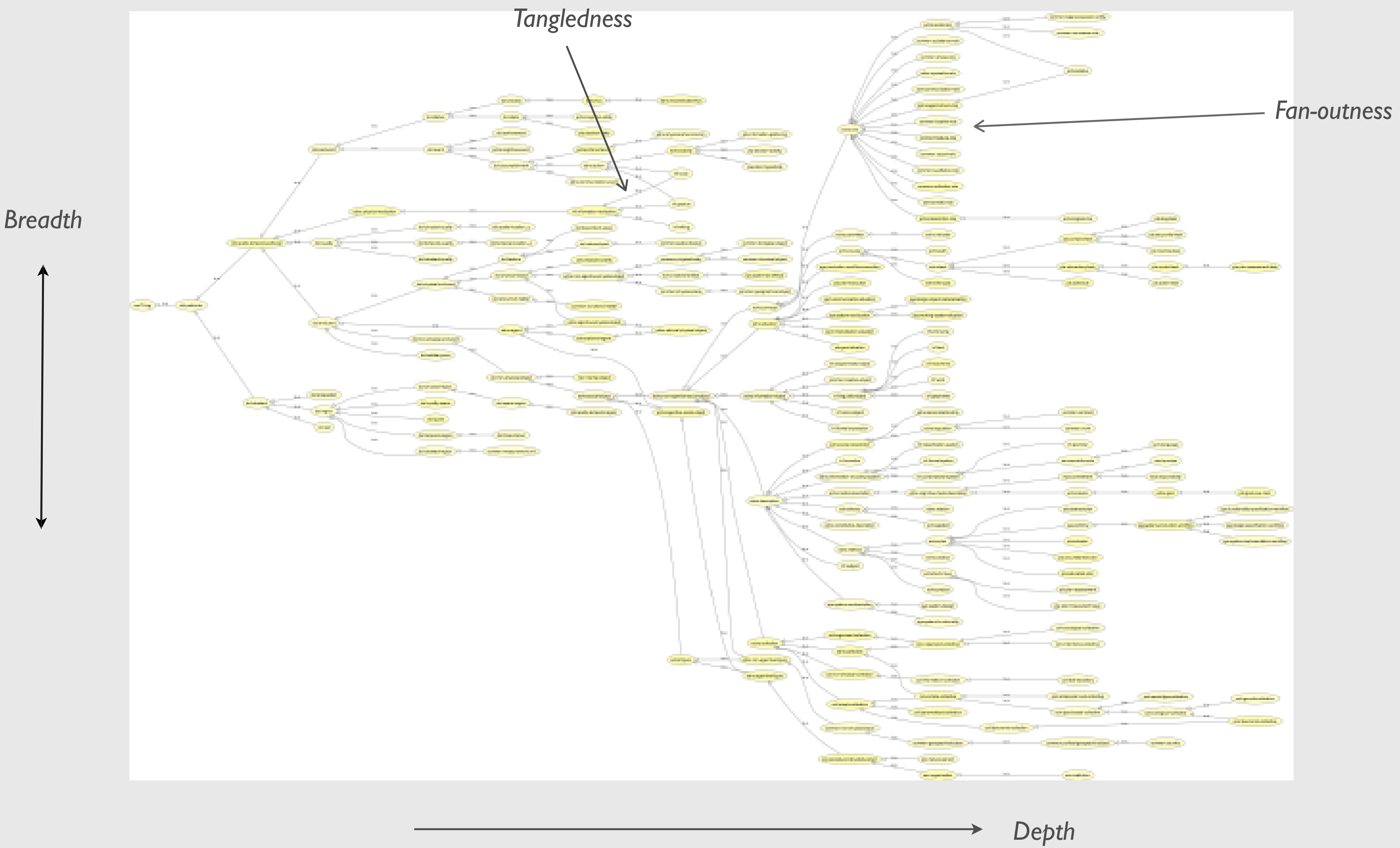
- Depth: absolute, maximal, average, e.g.
  - Average depth: $$m = \frac{1}{n_{P \subseteq g}} \sum_{j}^{P} N_{j \in P}$$
    - where $N_{j \in P}$ is the cardinality of each path *j* from the set of paths *P* in a graph *g*, and $n_{P \subseteq g}$ is the cardinality of *P*.
- Breadth: absolute, maximal, average

See also [Yao et al. 2005]

- Tangledness
  - Tangledness: $$m = \frac{n_G}{t_{\in G \wedge \exists a_1, a_2 (isa(m, a_1) \wedge (isa(m, a_2))}}$$
    - where n_G is the cardinality of the set *G* of graph nodes, and *t* is the cardinality of the set of nodes with more than one ingoing *isa* arc in *g*.
- Fan-outness: absolute, maximal, average, focused on siblings or on leaf-nodes; e.g.
  - Average sibling fan-outness: $$m = \frac{\sum_{j}^{SIB} N_{j \in SIB}}{n_G}$$

where $\sum_{j}^{SIB} N_{j \in SIB}$ is the absolute sibling set cardinality for the digraph *g*

Measure types on an ontology graph (DOLCE-Lite-Plus v.397)



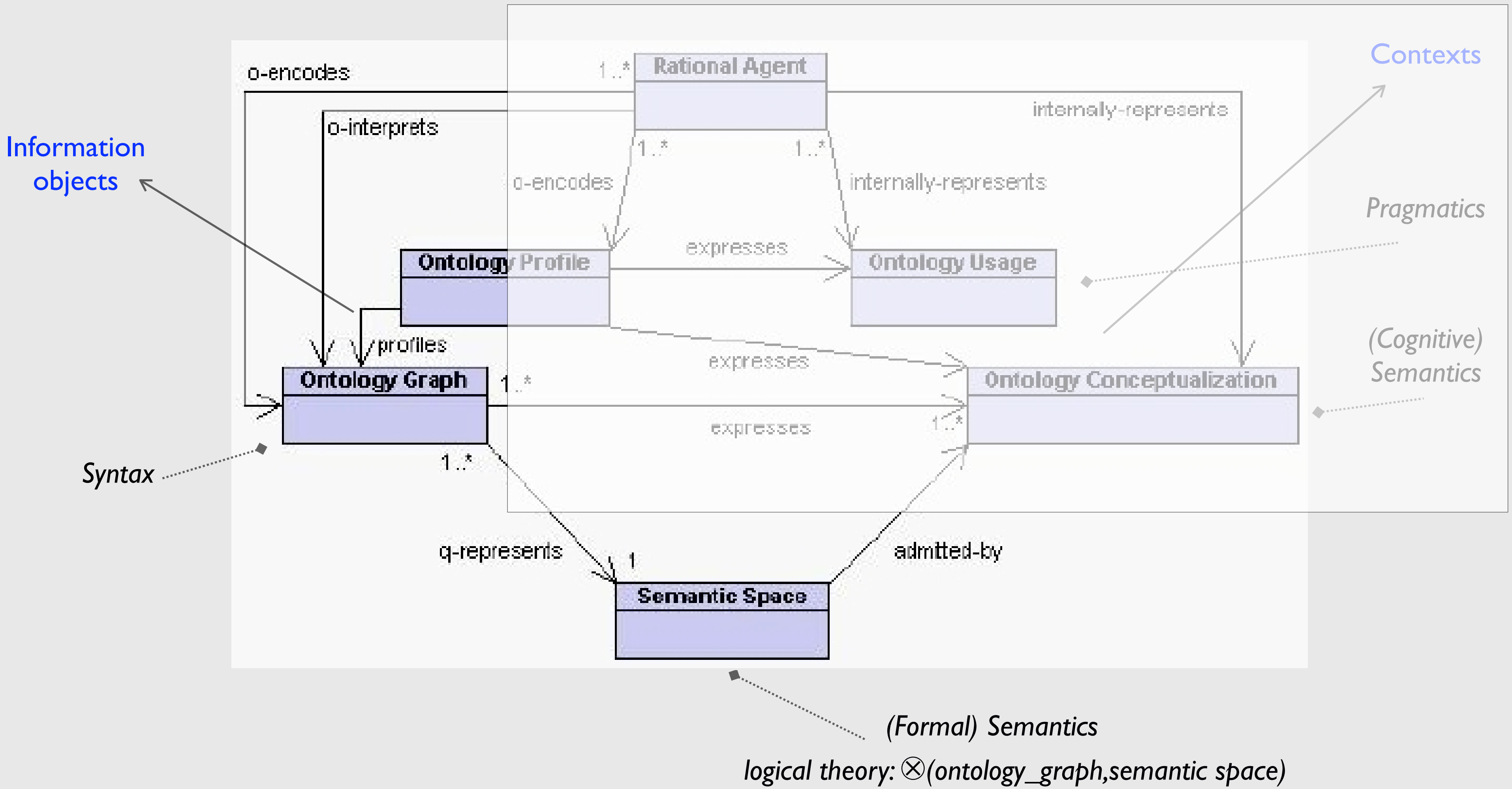
(cont'd)

- Specific criterion: relative, absolute, e.g.
  - Ratio of sibling sets featuring a shared specific difference among elements (e.g. *Object:agentive/non-agentive*):

$$m = \frac{n_{SIB(DF)}}{n_{SIB}}$$

  - where $n_{SIB(DF)}$ is the cardinality of the set *SIB(DF)* including only the sibling sets whose elements share a specific difference. More precisely, an element $x \in j$ (a sibling set from *SIB(DF)*) must have a same non-taxonomical relation ρ holding for different values from a same class φ; formally, for a sibling set *j*: $\forall x \in j \ \exists \rho, \varphi(\rho(x,y) \wedge \varphi(y))$. n_{SIB} is the cardinality of the set *SIB* for *g*.
- Density: presence of clusters of classes with many non-taxonomical relations holding among them (wrt to overall ontology graph)
  - usu. *core ontology patterns* (e.g. thematic roles, contracts, diagnoses)
- Modularity: absolute, partitioning, module overlapping, e.g.
  - Modularity rate: $m = \frac{n_M}{n_S}$
  - where n_M is the cardinality of *M* (the set of asserted modules), and n_S is the cardinality of *S* (the set of graph elements).

