Advanced Information Systems Laboratory

Bibliographic databases, an ontological perspective

Javier Lacasta, Javier Nogueras-Iso, Gilles Falquet, Jacques Teller, F. Javier Zarazaga-Soria

20 June 2012, OAI8, Genève

http://iaaa.unizar.es
Computer Science and Systems Engineering Department

Universidad Zaragoza

UNIVERSITÉ DE GENÈVE

Université de Liège

Advanced Information Systems Laboratory (IAAA)

- Computer Science and Systems Engineering Dept., University of Zaragoza, Spain http://iaaa.cps.unizar.es/
- Management of GeoSpatial Information
  - Application domains: environment, administration, emergency response
- Key topic : semantic interoperability
  - Information retrieval (multilingual): metadata generation, indexing, ranking
- Current focus:
  - Semantic Web technologies
    - Give information a well-defined meaning through shared reference to ontologies available on the Web
  - Ontology learning
    - Automatic development of domain ontologies
  - Geospatial Linked Open Data

An ontological perspective of bibliographic databases, applicability to urbanism

- Process to improve the descriptions of resources in digital libraries
  - Formalization of knowledge models used for classification
  - Alignment with existent formal ontologies

- Introduction and objectives
- Proposed process
- Experiments in the field of urbanism

Introduction and objectives

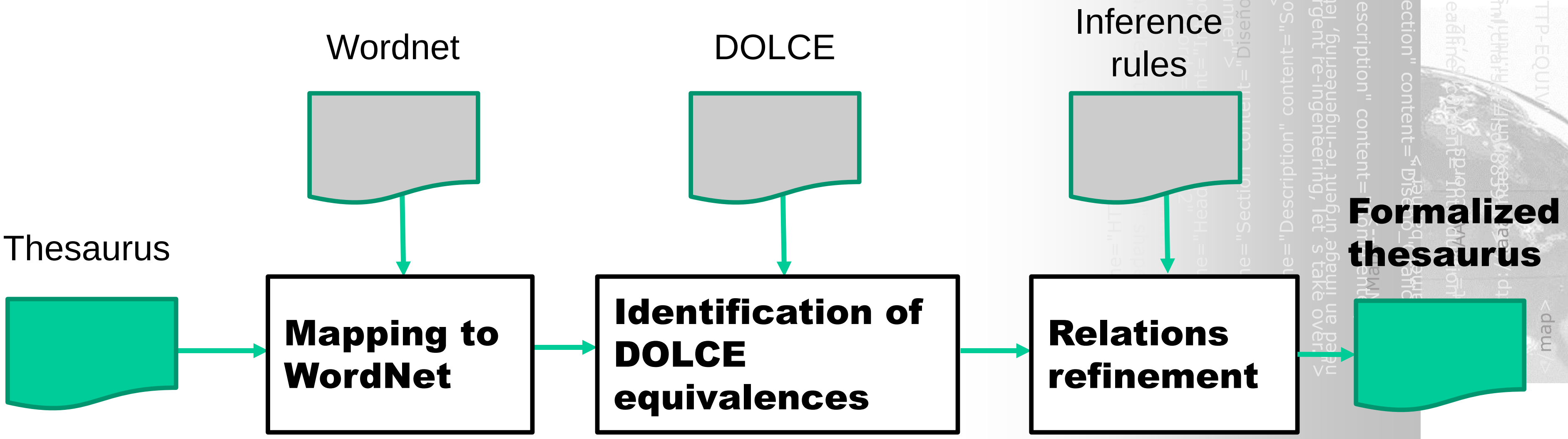
- Collections are frequently classified an searched using terms from thesauri
  - Reduce terminological heterogeneity
  - Facilitate users the selection of search terms
- Usability of the indexed collection is not as good as it could be due to the limited semantics
  - Ambiguity in the definition of concepts
  - Heterogeneity in interpretation of relations
    - Expansion of queries with vague narrower concepts can introduce wrong results
    - Browsing through an unclear hierarchy is difficult

Transform a thesaurus into an ontology

- Solution: Replace the thesaurus used for classification with an ontology
  - Formal definition of the concepts and the relations
  - There are no specialized ontologies in all the fields
- Create a formal ontology from scratch
  - Costly for models with thousands of concepts
- Add formalism to used thesaurus
  - Link the thesaurus with a top level ontology like DOLCE to provide additional semantics about the concepts
    - 3 families of DOLCE abstract categories
      - Perdurants: events, processes, phenomena, activities, states
      - Endurants: entities that maintain their identity along the time (physical objects, social objects such as society)
      - Qualities: entities that can be perceived or measured (color, shape)
    - It facilitates the refinement of vague relations

Alignment-based method for the formalization of thesauri

- **Need to cover the abstraction gap between the thesaurus and DOLCE**
    - **Thematic thesaurus concepts are too specific**
    - **DOLCE concepts are too general**
- **Our approach**
    - **Use WordNet lexical database as intermediate structure**
    - **Hyponym/hypernym Wordnet hierarchy allow connecting specific concepts with abstract categories of DOLCE**



