

Division of semantic labor over vocabulary and ontology layers

Piek Vossen, VU University Amsterdam

Flarenet-Silt workshop on Ontology and Lexicon

September-19th-2009, Pisa

Modeling knowledge in a domain

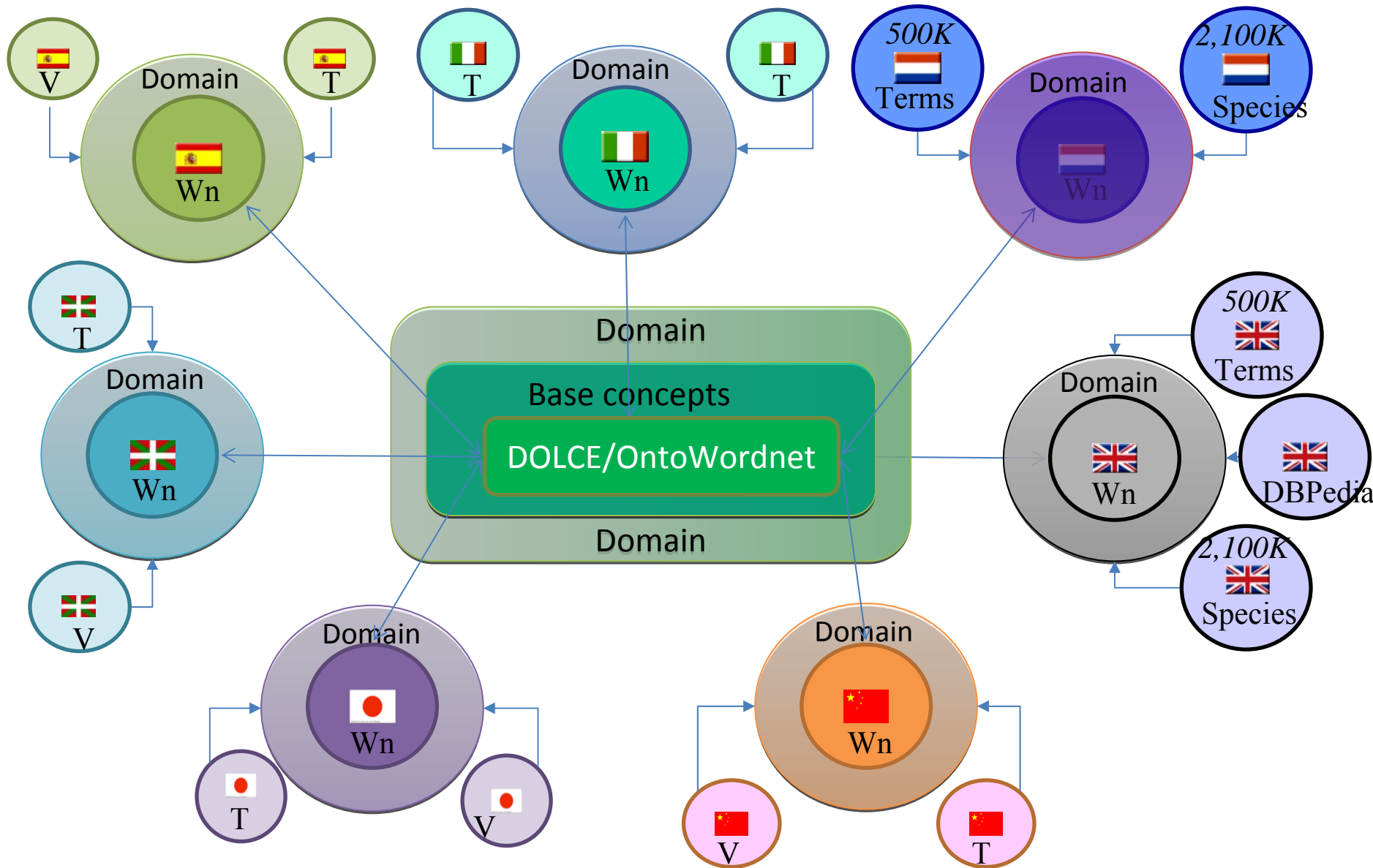
- Knowledge needs to be divided over different lexical and ontological layers because:
 - the volume of terms and concepts is too large
 - the terms are linguistically too diverse
- Division of knowledge over different layers implies:
 - Precisely define the relations between lexical and ontological layers
 - Precisely define the inferencing based on the distributed knowledge layers