A semiotic perspective: the formal aspect



Types of structural measures (logic-based)

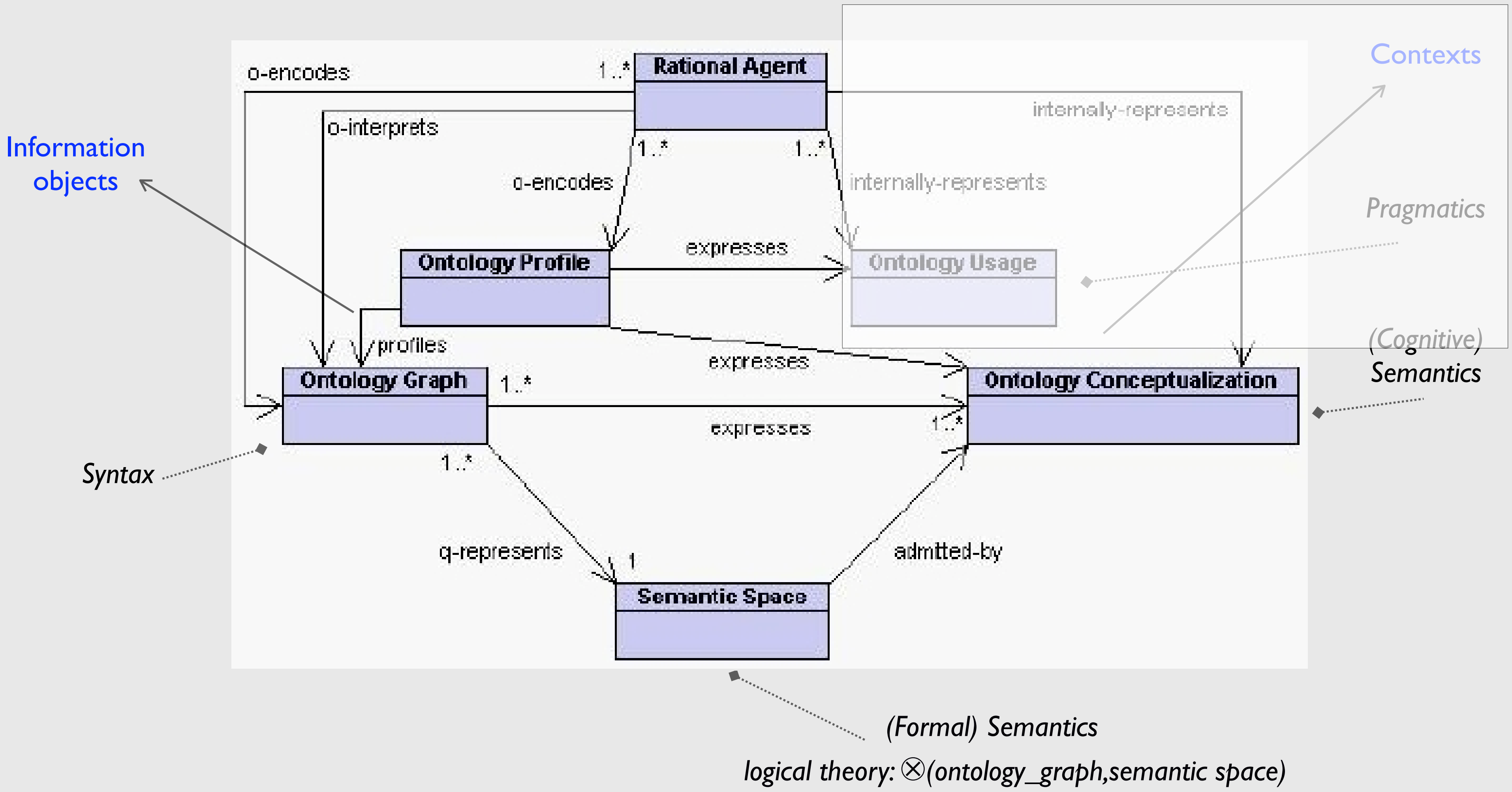
- Logical adequacy: consistency, complexity, anonymous classes, disjointness ratio, cycles, Class/Property/Restriction/Individual ratio
  - see also Sean Bechhofer's presentation in SSSW 2005
  - Consistency: proportion of incoherent concepts
  - Complexity: types of logical constructs used
  - Anonymous classes: proportion of unnamed logical constructs used by a classifier
  - Disjointness: proportion of mutually disjoint classes
  - Cycles: proportion of cyclical paths over the set of paths for *g*
  - Ontology element distribution: Class/Property ratio, etc.

(cont'd)

- Meta-logical adequacy: Guarino and Welty's OntoClean
  - classes classified by their *rigidity, unity, dependence* properties
  - attempt to reduce some functional measures to measurement of adequacy wrt series of possible worlds (states of affairs)
    - stability of a property across a series of temporal states for a same entity (<u>rigidity</u>) e.g. *person* vs. *student*
    - disjointness of sets of properties across a series of (different) topological states for a same entity or cluster (<u>unity</u>) e.g. *dog* vs. *rubbish*
    - stability of a property across a series of states featuring different relational properties for a same entity (<u>dependence</u>) e.g. *dog* vs. *dogtail*

Evaluation - 2
Functional evaluation -
Coverage, Precision and Accuracy

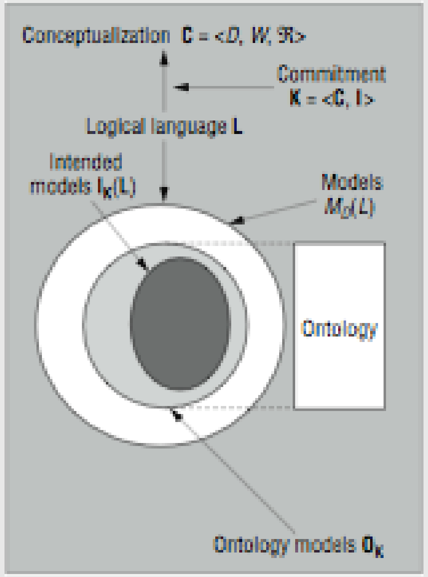
A semiotic perspective: the cognitive aspect



Functional evaluation

- An ontology graph and its formal interpretation must match an intended conceptualization [notice: design and reuse time have own issues]
- In the semiotic view of ontologies, a (syntactic) graph expresses a context-bound intended meaning
- The semantic interpretation of a graph should fit that meaning
- The pragmatics of an ontology is mostly encoded in extrinsic codes (annotations, profiles)

A formal model of functional evaluation [Guarino 2004]