Wordnet

DOLCE

Inference rules

Thesaurus

Mapping to WordNet

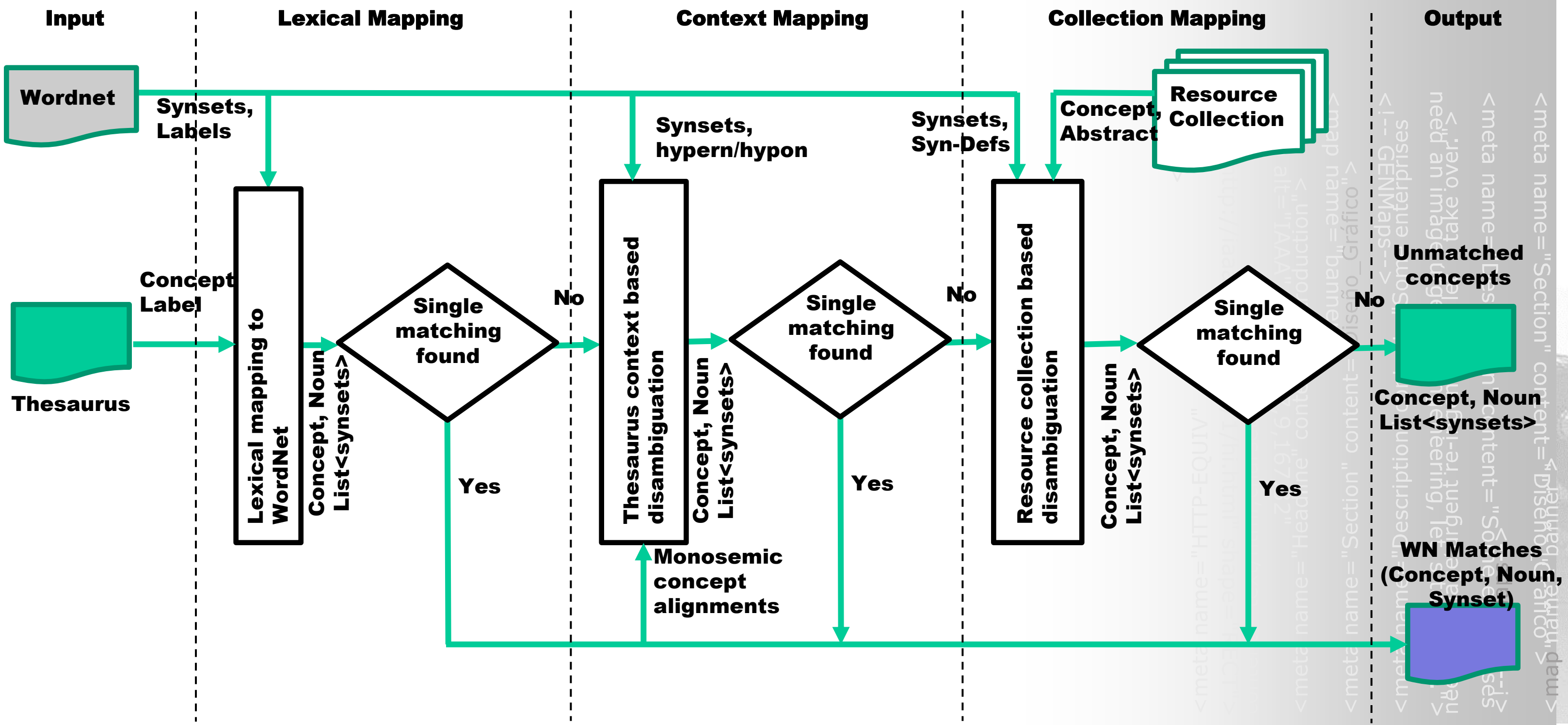
Identification of DOLCE equivalences

Relations refinement

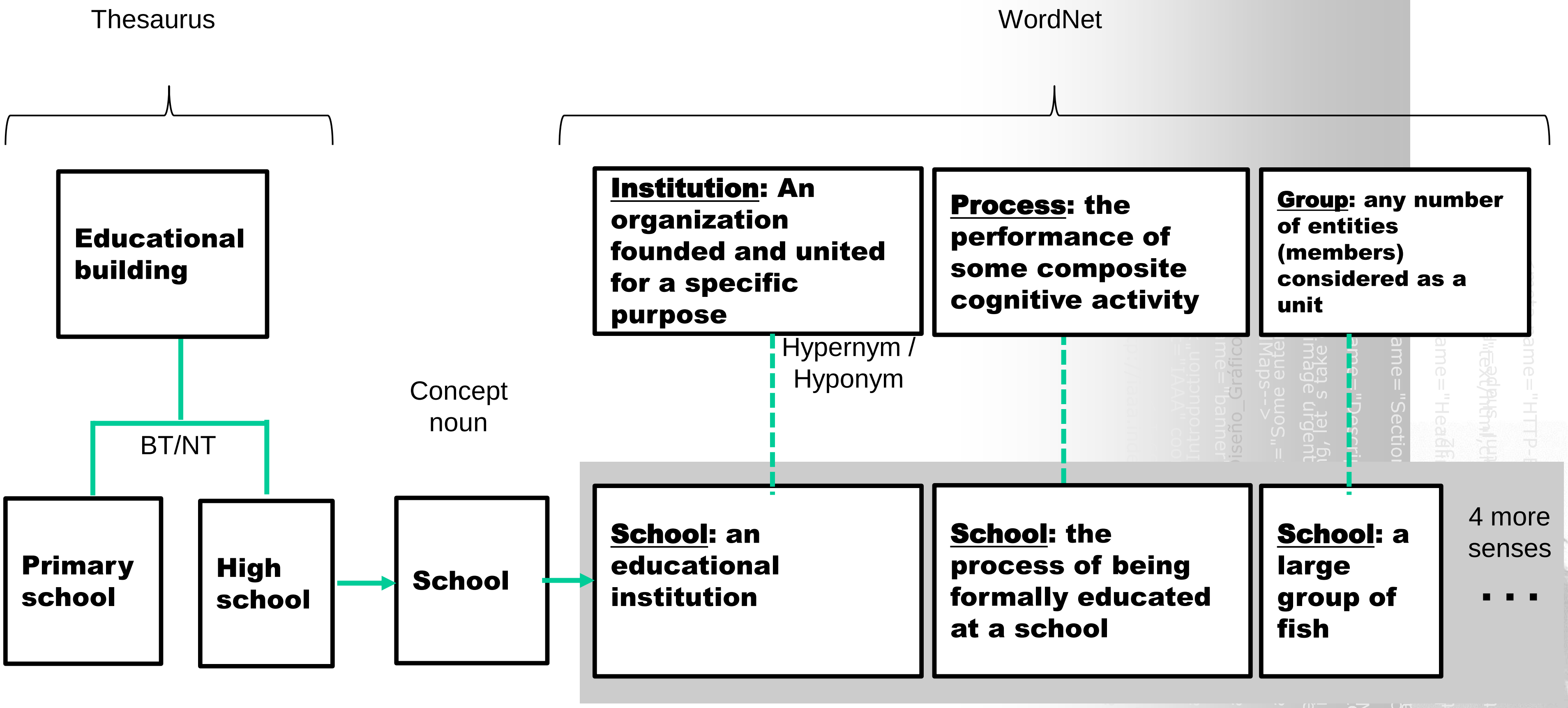
Formalized thesaurus

Mapping between a thematic thesaurus and WordNet

- Usually, thesaurus concepts haven’t got a direct and monosemic matching in Wordnet
- Thus, we need additional heuristics


Input
Lexical Mapping
Context Mapping
Collection Mapping
Output
Wordnet
Synsets, Labels
Synsets, hypern/hypon
Synsets, Syn-Defs
Concept, Abstract
Resource Collection
Concept Label
Thesaurus
Lexical mapping to WordNet
Concept, Noun List<synsets>
Single matching found
No
Yes
Thesaurus context based disambiguation
Concept, Noun List<synsets>
Single matching found
No
Yes
Resource collection based disambiguation
Concept, Noun List<synsets>
Single matching found
No
Yes
Monosemic concept alignments
Unmatched concepts
Concept, Noun List<synsets>
WN Matches (Concept, Noun, Synset)