Repositories for Kyoto project on the environment domain

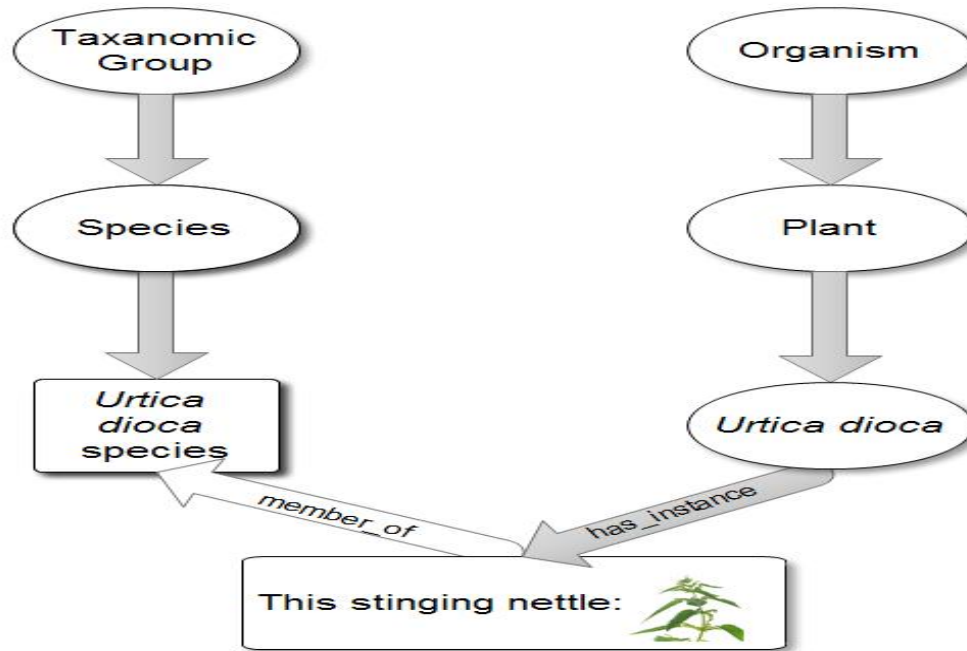
- **Term database:** 500,000 terms per 1,000 documents per language
- Open data project:
 - **DBPedia:** 2.6 million things, including at least 213,000 persons, 328,000 places, 57,000 music albums, 36,000 films, 20,000 companies.
The knowledge base consists of 274 million pieces of information (RDF triples).
 - **GeoNames:** 8 million geographical names and consists of 6.5 million unique features whereof 2.2 million populated places and 1.8 million alternate names.
- **Domain thesauri and taxonomies:** Species 2000: 2,1 million species
- **Wordnets** for 7 languages: about 50,000 to 120,000 synsets per language
- **Ontologies:** SUMO, DOLCE, SIMPLE



Kyoto Knowledge Base



Species in the ontology



- Implies to store 2.1 million species twice in the ontology.

Should all knowledge be stored in the central ontology?

- Vocabularies are too large for full inferencing with current reasoners
- Vocabularies are linguistically too diverse to be represented in an ontology
- Inferencing capabilities of formal ontologies is not needed for all levels of knowledge

Division of linguistic labor principle

- Putnam 1975:
 - No need to know all the necessary and sufficient properties to determine if something is "gold"
 - Assume that there is a way to determine these properties and that domain experts know how to recognize instances of these concepts.
 - Speakers can still use the word "gold" and communicate useful information

Division of semantic labor principle

- Digital version of Putnam (1975):
 - Computer does not need to have all the necessary and sufficient properties to determine if something is a "European tree frog"
 - Computer assumes that there is a way to determine this and that domain experts (people) know how to recognize instances of these concepts.
 - Computers can still reason with semantics and do useful stuff with textual data