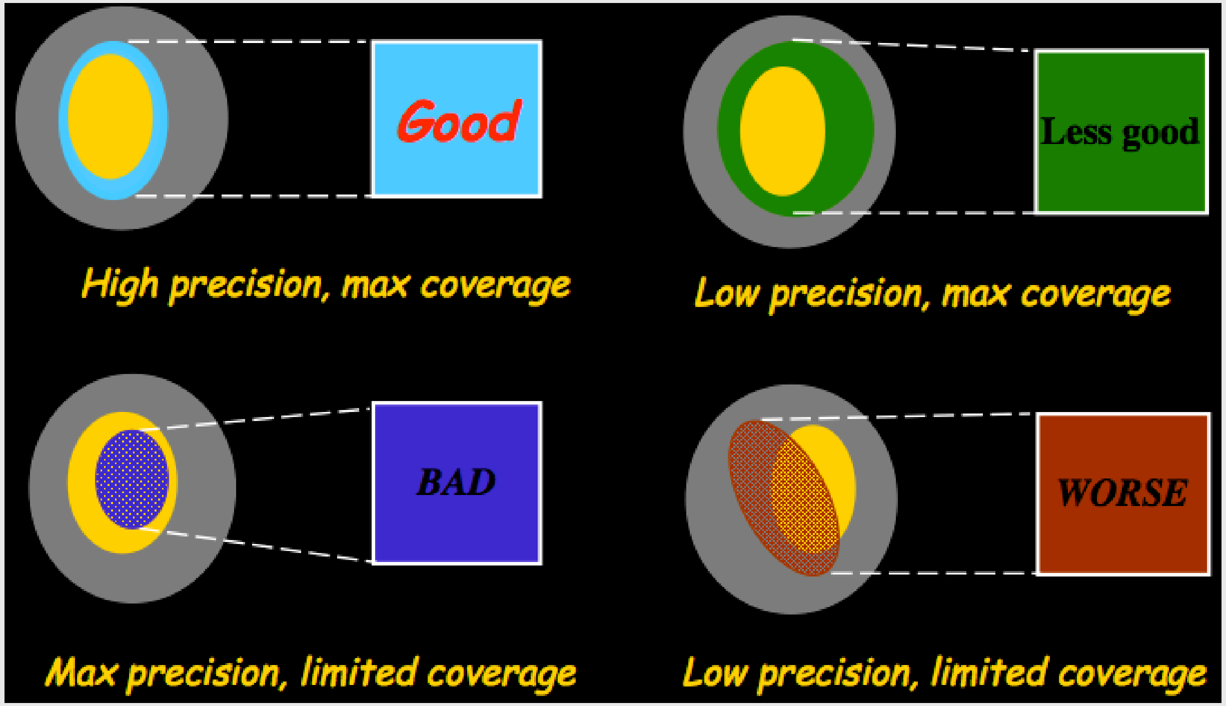


The relationship between an ontology and a conceptualization.
Given a logical language L that implicitly commits to a conceptualization C, an ontology's purpose is to capture the models of L that are compatible with C. These models are called the intended models.

(cont'd)

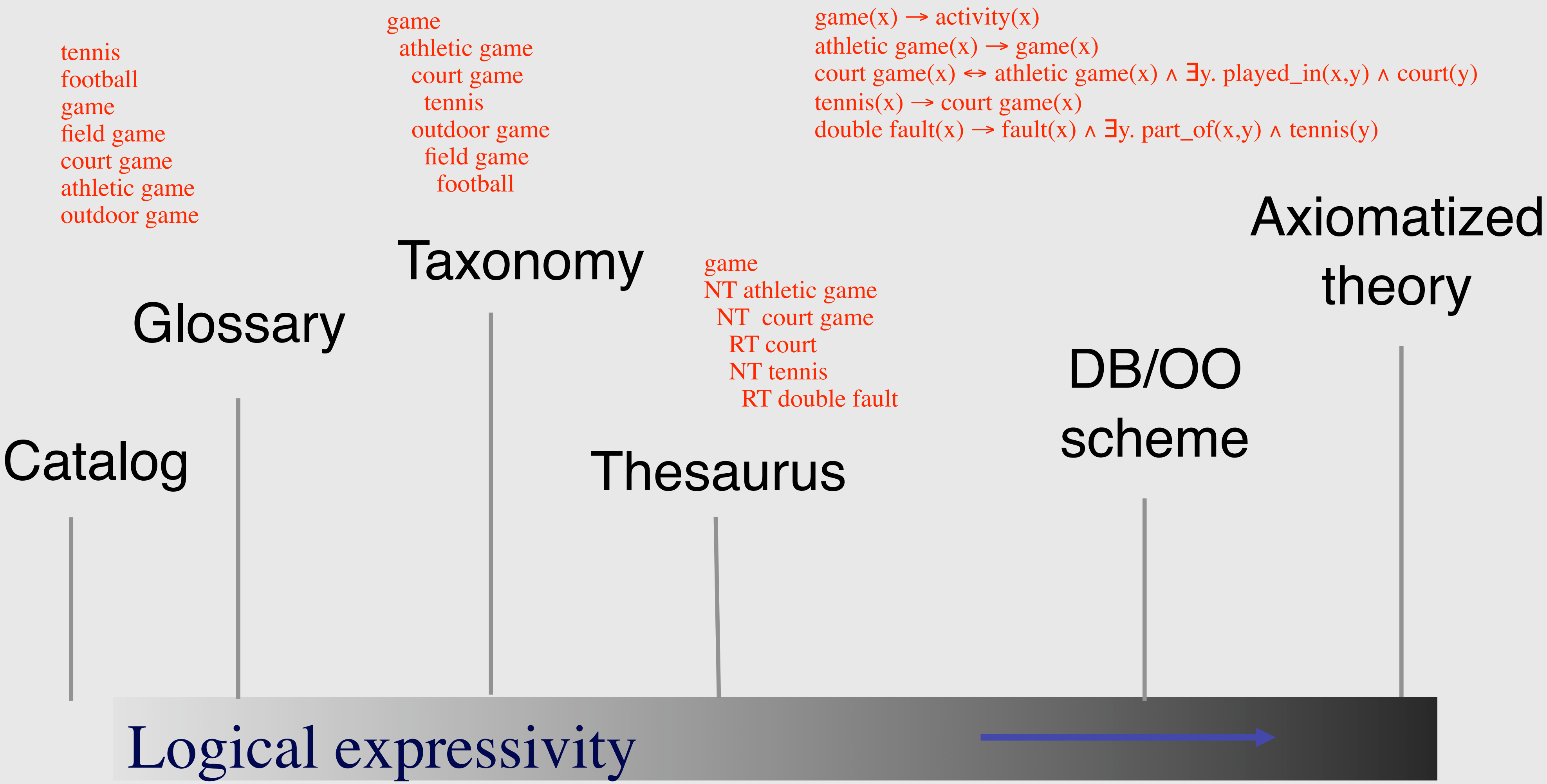
- The formal model in [Guarino 2004] uses an analogy with IR's *precision* and *recall* measures
- Given an intended conceptualization: C = <Δ, *W*, *R*>, where:
  - Δ is a set of relevant entities, *W* a set of possible worlds, and *R* a set of relations
- **Recall** (renamed <u>coverage</u>) = the proportion of intended *models* over Δ allowed by an ontology (intended as a logical theory)
- **Precision** = the proportion of intended *models* on all allowed models
- **Accuracy** = the proportion of intended *states of affairs* on all allowed states of affairs (intended as possible worlds)
  - Models can map different states of affairs

Precision and coverage



The grey ovals includes all models allowed by the logical language.
The yellow ovals include the intended models.
The oval projections of rectangular spaces include the models allowed by an ontology. [Guarino 2004]

Ontologies by expressivity



Is <u>precision</u> dependent on expressivity? That depends on the intended conceptualization: beyond a certain detail, expressivity is key

Accuracy and precision

- When a model catches a state of affairs that is intended, the accuracy of the ontology increases
- When a model catches a state of affairs that is not intended, the accuracy of the ontology decreases
- The precision of an ontology is at least equivalent to its accuracy
- The ideal accuracy requires one model for each state of affairs

An example: the $\mathcal{BWO}$

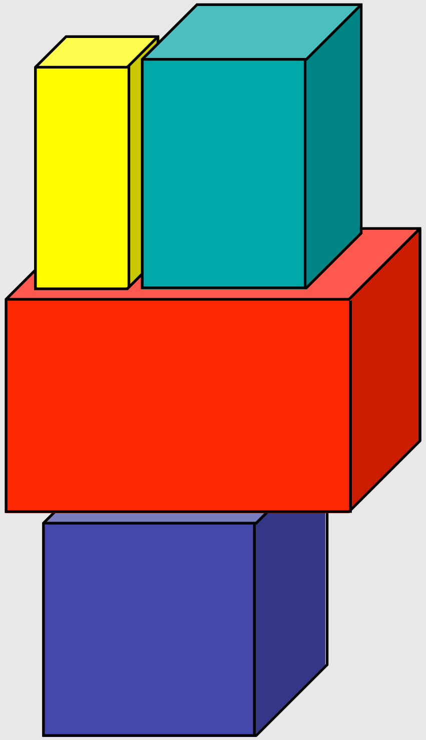
An axiomatic theory: the Blocks World Ontology ($\mathcal{BWO}$)

- Signature: {On, Block}
- Axioms:
- $On(x,y) \rightarrow Block(x) \wedge Block(y)$
- $On(x,y) \rightarrow \neg On(y,x)$ *(antisymmetry)*
- $(On(x,y) \wedge On(y,z)) \rightarrow On(x,z)$ *(transitivity)*

A model $\mathcal{M}$ in $\mathcal{BWO}$:

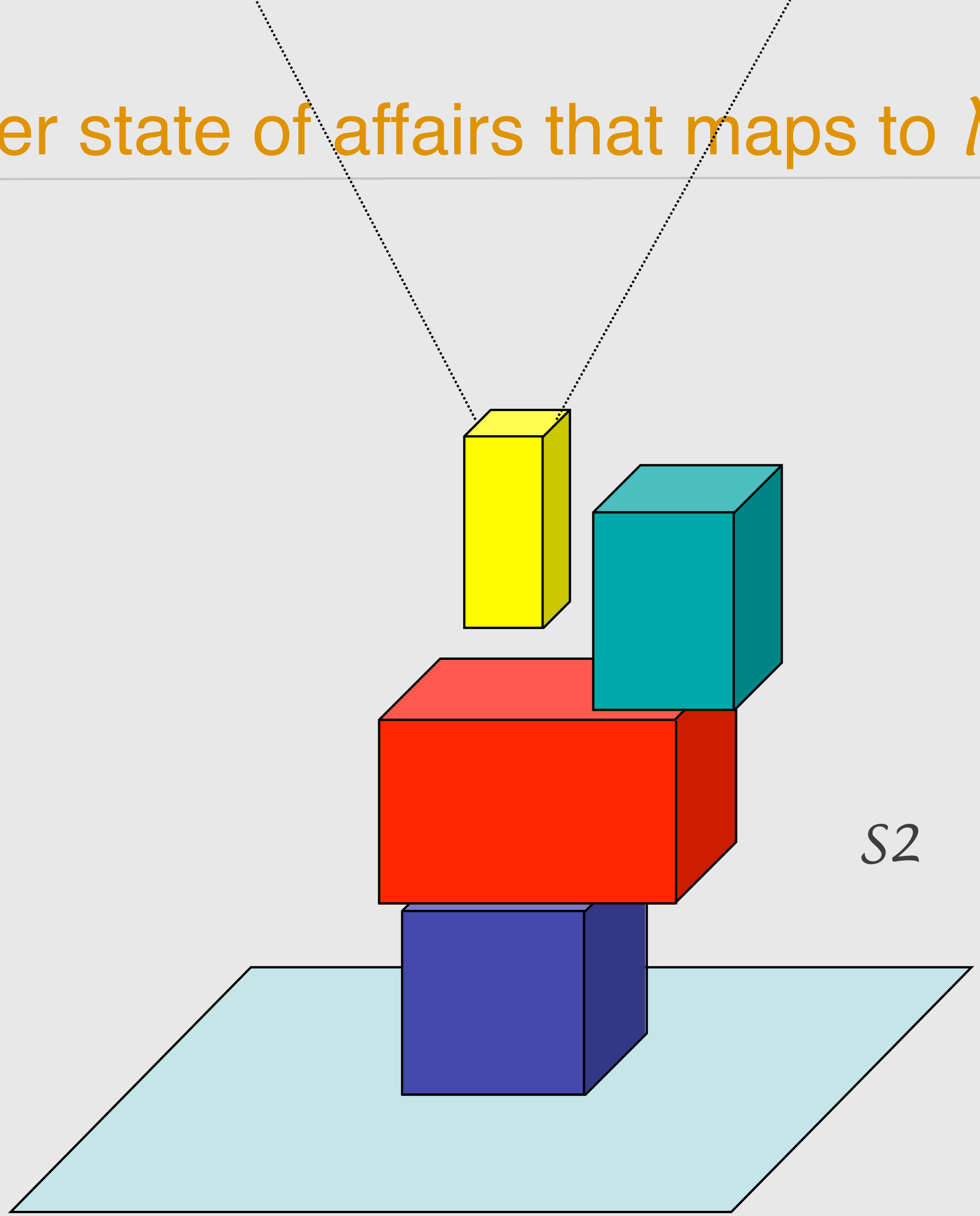
- On(red_block#1, blue_block#1)
- On(green_block#1, red_block#1)
- On(yellow_block#1, red_block#1)

A state of affairs that maps to $\mathcal{M}$



$S1$

Another state of affairs that maps to $\mathcal{M}$



Quality checking of *BWO*

- Does *BWO* catch all the intended *meanings*?
- Is *BWO* so precise in order to exclude non-intended models (if any)?:
  - who knows? we need (counter)examples or qualified expressions of the expertise
- Is *BWO* so accurate in order to exclude non-intended states of affairs (if any)?
  - “Non-Vertical” *On*?
  - “Disconnected” *On*?
  - Time and space? and what space?
  - Supporting plane and cables?
  - Colors?

(cont'd)

- Is $\mathcal{BWO}$ flexible enough?
  - Can we specialize the theory in order to make it more accurate (if needed)?
    - Can we represent supplementary entities ("surface", "cable")?
    - Can we represent time and space primitives?
  - Can we encode alternative axioms for the same signature, in the same or a different theory?

Applicability of functional measures

- How to retrieve the intended conceptualization?
  - Where is it?
- How to distinguish intended models vs. states of affairs in actual ontology projects?
  - Frequent intertwining of formal and informal drafting
- We need to resort again to the semiotic perspective sketched in the O^2 pattern
  - Intended conceptualization and expectations related to states of affairs concern *cognitive*, not *formal* semantics

Intended conceptualization as expertise

- An intended meaning (or conceptualization) can be considered part of the expertise of the ontology intended users (cf. [Steels 1990])
- A state of affairs can be either intended or not based on that expertise
- Models can be admitted or not by the users based on the states of affairs that experts can envisage (e.g. wrt *prototypical examples* and *counter-examples*)
- Alternatively, one could assume a set of data as a *qualified expression* of the expertise, e.g. texts, pictures, diagrams, db records, terminologies, metadata schemas

Qualified expressions of intended conceptualization: some measurement methods for P/R

- *Agreement* assessment (black-box evaluation)
  - through assessment made by experts'
- *Task* assessment (what tasks must be supported by the ontology)
  - wrt (computational) service specification
  - wrt gold-standard
  - wrt (social) task specification
- *Topic* assessment (what are the boundaries of the ontology)
  - wrt to existing metadata resources (e.g. a terminology)
  - wrt to extracted patterns of information (e.g. from a corpus)
- *Modularity* assessment
  - wrt reusable modules

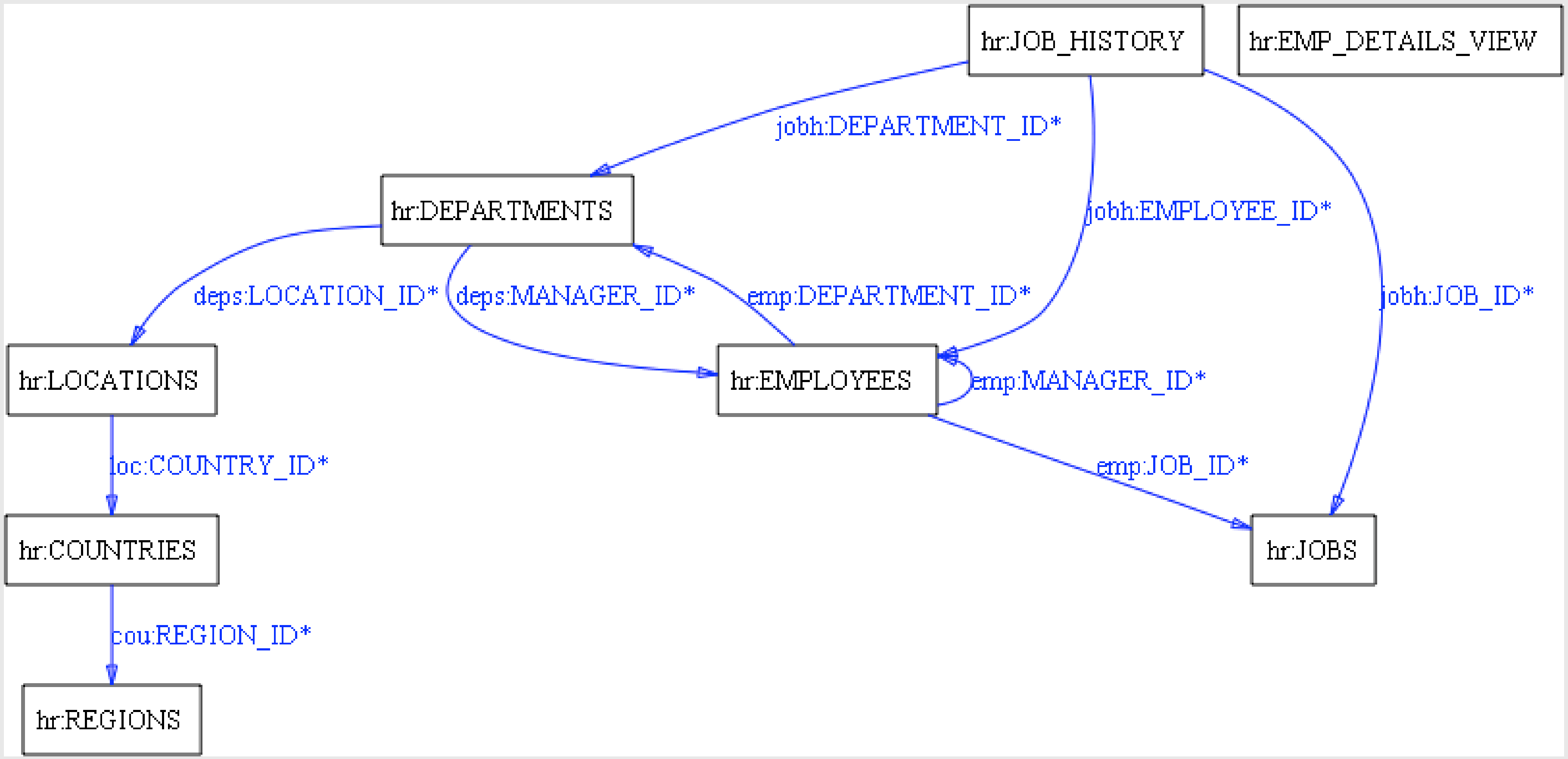
Agreement assessment

- Measuring P/R wrt consensus
  - Black-box measurement: intended conceptualization is left in the experts' mind
  - Requires organized experts and their availability
  - Proposing examples and counter-examples is a typical technique
  - Many other rhetorical/argumentation issues
  - Relation to user-satisfaction
  - Incremental agreement
    - Top-down: see modularity assessment
    - Bottom-up: cf. consensus reaching methodologies (e.g. Uren et al. 2004)