An example of the sense disambiguation problem



Thesaurus

WordNet

Educational building

BT/NT

Primary school

High school

Concept noun

School

Institution: An organization founded and united for a specific purpose

Process: the performance of some composite cognitive activity

Group: any number of entities (members) considered as a unit

Hypernym / Hyponym

School: an educational institution

School: the process of being formally educated at a school

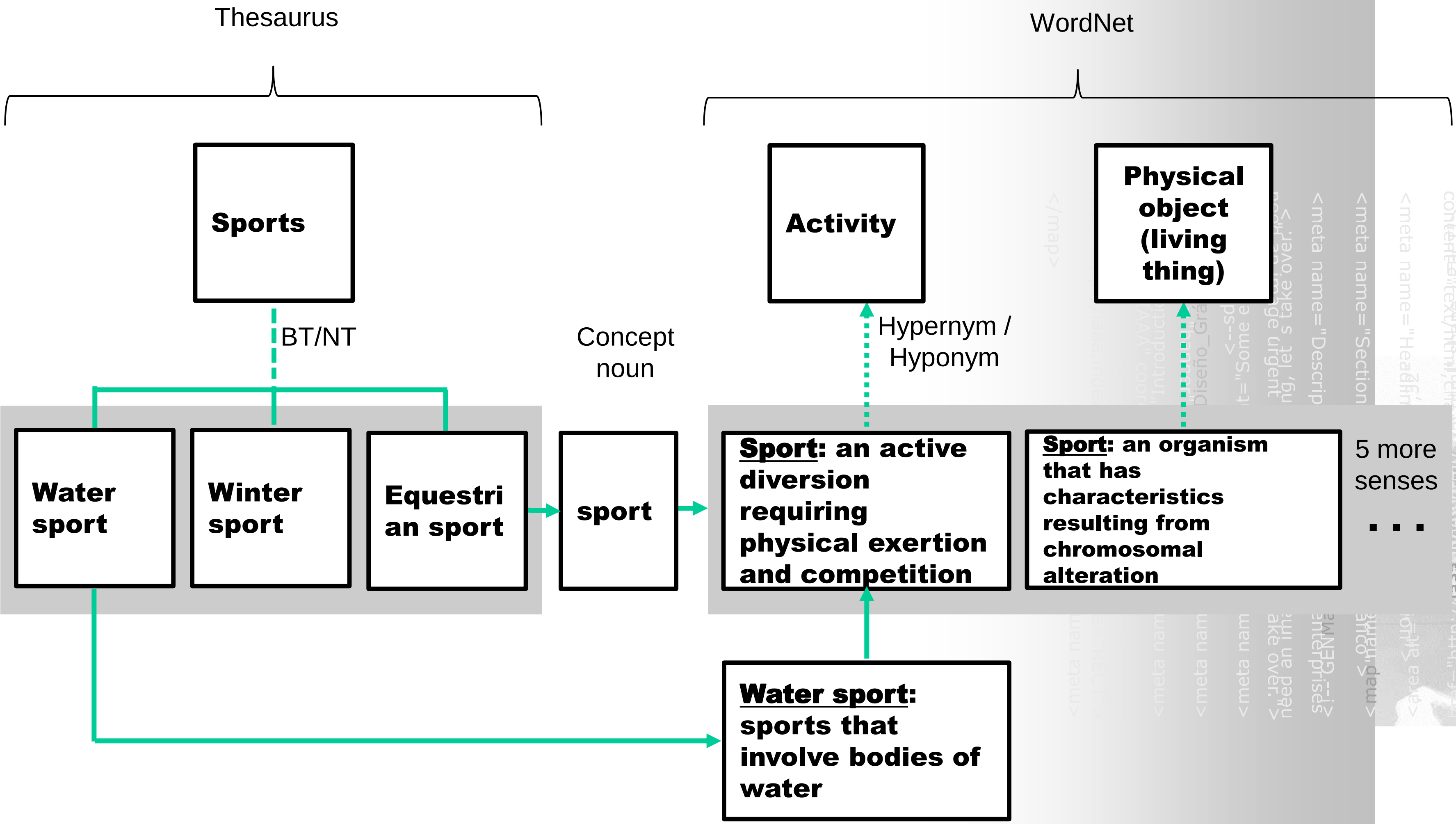
School: a large group of fish

4 more senses

. . .

Thesaurus context based disambiguation

- If we find a monosemic matching in Wordnet, we use it to decide matching for related thesaurus concepts



Thesaurus

WordNet

Sports

BT/NT

Water sport

Winter sport

Equestrian sport

Concept noun

sport

Activity

Physical object (living thing)

Hypernym / Hyponym

Sport: an active diversion requiring physical exertion and competition

Sport: an organism that has characteristics resulting from chromosomal alteration

5 more senses

. . .

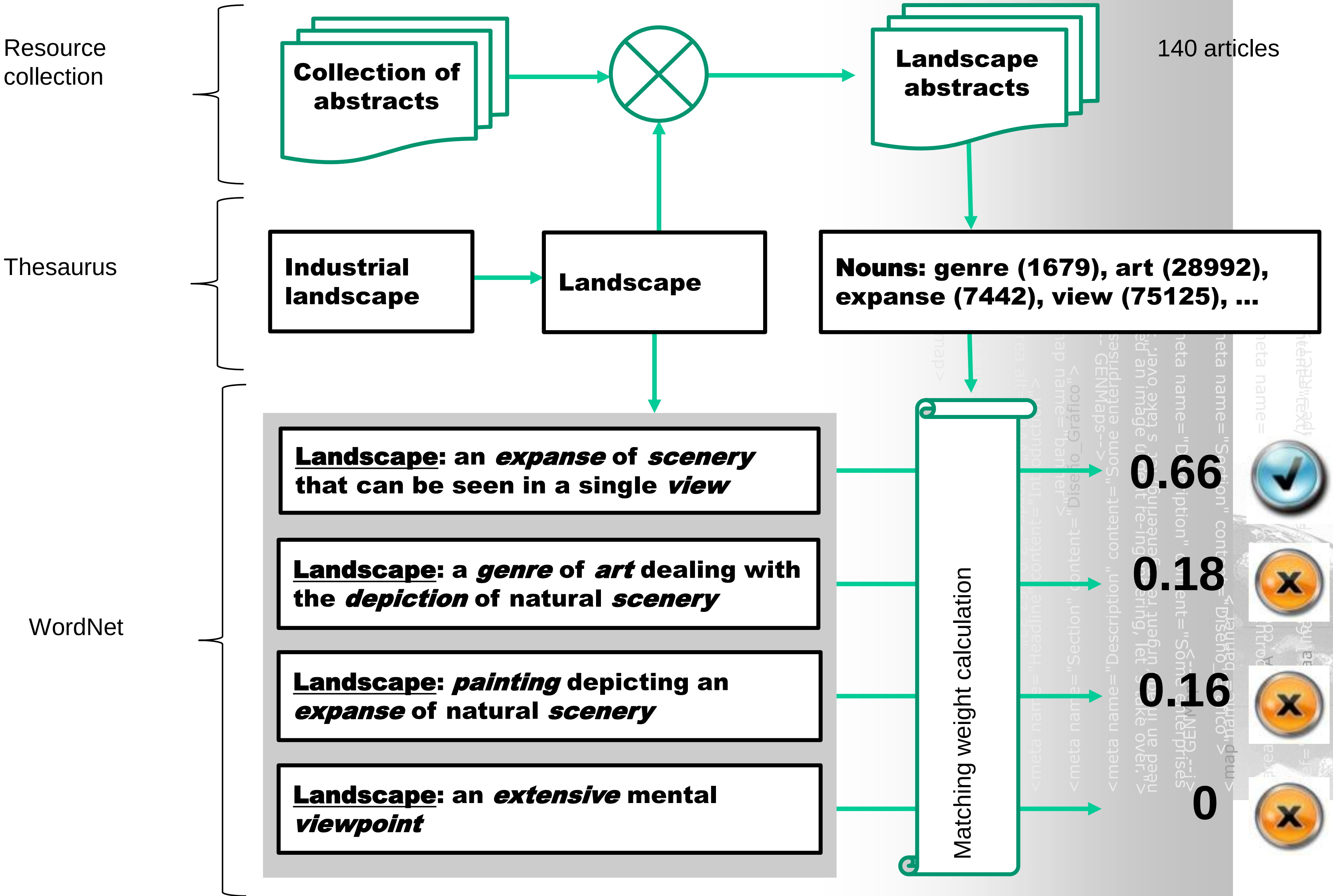
Water sport: sports that involve bodies of water

Resource collection based disambiguation (I)

- We use the abstracts (articles) classified with the thesaurus concept as context for the disambiguation
- Idea: An abstract classified according to a thesaurus concept contain terms (nouns) thematically related to the concept.
  - These nouns can be used to identify the intended meaning of the thesaurus concept
  - They may be contained in the definitions of the possible synsets
- Similarity is measured in a similar way to query-document relevance in vector-space information retrieval model

$$Sim(s,c) = \frac{\sum_{n_i \in SN(s) \cap AN(c)} (occur(n_i, SN(s)) * occur(n_i, AN(c)))}{\sqrt{\sum_{n_i \in SN(s)} (occur(n_i, SN(s))^2)} * \sqrt{\sum_{n_i \in AN(c)} (occur(n_i, AN(c))^2)}}$$

Resource collection based disambiguation (II)



Resource collection

Collection of abstracts

Landscape abstracts

140 articles

Thesaurus

Industrial landscape

Landscape

Nouns: genre (1679), art (28992), expanse (7442), view (75125), …

WordNet

Landscape: an ***expanse*** of ***scenery*** that can be seen in a single ***view***