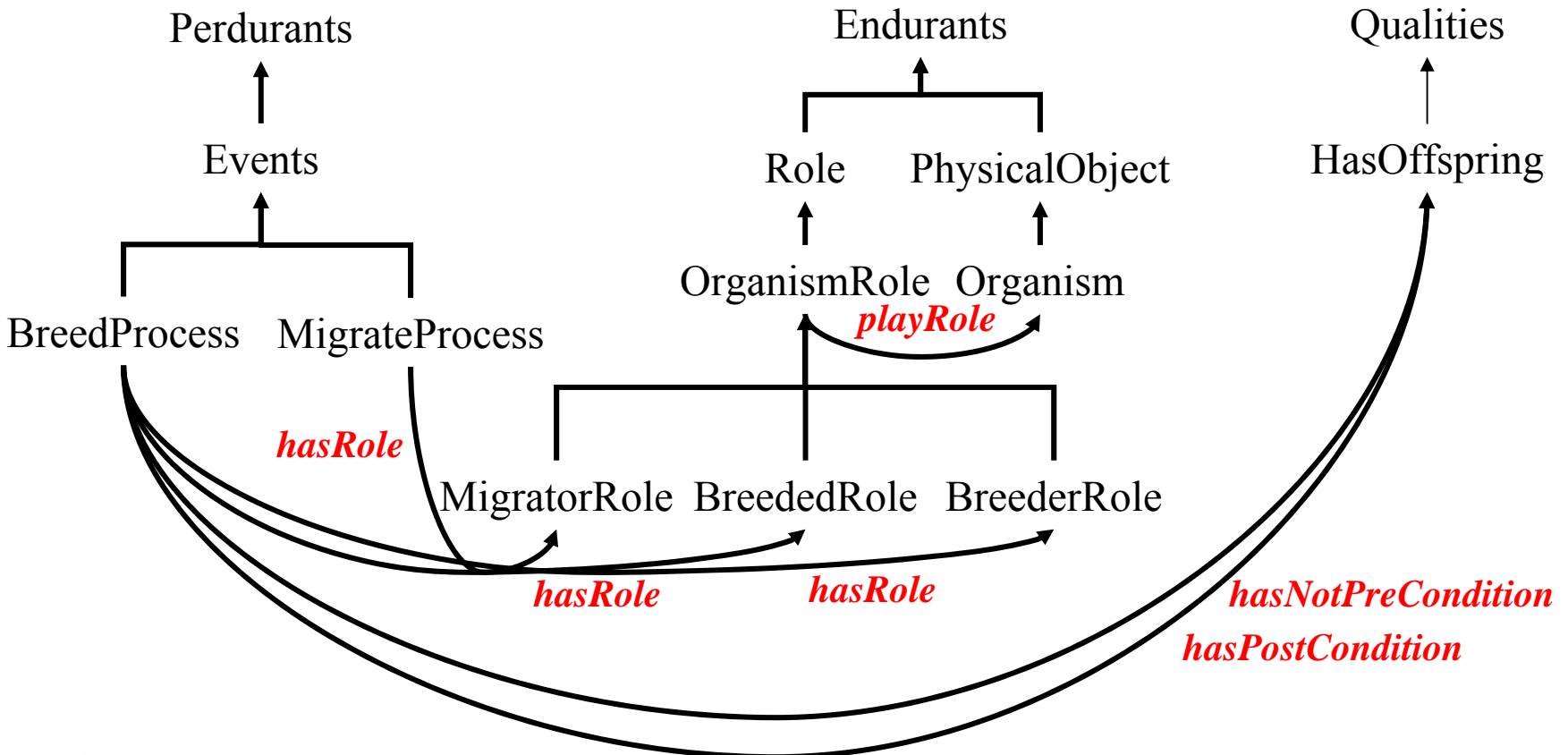
What does the computer need to know?