Task assessment

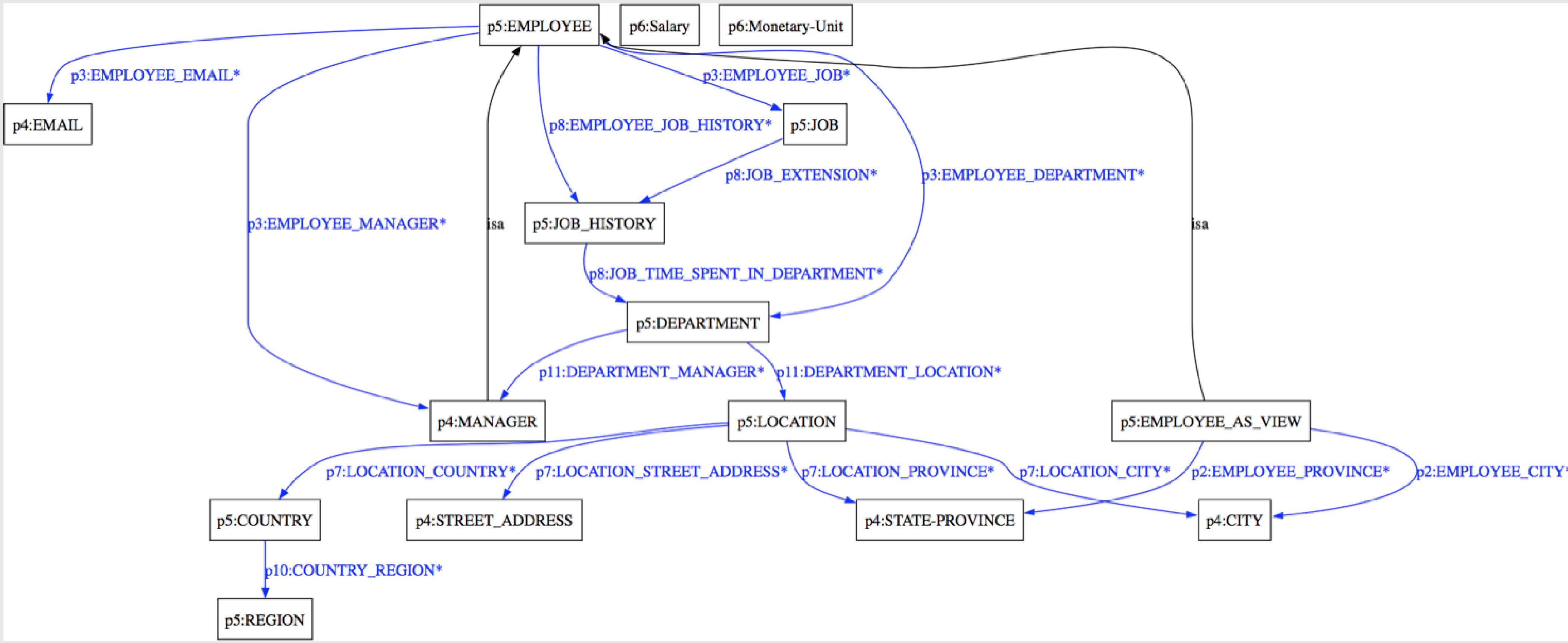
- Measuring P/R wrt fitness to task
  - Requires (explicit or implicit) task specification
  - Three approaches
    - Service specification
      - Profile- or process-model of an application functionality
      - Could miss the relevant social aspects of competency (task ≠ application) ---> The Oracle HR schema example
    - Gold-standard-solution-based
      - Take a validated corpus of answers for the task, and check the performance of an ontology-driven system to those answers (cf. [Porzel et al. 2004])
    - Explicit-task-based
      - Task specification must include reference to domain ontology (combines with topic assessment), and can be generic
      - Competency questions [Grüninger et al. 1996], WSMO [Vasiliu et al. 2004], COS [Oberle et al. 2004] are examples of specification frameworks for tasks
      - Match ontology to task specification

The Oracle HR schema in OWL (application-oriented)



But 40 datatype properties more! (not visualized here): Should an ontology be evaluated also against foreign keys?
This reversely-engineered ontology catches the intended meaning underlying the HR application

The Oracle HR schema in OWL (task-oriented)



Other properties are conveyed by means of restrictions (not visualized here).
This reversely-engineered ontology catches the intended meaning underlying the HR task

Topic assessment

- Measuring P/R wrt fitness to existing information realizations
  - Requires enrichment of ontology with information objects that can be matched against information realizations (e.g. lexical occurrences in texts)
  - Three approaches
    - Choose directory and annotate: operated on a subject directory by annotating the ontology with a subject label: it’s a black-box technique
    - Reengineer and match: operated on metadata repositories, such as terminologies, informal diagrams, DB and OO schemas, etc.
    - Extract and match: operated on data repositories, such as linguistic or image corpora, databases, by extracting information patterns and matching them to ontology graph

Methodologies

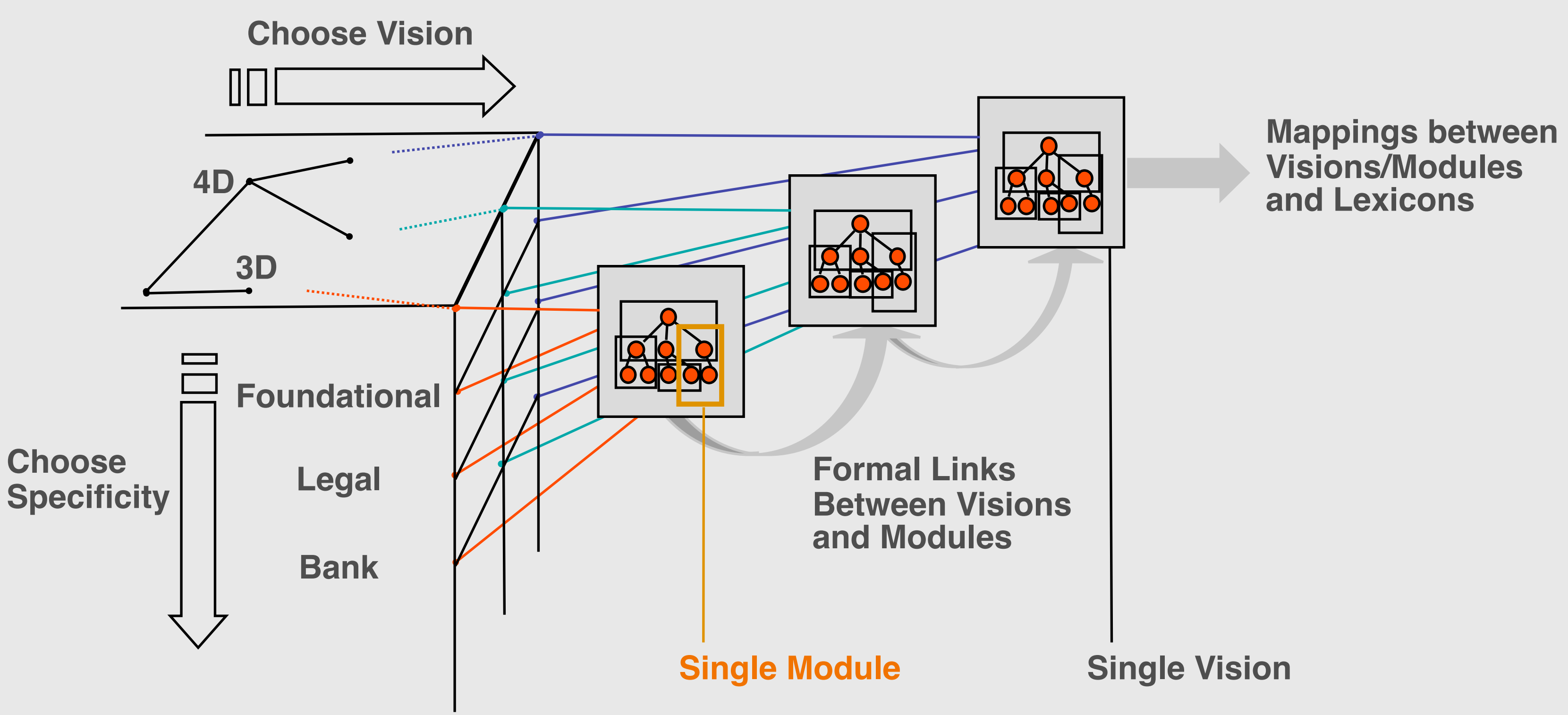
- Reengineer and match
  - Reengineering is based on best practices (e.g. thesauri to OWL), formal translations (e.g. FOL to OWL), and a lot of customization
  - Once reengineered, a source can be either imported or mapped: the degree of difficulty in devising such activities provides P/R measures of the ontology
  - A more directed measure can be obtained by considering the source as a gold-standard-model (cf. [Maedche&Staab 2002])
- Extract and match
  - Extraction is based mostly on learning techniques
  - Matching is controversial, because it depends on the way data are parsed; e.g. a same text can be parsed in different ways, thus obtaining different patterns
  - E.g. in NLP, parsing terms (how complex?) vs. syntactic structures
  - E.g. parsing statistically vs. rule-based
  - An experiment with statistical parsing of syntactic patterns is [Ciaramita et al. 2005]. Other work in e.g. [Brewster et al. 2004].