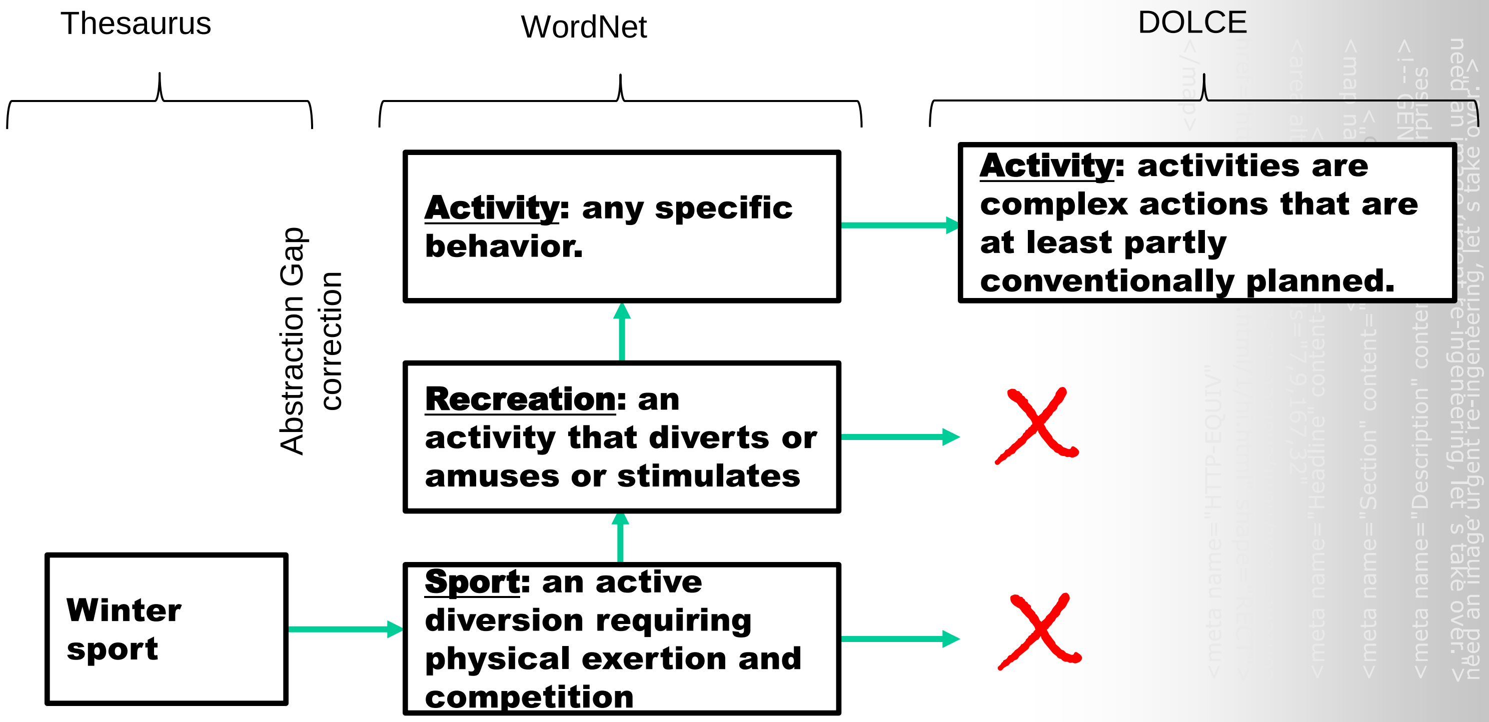
Landscape: a ***genre*** of ***art*** dealing with the ***depiction*** of natural ***scenery***

Landscape: ***painting*** depicting an ***expanse*** of natural ***scenery***

Landscape: an ***extensive*** mental ***viewpoint***

Identification of DOLCE equivalences

- **First, we have defined a lexical mapping between Wordnet and DOLCE.**
- **Using it and the Wordnet hierarchy the DOLCE concepts can be automatically assigned as superclasses of thesaurus concepts**



Winter sport

Sport: an active diversion requiring physical exertion and competition

Recreation: an activity that diverts or amuses or stimulates

Activity: any specific behavior.

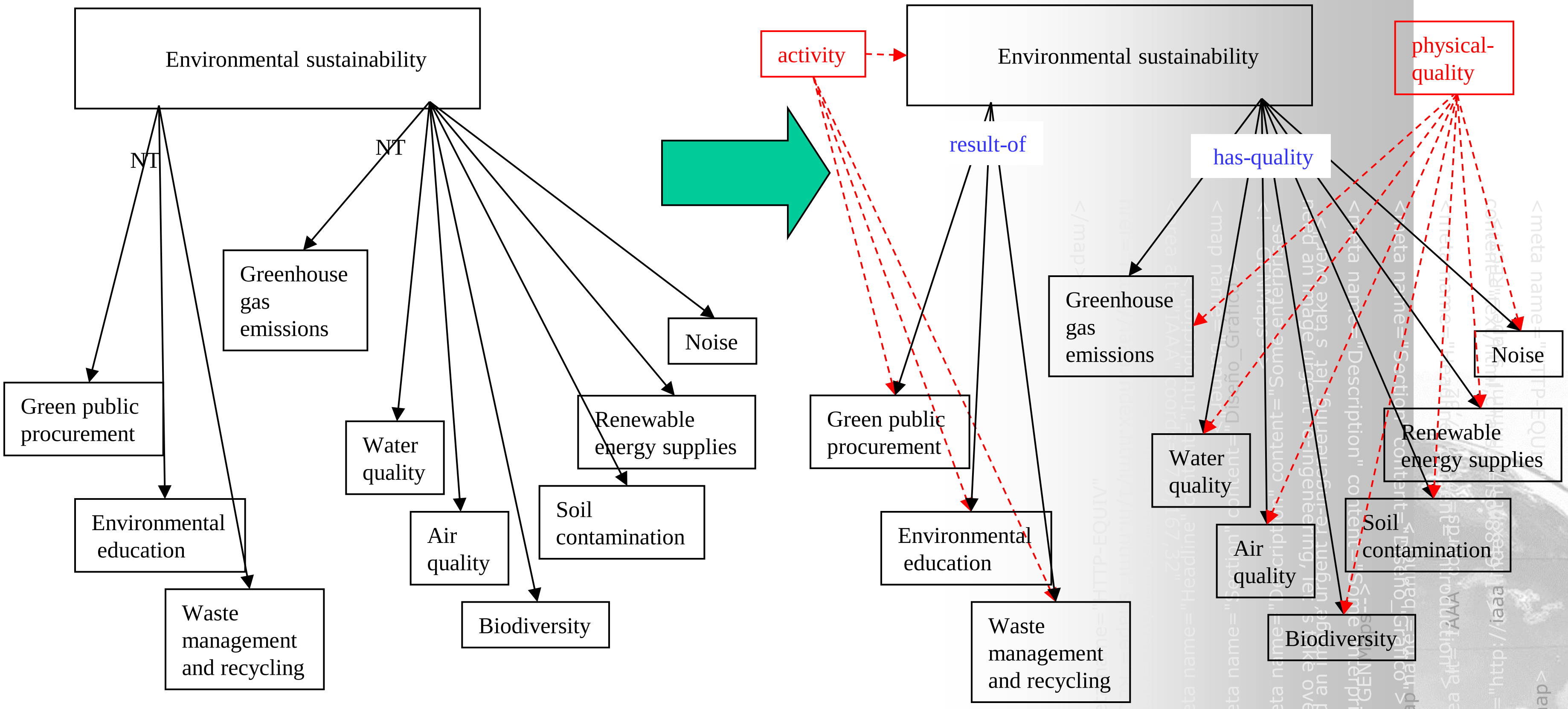
Activity: activities are complex actions that are at least partly conventionally planned.

Relations refinement

- DOLCE may provide several relations between two classes
  - Definition of inference rules

Pairs of DOLCE classes identified as superclasses of two concepts holding a BT/NT relation	Inferred relation
(activity → physical/abstract-quality) (geographical/physical/information-object → abstract-quality) (rational-agent → abstract-quality) (regulation → abstract-quality) (plan → abstract-quality) (physical-quality → abstract-quality) (physical-quality → physical-quality)	has-quality
(activity → rational-agent) (activity → information/physical-object) (activity → regulation) (activity → principle) (phenomenon → geographic-object)	participant
(abstract-quality → abstract-quality) (activity → plan) (phenomenon → activity) (geographic-object → geographic-object) (regulation → plan)	part
(plan → activity) (rational-agent → information-object) (rational-agent → physical-object) (rational-agent → plan) (norm → system-design)	generic-dependent
(geographical-object → physical-object) (rational-agent → rational-agent) (regulation → regulation) (information-object → information-object)	subclass-of
(physical-object → activity) (physical-object → plan)	instrument-of
(activity → activity)	result-of