- Distinction between rigid and non-rigid (Welty & Guarino 2002):
 - being a "*cat*" is essential to individual's existence and therefore rigid
 - being a "*pet*" is a temporarily role and therefore non-rigid; a cat can become a pet and stop being a pet without ceasing to exist
 - Felix is born as a cat and will always be a cat, but during some period Felix can become a pet and stop being a pet while it continuous to exist
- All 2.1 million species are rigid concepts

What does the computer need to know?

- Roles and processes in documents have more *information value* than the defining properties of species:
 - Species defined in terms of physical properties already known to expert;
 - Roles such as "invasive species", "migration species", "threatened species" express THE important properties of instances of species
- Roles are typically the terms we learn from the text not the species!

Ontology relations based on Dolce



Wordnet-ontology-relations

- Rigid synsets:
 - Synset:Endurant; Synset:Perdurant; Synset:Quality:
 - `sc_equivalenceOf` (= relation in WN-SUMO) or `sc_subclassOf` (+ relation in WN-SUMO)
- Non-rigid synsets:
 - Synset:Role; Synset:Endurant
 - `sc_domainOf`: range of ontology types that restricts a role
 - `sc_playRole`: role that is being played
- Rigidity can be detected automatically (Rudify, 80% precision, IAG 80%) and is stored in wordnets as attributes to synsets