Modularity assessment

- Measuring P/R wrt fitness to existing reusable ontologies
  - Requires (one or more) libraries of ontologies, with indications of their provenance, specificity, application history, etc. (cf. also usability)
  - Depends on topic assessment: what theories are needed? [Fernandez-Lopez et al. 2004]
    - But also depends on task assessment: how much of a reusable theory is needed? (inherent circularity)
  - Partial matching also possible through content design patterns [Gangemi 2005]
  - Examples of modularization architectures with *reference ontologies*, which allow to factorize ontology projects, as well as their quality assessment

Vision-based vs. topic-based modules

- Modules could be organized along two dimensions:
  - *visions*, related to basic choices made (e.g. wrt *temporal, spatial, identity, information* theories)
  - *specificity*, related to domain stratification of topics (e.g. *foundational*, *legal*, *bank* in a project for the management of *banking regulations*)



A toy example of stratification

In an ontology for historical works of art, domain ontologies define a vocabulary that is typically stratified in foundational, core, and domain layers

Foundational ontology (domain-independent)

{Object, Process, Part, Time, Location, Representation, Plan, …}

↑ *inherits from*

Core ontology (specific-domain-independent)

{Work of art, Painting technique, Author, Artistic period, Plastic art, Interpretation, …}

↑ *inherits from*

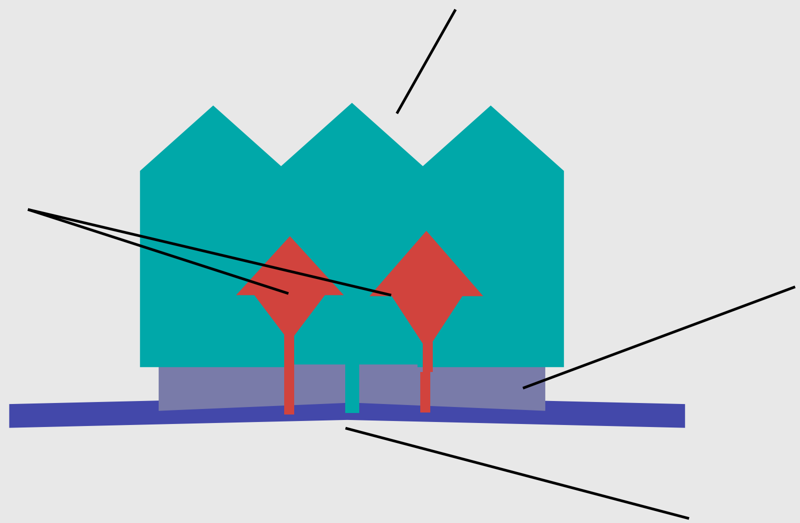
Domain ontology

{Sculpture,Restoration, Mythical being, Caryatid, Doric order, Armilla, Fresco, …}

Stratification in practice: a library for the *Fishery Ontology Service*

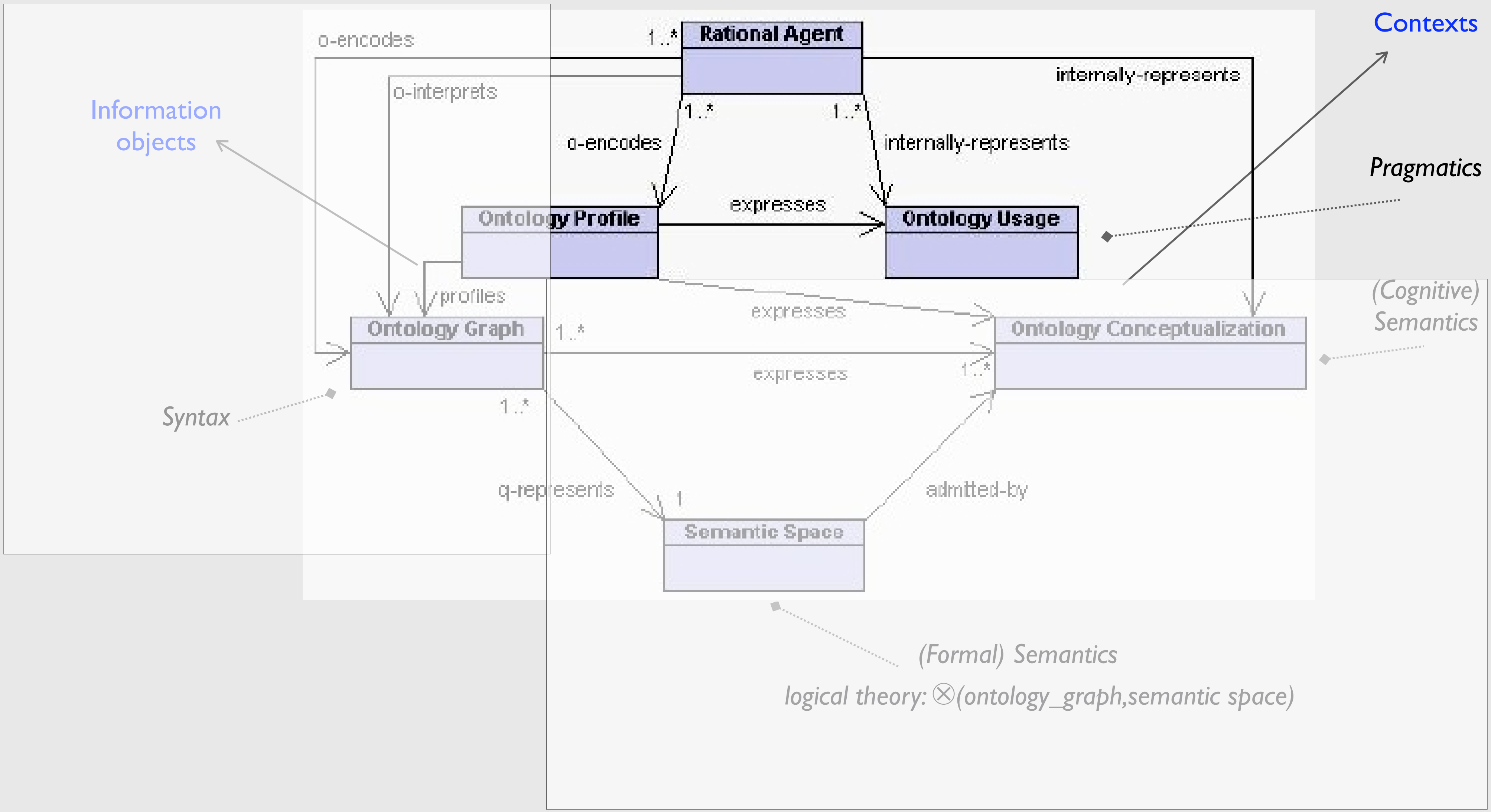
The "toyhouse".
In realistic projects, stratification is usually less 'pristine'. Domain modules often contain bits and pieces of other domains (e.g. geology and law in fishery), thus requiring the reuse of general purpose ontologies like OntoWordNet.

- Biological entities
- Continental and water areas
- Ecosystems
- Techniques (capture, culture)
- Vessels and gears
- Resources, stocks, and management
- Commodities and commercialization
- Institutions



Evaluation - 3
Usability evaluation -
Recognition, Efficiency and Interfacing

A semiotic perspective: the pragmatic aspect



Recognition

- Annotations about the ontology structure
  - Graph measures
  - Logic-type and computational complexity
  - Modularization (e.g. owl:imports)
- Annotations about the ontology function
  - Lexical annotation of ontology elements (incl. multilingual)
  - Glosses (e.g. rdfs:comment)
  - Task
  - Topic (e.g. rdf:about)
  - Modularization design
- Annotations about the ontology lifecycle
  - Use-case-related
  - Methods employed
  - Provenance and Trust rating
  - Versioning (e.g. owl:versionInfo)
  - Customer satisfaction (e.g. http://smi-protege.stanford.edu:8080/KnowledgeZone/)

Amount, *completeness*, and *reliability* of annotations are usability measures ranging on recognition annotations.

Economic efficiency

- Annotations (either on the overall ontology, or on ontology elements) about the organizational design of a modularized ontology, and about the middleware that allows its deployment
- Annotations about the commercial (trading, pricing) and legal (policy, disclaimer) semantics
- Annotations about the application history -with reference to development effort (task- or topic-specificity applied to a token scenario) of an ontology

Presence, *completeness*, and *reliability* of annotations are usability measures ranging on efficiency annotations.