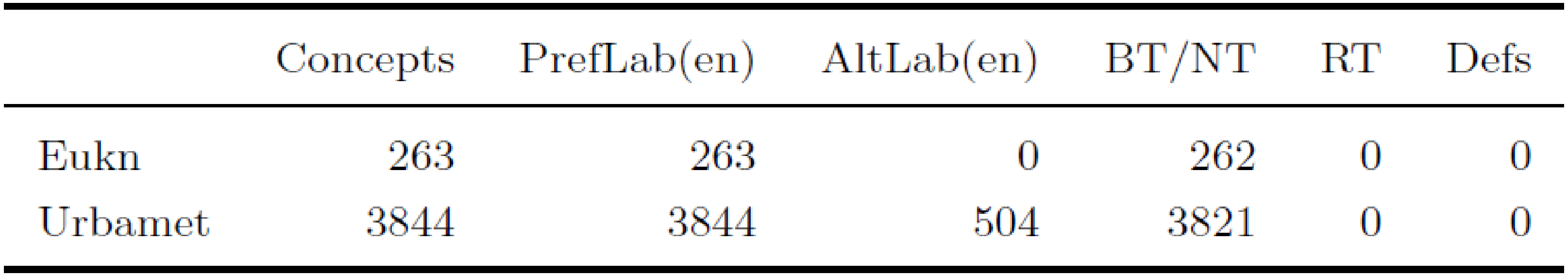
An example of relations refinement



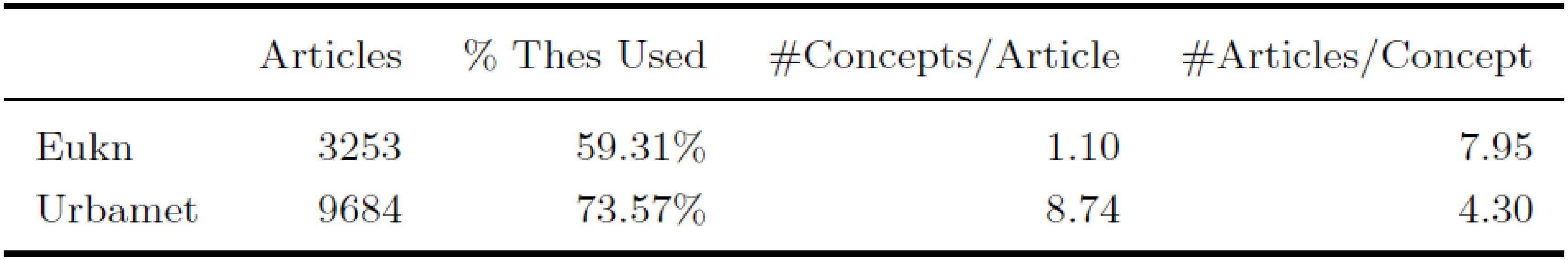
Experiments and tests on the formalization process

- Collection of resources in the European Knowledge Network (EUKN) and its associated thesaurus
- URBAMET bibliographic database (2005-2006) and its associated thesaurus
  - Reviewed 208 concepts of the “urban planning development” branch

Table 2: Comparison of Urbamet and EUKN thesaurus



	Concepts	PrefLab(en)	AltLab(en)	BT/NT	RT	Defs
Eukn	263	263	0	262	0	0
Urbamet	3844	3844	504	3821	0	0



	Articles	% Thes Used	#Concepts/Article	#Articles/Concept
Eukn	3253	59.31%	1.10	7.95
Urbamet	9684	73.57%	8.74	4.30

Results: Thesaurus – Wordnet mapping

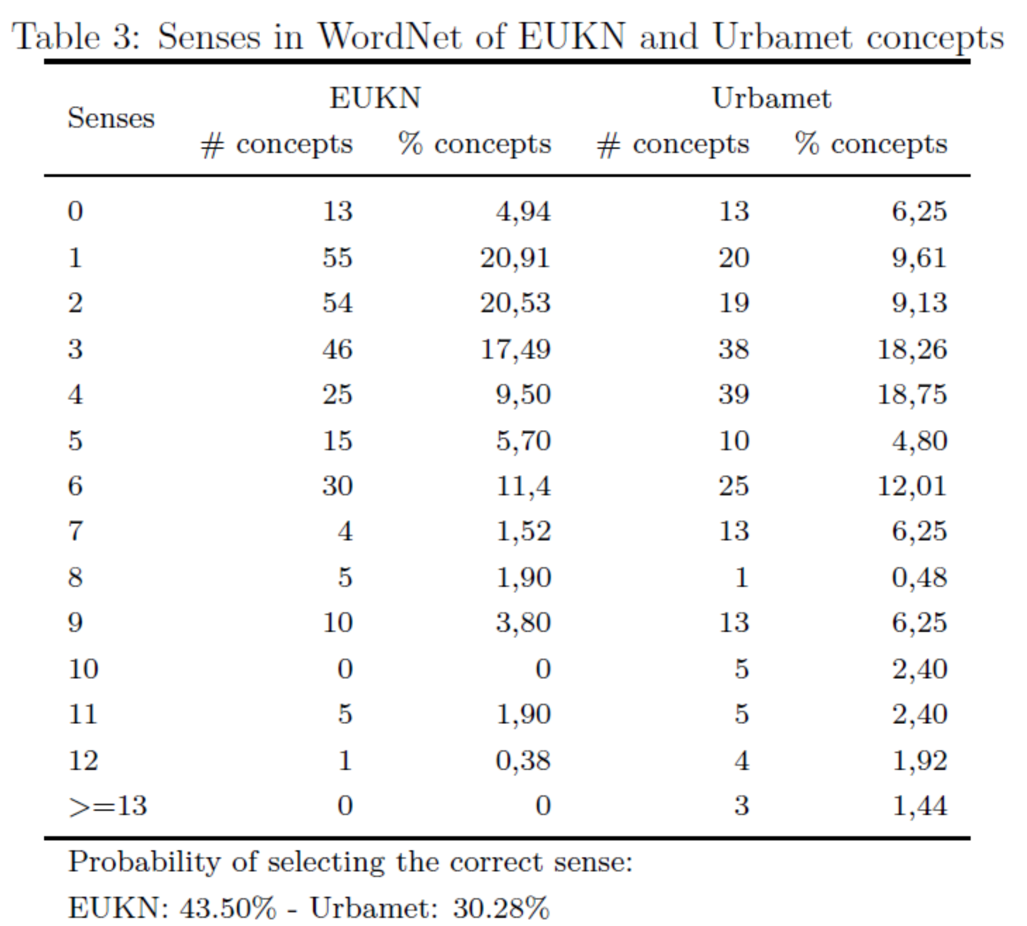


Table 3: Senses in WordNet of EUKN and Urbamet concepts

Senses	EUKN		Urbamet	
	# concepts	% concepts	# concepts	% concepts
0	13	4,94	13	6,25
1	55	20,91	20	9,61
2	54	20,53	19	9,13
3	46	17,49	38	18,26
4	25	9,50	39	18,75
5	15	5,70	10	4,80
6	30	11,4	25	12,01
7	4	1,52	13	6,25
8	5	1,90	1	0,48
9	10	3,80	13	6,25
10	0	0	5	2,40
11	5	1,90	5	2,40
12	1	0,38	4	1,92
>=13	0	0	3	1,44

Probability of selecting the correct sense:
EUKN: 43.50% - Urbamet: 30.28%

- ❑ An increase in alignment coverage
- ❑ An increase in precision with respect to probability of assigning correct sense

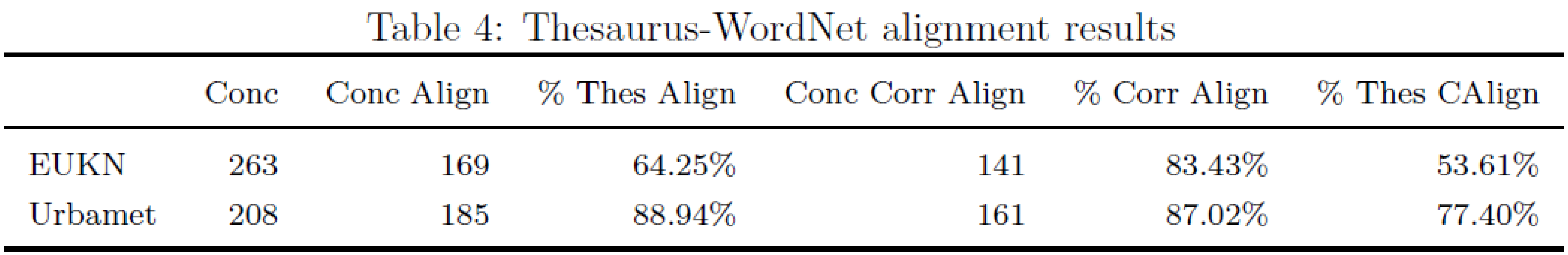


Table 4: Thesaurus-WordNet alignment results

	Conc	Conc Align	% Thes Align	Conc Corr Align	% Corr Align	% Thes CAlign
EUKN	263	169	64.25%	141	83.43%	53.61%
Urbamet	208	185	88.94%	161	87.02%	77.40%