Lexicalization of process-related concepts

{obstruct, obturate, impede, occlude, jam, block, close up} Verb, English

-> sc_equivalenceOf *ObstructionPerdurant*

{obstruction, obstructor, obstructer, impediment, impedimenta} Noun, English

-> sc_domainOf *PhysicalObject*

-> sc_playRole *ObstructingRole*

{migration birds} Noun, English

-> sc_domainOf *Bird*

-> sc_playRole *MigratorRole*

{migration} Verb, English

-> sc_equivalenceOf *MigrationProcess*

{migration area} Noun, English

-> sc_domainOf *PhysicalObject*

-> sc_playRole *TargetRole*

Lexicalization of process-related concepts

{create, produce, make} Verb, English

-> sc_equivalenceOf **ConstructionProcess**

{artifact, artefact} Noun, English

-> sc_domainOf **PhysicalObject**

-> sc_playRole **ConstructedRole**

{kunststof} Noun, Dutch // lit. *artifact substance*

-> sc_domainOf **AmountOfMatter**

-> sc_playRole **ConstructedRole**

{meat} Noun, English

-> sc_domainOf **Cow, Sheep, Pig**

-> sc_playRole **EatenRole**

{名肉, 食物, 餐} Noun, Chinese

-> sc_domainOf **Cow, Sheep, Pig, Rat, Mole, Monkey**

-> sc_playRole **EatenRole**

{طعام, لحم, غذاء} Noun, Arabic

-> sc_domainOf **Cow, Sheep**

-> sc_playRole **EatenRole**

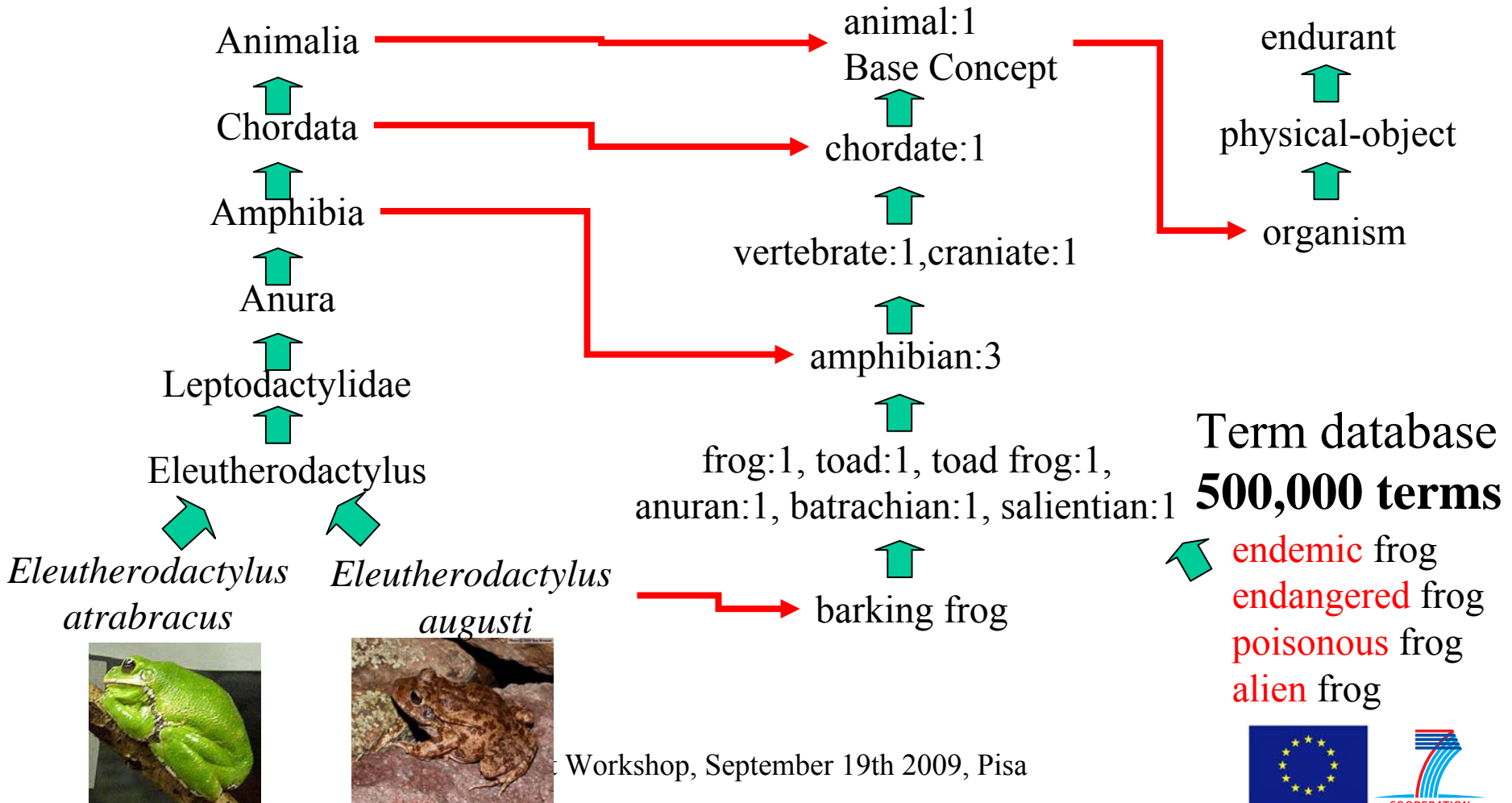


Division of labor in knowledge sources

Skos database
2.1 million species

Wordnet-LMF
100,000 synsets

Ontology-OWL-DL
1,000 types



How to make inferences?

- Sparql queries to large Virtuoso databases:
Aligned Species 2000, DBPedia
- Sql queries to term database
- Graph matching on wordnets
- Reasoning on a small ontology

Relations in the ontology

“Highways in the Humber Estuary obstruct the migration of birds.”

// endurants

(subclass, Road, PhysicalObject)

(subclass, Organism, PhysicalObject)

// roles

(subclass, LocationRole, Role)

(subclass, MigratorRole, Role)

(subclass, MigrationTargetRole, Role)

(subclass, ConstructorRole, Role)

(subclass, ConstructedRole, Role)

(subclass, ObstructingRole, Role)

(subclass, ObstructedRole, Role)

// perdurants

(subclass, ObstructionPerdurant, Perdurant)

(hasRole, ObstructionPerdurant, ObstructingRole)

(playedBy, ObstructingRole, PhysicalObject)

(subclass, MigrationProcess, Process)

(hasRole, MigrationProcess, MigratorRole)

(hasRole, MigrationProcess, MigrationTargetRole)

(playedBy, MigratorRole, Organism)



Instantiation of the ontology

“Highways in the Humber Estuary obstruct the migration of birds.”