Interfacing

- The interfacing level concerns the process of matching an ontology to a user interface.
- As far as evaluation is concerned, we are only interested in the case when an ontology includes annotations to interfacing operations.

Presence, *completeness*, and *reliability* of annotations are usability measures ranging on interface annotations.

Validation -
Principles, parameters, and preferential ordering

Some principles for ontology quality

- Cognitive ergonomics
- Transparency (explicitness of organizing principles)
- Computational integrity and efficiency
- Flexibility (context-boundedness)
- Compliance to expertise
- Compliance to procedures for extension, integration, adaptation, etc.
- Accessibility (computational as well as commercial)
- Cost-effectiveness

Faster-Better-Cheaper

Parameters on measures, constrained by principles (examples)

- Cognitive ergonomics: -depth, -breadth, -tangledness, +class/property ratio, +annotations (esp. lexical, glosses, topic), -anonymous classes, +interfacing, +patterns
- Transparency: +modularity, +axiom/class ratio, +(dense)patterns, +specific differences, +partitioning, +accuracy, +complexity, +anonymous classes, +modularization design
- Computational integrity and efficiency: +logical consistency, +disjointness ratio, -tangledness, -restrictions, -cycles
- Flexibility: +modularity, +partitioning, +context-boundedness
- Compliance to expertise: +precision, + recall, +accuracy
- Compliance to procedures for extension, integration, adaptation, etc.: +accuracy(?), +linguistic adequacy, +modularity, +annotations, -tangledness(?) ... *see also transparency*
- Cost-effectiveness: +applicability, +cost measure, +customer satisfaction

Preferential ordering and trade-off

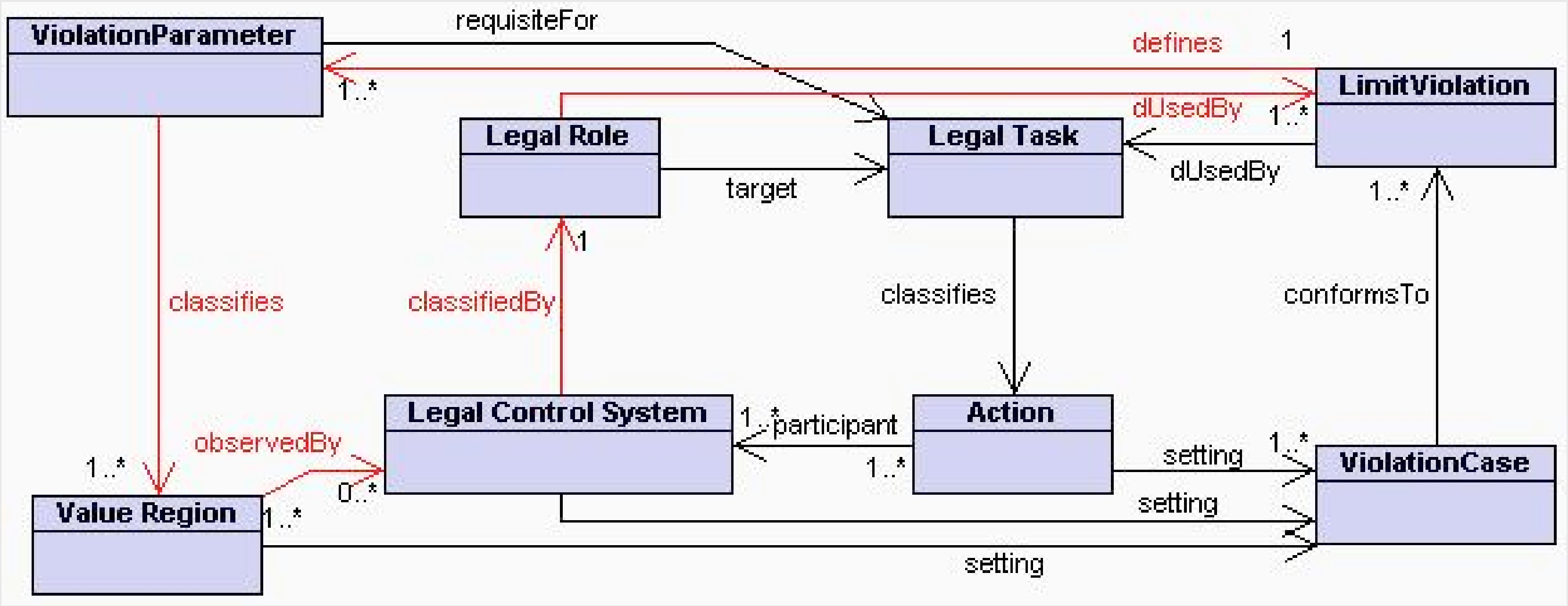
- Due to partly mutual independence of principles, an ontology project should explicitly define its own preferential ordering of quality parameters
- Based on resources, expertise, business relations, available tools, etc.
- Preferential ordering typically arises when a trade-off is needed
- Trade-offs are needed when two or more principles should be composed
- OntoMetric [Lozano-Tello et al. 2004] is an example of a tool that supports measurement based on a preferential ordering
- Ongoing work on defining preferential orderings over typical (generic) use cases and trade-offs

An example: patterns vs. cycles

- *Transparency* and *compliance to expertise* principles usually require *dense* content ontology patterns (cf. Gangemi, 2005), involving hub nodes (cf. Noy, 2004)
- Hubs usually need the definition of sets of (usually existential) axioms that induce complex (in)direct cycles
- *Computational efficiency* principle requires a low rate of cycles (cf. also Berardi et al., 2001 for the complexity of DL ports of UML models)
- Therefore, a trade-off is needed, often leading to tuning practices e.g. involving generalization over restrictions, or even suppression of a restriction when a similar inverse one exists

- ---> An example in legal ontologies

An example of the pattern/cycle trade-off

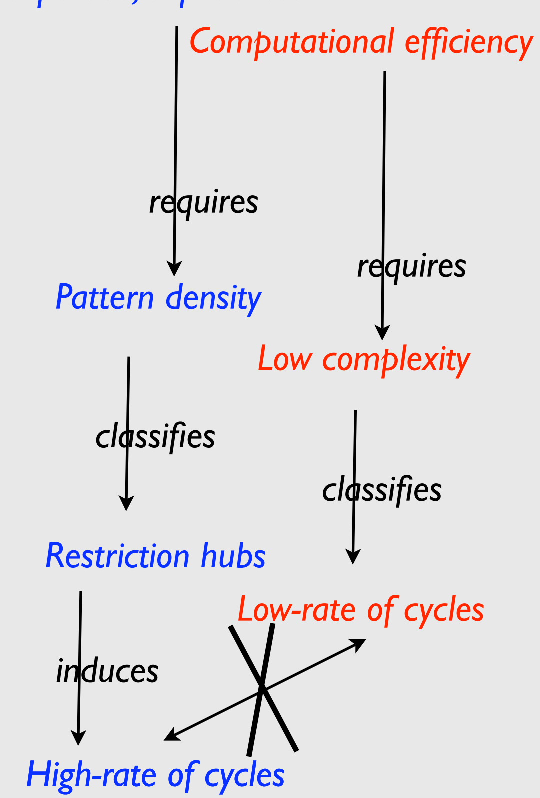


Class(LimitViolation partial restriction(defines someValuesFrom(ViolationParameter)))
Class(ViolationParameter partial restriction(classifies someValuesFrom(ValueRegion)))
Class(ValueRegion partial restriction(observedBy allValuesFrom(LegalControlSystem)))
Class(LegalControlSystem partial restriction(classifiedBy someValuesFrom(LegalRole)))
Class(LegalRole partial restriction(d-used-by someValuesFrom(LimitViolation)))