Results: WordNet-DOLCE alignment

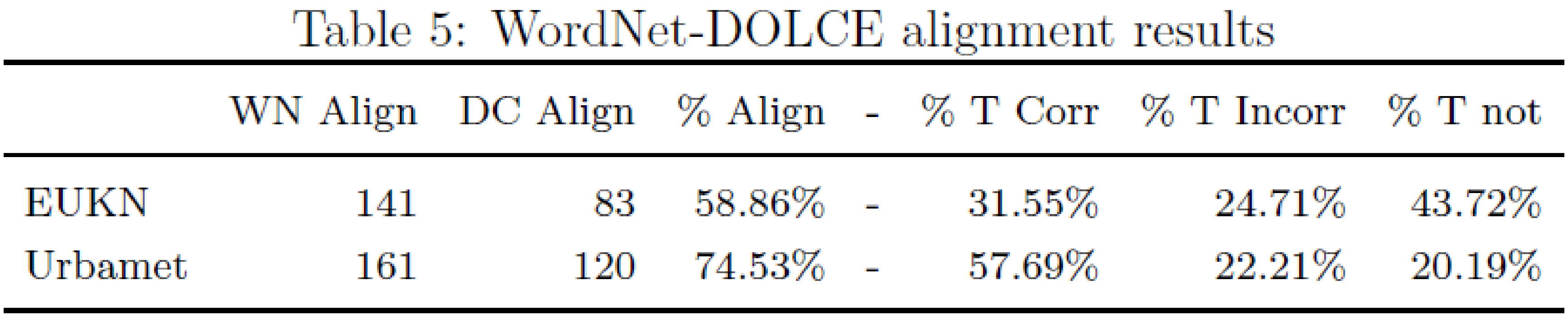


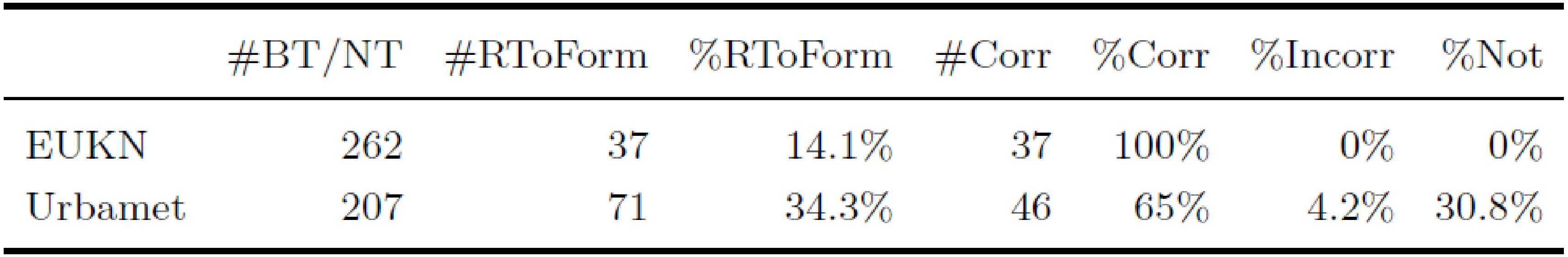
Table 5: WordNet-DOLCE alignment results

	WN Align	DC Align	% Align	-	% T Corr	% T Incorr	% T not
EUKN	141	83	58.86%	-	31.55%	24.71%	43.72%
Urbamet	161	120	74.53%	-	57.69%	22.21%	20.19%

- **Why UBAMET results are much better than EUKN?**
  - **EUKN concepts are matched with WordNet areas with worse DOLCE alignment**
  - **EUKN thesaurus concepts are more complex**
    - **Multiple concept terms, difficult to align with WordNet**
  - **40% of EUKN concepts have been never used for classification of resources**
    - **the disambiguation context isn’t so rich as in URBAMET**

Relations refinement

Table 6: Relations refinement

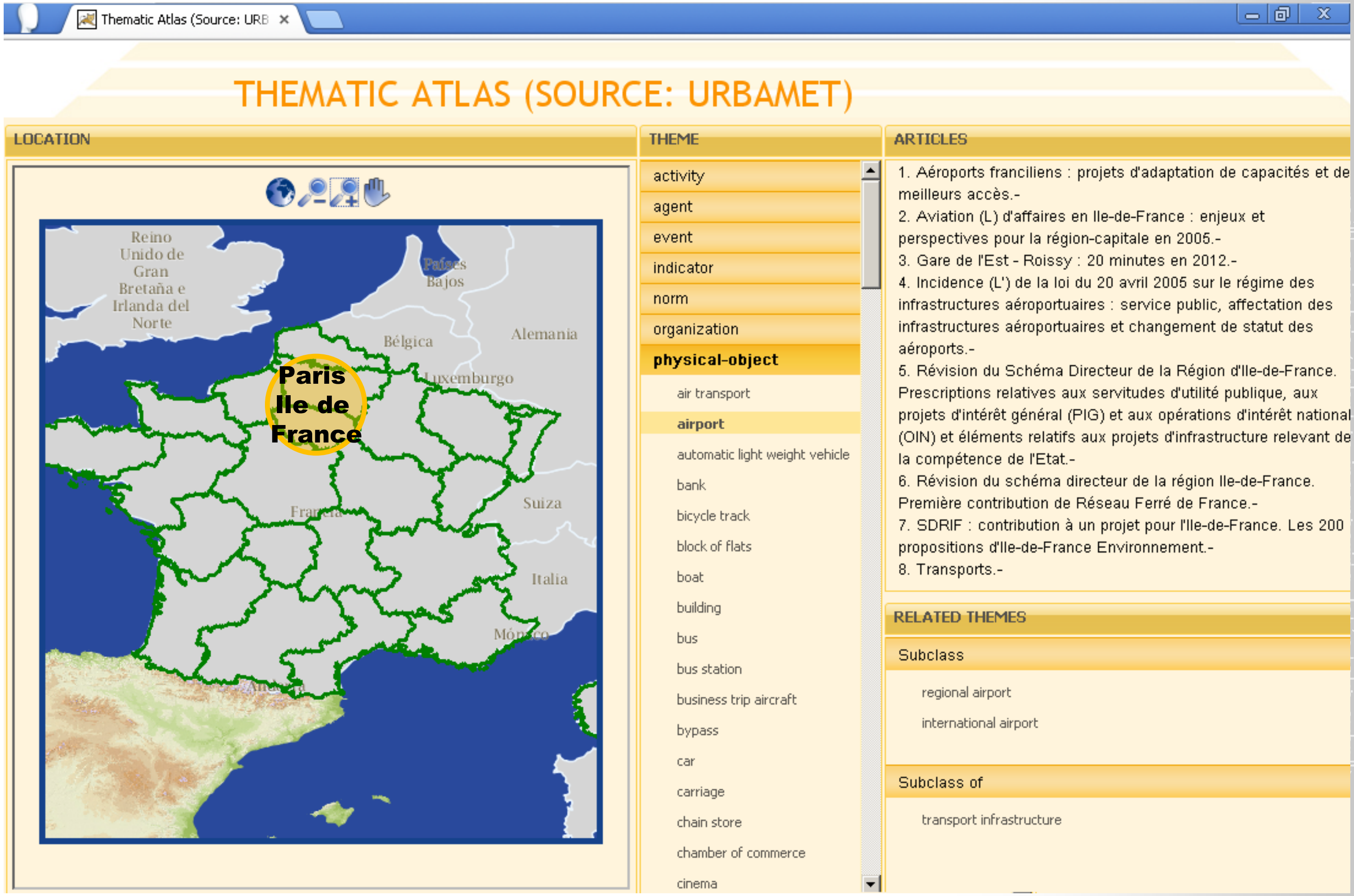


	#BT/NT	#RToForm	%RToForm	#Corr	%Corr	%Incorr	%Not
EUKN	262	37	14.1%	37	100%	0%	0%
Urbamet	207	71	34.3%	46	65%	4.2%	30.8%