(instanceOf, 0, Location)	<!--obstruction -->
(instanceOf, 1, Road)	(instanceHasRole, 3, 5) //involves obstructing role
(instanceOf, 2, Organism)	(instanceHasRole, 3, 6) //involves obstructed role
(instanceOf, 3, ObstructionPerdurant)	(instanceHasRole, 3, 8) //takes place in location
(instanceOf, 4, MigrationProcess)	(instancePlay, 1, 5) //highways play this obstructing role
(instanceOf, 5, ObstructingRole)	(instancePlay, 2, 6) //birds play this obstructed role
(instanceOf, 6, ObstructedRole)	(instancePlay, 0, 8) //Humber Estuary plays LocationRole
(instanceOf, 7, MigratorRole)	<!--migration -->
(instanceOf, 8, LocationRole)	(instanceHasRole, 4, 7) //involves a migrator role
	(instanceHasRole, 4, 9) //involves target location
	(instanceHasRole, 4, 10) //has LocationRole
	(instancePlay, 2, 7) //birds play this migrator role
	(instancePlay, 0, 8) //Humber Estuary plays location role



Ontology relations (DOLCE)

- Endurant (objects), Perdurant (processes), Quality
- subClassOf, equivalentTo, generic-constituent relations:
Endurant:Endurant, Perdurant:Perdurant, Quality:Quality
- *Role* hierarchy below *endurant*:
 - *OrganismRole* -> *BreedingRole*
 - *MigrationRole* -> *BirdMigrationRole*
- playedBy relation: *Role:Endurant*
- hasRole relation: *Perdurant:Role*
- instanceOf: *Instance:Endurant/Perdurant*
- instancePlay: *Instance:Role*

How to integrate the data?

- Species 2000 vocabulary: 2,171,281 concepts in MySQL database with parent relations:
 - Kingdom -> Class -> Order -> Family -> Genus -> Species -> Infra species
 - Animalia -> Chordata -> Amphibia -> Anura -> Leptodactylidae -> Eleutherodactylus -> Eleutherodactylus augusti
- Converted to SKOS format
- Aligned with DBPedia for language labels
- Aligned with Wordnet using vocabulary and relation mappings
- Published in Virtuoso, accessed with SPARQL queries

How to integrate data?

Extending language labels using DBPedia

Language	Species 2000	DBPedia extension
English	69,045	834,821
Spanish	1,731	358,499
Italian	17,552	215,511
Dutch	5,397	185,437
Chinese	58,774	83,756
Japanese	4,625	139,754

How to integrate data?

Alignment Species 2000 with wordnet

- Vocabulary match with Wordnet synsets
- If polysemous then SSI-Dijkstra weighting of senses based on the hyperonym chain
- Results still to be evaluated:
 - Animalia (*animal:1*)-> Chordata (*chordate:1*) -> Amphibia (*amphibian:3*) -> Anura -> Leptodactylidae -> Eleutherodactylus -> Eleutherodactylus augusti (*barking frog:1*)

How to integrate data?

Alignment of terms with wordnet

- Word-sense-disambiguation is applied to terms in KAF (Kyoto Annotation Format)
- Term hierarchy is extracted from KAF:
 - land:5
 - grassland:1 -> *biome:1*
 - woodland:1 -> *biome:1*
 - cropland
 - urban land
- Results still to be evaluated: SemEval2010

Should all knowledge be stored in the central ontology?

- Vocabularies are too large for full inferencing
- Vocabularies are linguistically too diverse to be represented in an ontology
- Inferencing capabilities of formal ontologies is not needed for all levels of knowledge
- A model of division of labor (along the lines of Putnam 1975) in which knowledge is stored in 3 layers:
 - SKOS vocabularies and term databases
 - wordnet (WN-LMF)
 - ontology (OWL-DL),
- Each layer supports different types of inferencing ranging from Sparql queries, graph algorithms to reasoning.
- Mapping relations that support the division of labour and different types of inferencing and that allow for the encoding of language-specific lexicalizations and restrictions.