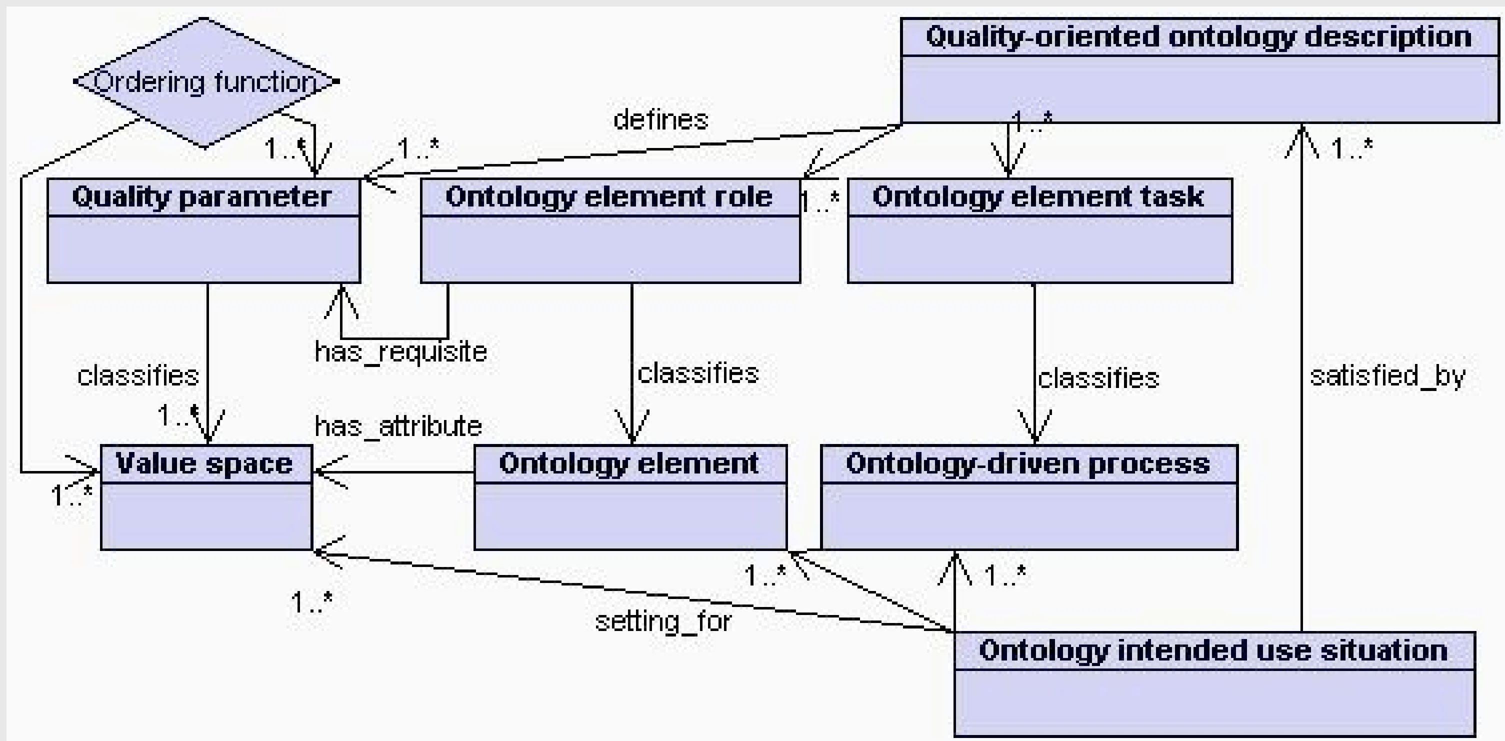
Class(LimitViolation partial restriction(defines someValuesFrom(ViolationParameter)))
Class(ViolationParameter partial restriction(classifies someValuesFrom(ValueRegion)))
**** Class(ValueRegion partial restriction(observedBy allValuesFrom(ControlSystem)))*
**** Class(LegalControlSystem partial restriction(classifiedBy allValuesFrom(Role)))*
Class(LegalRole partial restriction(d-used-by someValuesFrom(LimitViolation)))

Back to the οquαl pattern

****Which principle/parameter should get a higher rank?*



Limit-violation-ontology diagnosis



Limit-violation usage situation

Conclusions

Thanks to too many people (see references for a sample)

and look:

There is nothing so practical as a good theory

(Ludwig Boltzmann, 1844-1906)

Thanks!

Some references

Almuhareb A. and Poesio M., 2004: “Attribute-based and value-based clustering: an evaluation”. In Proceedings of the Conference on Empirical Methods in Natural Language Processing.

Berardi D., Calvanese D., De Giacomo G.: “Reasoning on UML Class Diagrams using Description Logic Based Systems”. In Proc. of the KI'2001 Workshop on Applications of Description Logics, 2001.

Brewster C., Alani H,, Dasmahapatra S. and Wilks Y.: “Data-driven ontology evaluation”. Proceedings of LREC 2004.

Catenacci C., Ciaramita M. Gil R., Gangemi A., Guarino N., and Lehmann J., “Ontology evaluation: A review of methods and an integrated model for the quality diagnostic task”, Deliverable of the OntoDev Italian Ministry of Research Project, available as TechReport at http://www.loa-cnr.it/Publications.html, 2005.

Ciaramita M., Gangemi A., Ratsch E., Saric J., and Rojas I., 2005: “Unsupervised Learning of Semantic Relations between Concepts of a Molecular Biology Ontology”. In Proceedings of the 19th International Joint Conference on Artificial Intelligence.

Fernández-López M. and Gómez-Pérez A., 2004: "Searching for a Time Ontology for Semantic Web Applications". In Proceedings of the 3rd FOIS Conference, A.C. Varzi and L. Vieu (Eds.), IOS Press.

Gangemi A., Borgo S., Catenacci C., and Lehmann J., “Task taxonomies for knowledge content”, Deliverable D07 of the EU FP6 project “Metokis”, available at http://www.loa-cnr.it/Publications.html, 2004.

This bibliography is just a sample.
The most obvious ones, e.g. for OWL, description logics, graph theory, model-theoretic semantics, etc., are not included here.
Many other relevant references are missing, please refer to those cited here and the usual repositories.

Some references

Gangemi A.: “Ontology Design Patterns for Semantic Web Content”. In Motta E. and Gil Y., Proceedings of the Fourth International Semantic Web Conference, 2005.

Gluck M. and Corter J., 1985: “Information, Uncertainty and the Utility of Categories”. In Proceedings of the 7th Annual Conference of the Cognitive Science Society.

Gómez-Pérez A.: “Ontology Evaluation”, in Handbook on Ontologies, S. Staab and R. Studer, eds., Springer-Verlag, 2003,pp. 251–274.

Guarino N.: “Towards a Formal Evaluation of Ontology Quality”. IEEE Intelligent Systems 1541 -1672, 2004.

Guarino N. and Welty C.: “Evaluating Ontological Decisions with OntoClean”, Comm. ACM, vol. 45, no. 2, 2002, pp. 61–65.

Hartmann J., Spyns P., Giboin A., Maynard D., Cuel R., Suárez-Figueroa M.C., and Sure Y.: "Methods for ontology evaluation". Knowledge Web Deliverable D1.2.3, v. 0.1 (2004).

Huang Z., van Harmelen F., and ten Teije A., 2005: “Reasoning with Inconsistent Ontologies”, Proceedings of IJCAI05.

Jakobson R.: “Linguistics and Poetics: Closing Statement”. In: Style in Language. MIT Press, Cambridge, MA (1960)

Lozano-Tello, A. and Gomez-Perez A., 2004: “ONTOMETRIC: A method to choose the appropriate ontology”, Journal of Database Management, 15(2).

Maedche A. and Staab S.: “Measuring the similarity between ontologies”. Proc. of the 13th Conf. on Knowledge Engineering and Knowledge Management, Springer, Berlin, 2002.

Some references

Noy, N.: “Defining N-ary Relations on the Semantic Web: Use With Individuals”, W3C Working Draft: http://www.w3.org/TR/swbp-n-aryRelations/ (2005).

Noy, N.: “Evaluation by Ontology Consumers”. IEEE Intelligent Systems 1541-1672, 2004.

Oberle D., Gangemi A., Mika P., and Sabou M.: “Foundations for service ontologies: Aligning OWL-S to DOLCE”, in Staab S. and Patel-Schneider P. (eds.), Proceedings of the World Wide Web Conference (WWW2004), Semantic Web Track, (2004).

Pantel P. and Ravichandran D., 2004: “Automatically Labeling Semantic Classes”. In Proceedings of HLT-NAACL 2004.

Pasca M. and Harabagiu S.H., 2001: “The Informative Role of WordNet in Open-Domain Question Answering”. In NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations.

Porzel R. and Malaka R.: “A Task-based Approach for Ontology Evaluation”. Proc. of ECAI 2004.

Riloff E., 1996: “An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains”. Artificial Intelligence, 85.

Steels L.: “Components of Expertise”, AI Magazine, 11, 2, 1990, pp. 30-49.

Sugiura N., Shigeta Y., Fukuta N., Izumi N., and Yamaguchi T., 2004: “Towards On-the-fly Ontology Construction Focusing on Ontology Quality Improvement”, ESWC04.

Sure Y. (ed.), 2004: “Why Evaluate Ontology Technologies? Because It Works!”, IEEE Intelligent Systems 1541-1672.

Some references

Uren V., Buckingham Shum S, Mancini C. and Li G.: "Modelling Naturalistic Argumentation in Research Literatures". Proceedings of the 4th Workshop on Computational Models of Natural Argument, 2004.

Uschold U. and Gruninger M.: "Ontologies: Principles, Methods, and Applications," Knowledge Eng. Rev.,vol. 11,no. 2,1996, pp. 93–155.

Vasiliu L., Moran M., Bussler C., Roman D., "WSMO in DIP", DIP EU FP6 Project Deliverable D19.1, v0.1, available at http://www.wsmo.org/2004/d19/d19.1/v0.2/, 2004

Welty C., Kalra R., and Chu-Carroll J., 2003: "Evaluating Ontological Analysis". In Proceedings of the ISWC-03 Workshop on Semantic Integration.

Yao H., Orme A.M., and Etzkorn L., 2005: "Cohesion Metrics for Ontology Design and Application", Journal of Computer Science, 1(1): 107-113, 2005.