- **The refinement of relations requires**
  - **The two concepts involved in the relation have been correctly matched to DOLCE**
  - **There is a relation in DOLCE between the matched concepts**
- **Fewer relations than expected fullfill these restrictions**
- **The quality of the assignements is high**

Applicability: transformation of a bibliographic database into a semantic repository

- Browse the bibliographic database as a thematic atlas
  - Exploiting themes and location of bibliographic records



THEMATIC ATLAS (SOURCE: URBAMET)

LOCATION

THEME

- activity
- agent
- event
- indicator
- norm
- organization
- **physical-object**
  - air transport
  - **airport**
  - automatic light weight vehicle
  - bank
  - bicycle track
  - block of flats
  - boat
  - building
  - bus
  - bus station
  - business trip aircraft
  - bypass
  - car
  - carriage
  - chain store
  - chamber of commerce
  - cinema

ARTICLES

1. Aéroports franciliens : projets d'adaptation de capacités et de meilleurs accès.-
2. Aviation (L) d'affaires en Ile-de-France : enjeux et perspectives pour la région-capitale en 2005.-
3. Gare de l'Est - Roissy : 20 minutes en 2012.-
4. Incidence (L') de la loi du 20 avril 2005 sur le régime des infrastructures aéroportuaires : service public, affectation des infrastructures aéroportuaires et changement de statut des aéroports.-
5. Révision du Schéma Directeur de la Région d'Ile-de-France. Prescriptions relatives aux servitudes d'utilité publique, aux projets d'intérêt général (PIG) et aux opérations d'intérêt national (OIN) et éléments relatifs aux projets d'infrastructure relevant de la compétence de l'Etat.-
6. Révision du schéma directeur de la région Ile-de-France. Première contribution de Réseau Ferré de France.-
7. SDRIF : contribution à un projet pour l'Ile-de-France. Les 200 propositions d'Ile-de-France Environnement.-
8. Transports.-

RELATED THEMES

Subclass

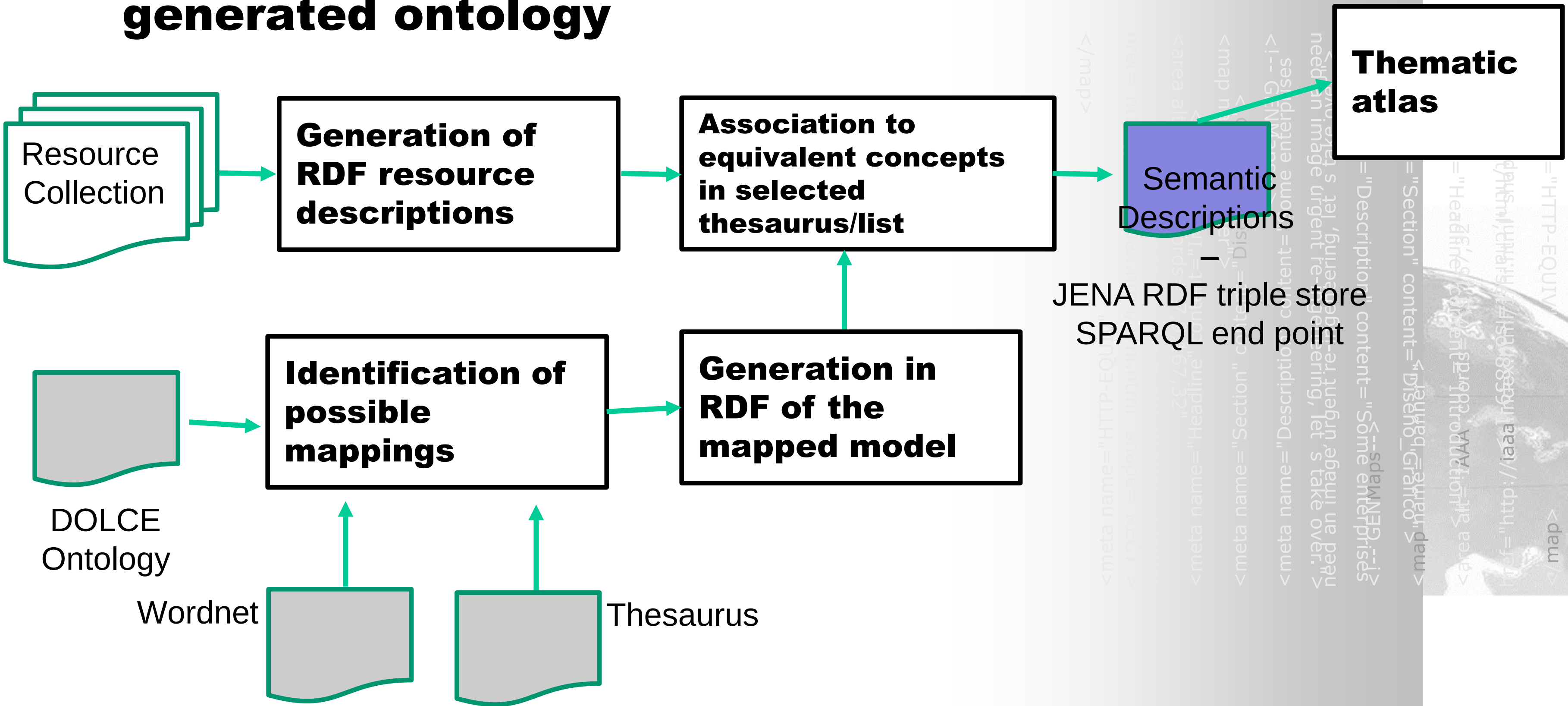
- regional airport
- international airport

Subclass of

- transport infrastructure

How can we create this semantic repository?

- Conversion of the collection descriptions to RDF (Dublin Core)
- Transform the thesaurus used for classification into an ontology
- Link the terms in the collection descriptions with the generated ontology



Resource Collection

Generation of RDF resource descriptions

Association to equivalent concepts in selected thesaurus/list

Semantic Descriptions

–

JENA RDF triple store
SPARQL end point

Thematic atlas

DOLCE Ontology

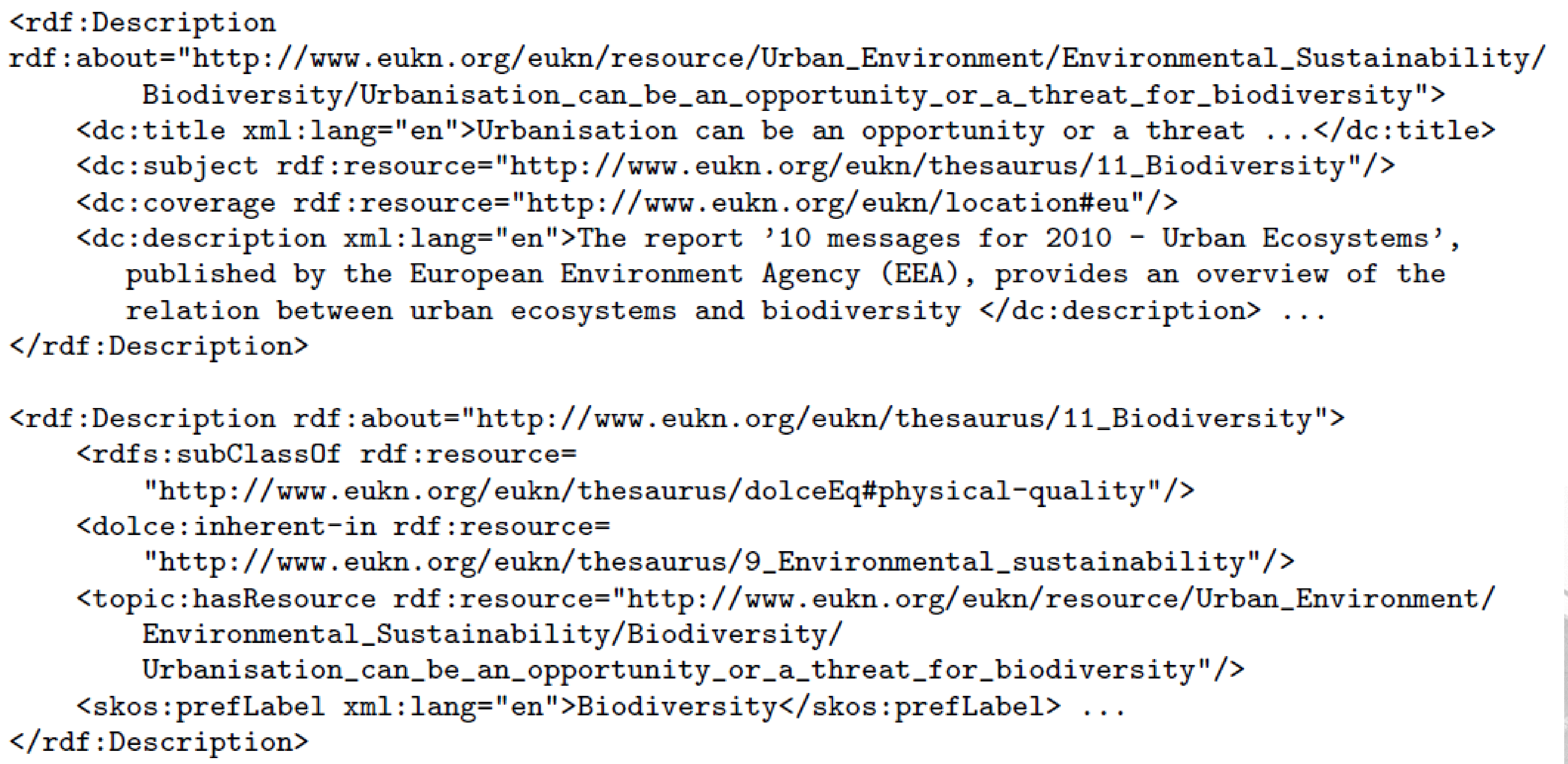
Identification of possible mappings

Generation in RDF of the mapped model

Wordnet

Thesaurus

Example of mapped model

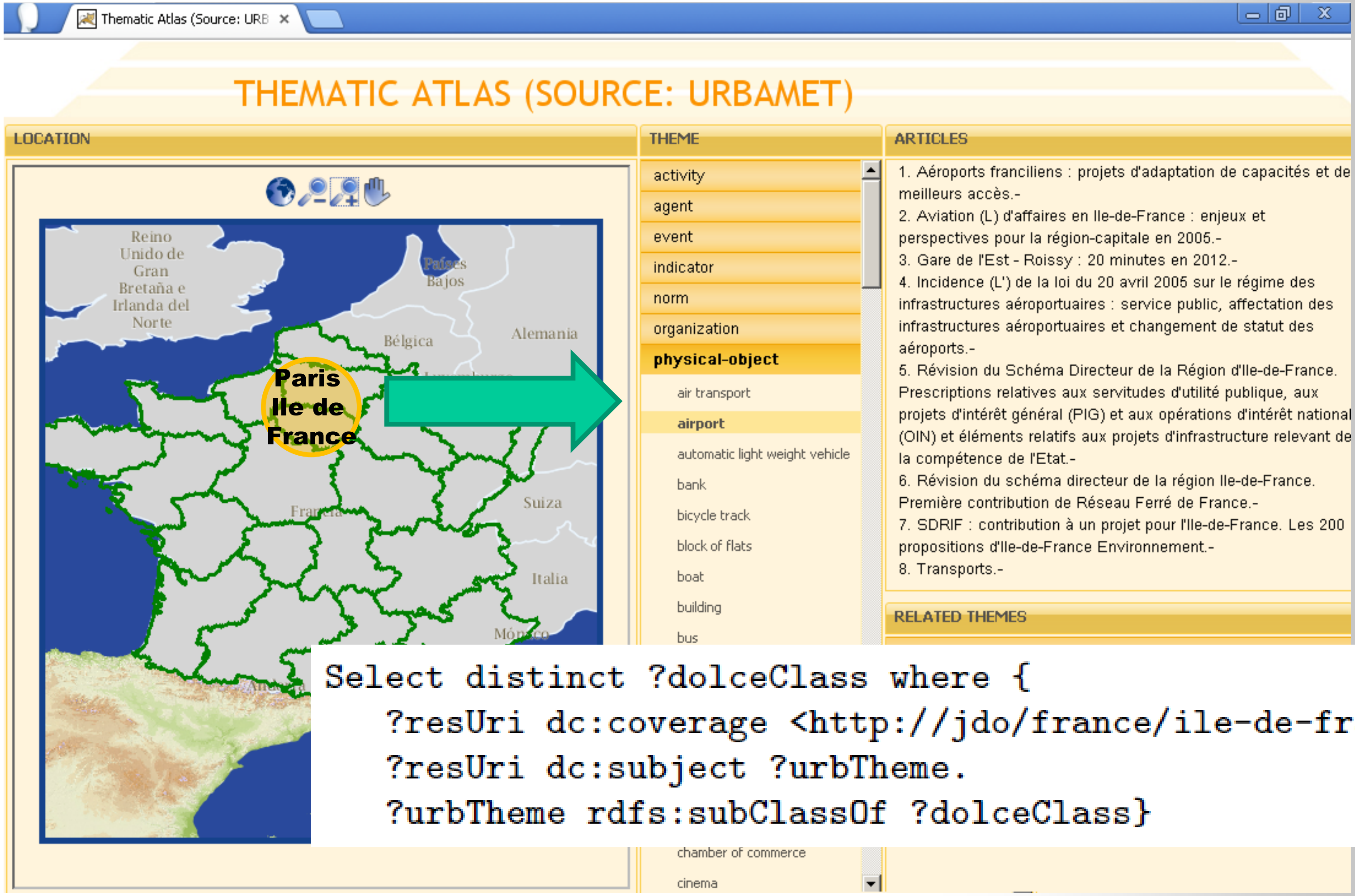


```
<rdf:Description
rdf:about="http://www.eukn.org/eukn/resource/Urban_Environment/Environmental_Sustainability/
        Biodiversity/Urbanisation_can_be_an_opportunity_or_a_threat_for_biodiversity">
    <dc:title xml:lang="en">Urbanisation can be an opportunity or a threat ...</dc:title>
    <dc:subject rdf:resource="http://www.eukn.org/eukn/thesaurus/11_Biodiversity"/>
    <dc:coverage rdf:resource="http://www.eukn.org/eukn/location#eu"/>
    <dc:description xml:lang="en">The report '10 messages for 2010 - Urban Ecosystems',
       published by the European Environment Agency (EEA), provides an overview of the
       relation between urban ecosystems and biodiversity </dc:description> ...
</rdf:Description>

<rdf:Description rdf:about="http://www.eukn.org/eukn/thesaurus/11_Biodiversity">
    <rdfs:subClassOf rdf:resource=
        "http://www.eukn.org/eukn/thesaurus/dolceEq#physical-quality"/>
    <dolce:inherent-in rdf:resource=
        "http://www.eukn.org/eukn/thesaurus/9_Environmental_sustainability"/>
    <topic:hasResource rdf:resource="http://www.eukn.org/eukn/resource/Urban_Environment/
        Environmental_Sustainability/Biodiversity/
        Urbanisation_can_be_an_opportunity_or_a_threat_for_biodiversity"/>
    <skos:prefLabel xml:lang="en">Biodiversity</skos:prefLabel> ...
</rdf:Description>
```

How to build the thematic atlas?

- Take advantage of SPARQL and inference


Thematic Atlas (Source: URB
Reino Unido de Gran Bretaña e Irlanda del Norte
Países Bajos
Bélgica
Alemania
Paris
Ile de
France
Francia
Suiza
Italia
Mónaco

THEMATIC ATLAS (SOURCE: URBAMET)

LOCATION

THEME

activity
agent
event
indicator
norm
organization
physical-object
air transport
airport
automatic light weight vehicle
bank
bicycle track
block of flats
boat
building
bus
chamber of commerce
cinema

ARTICLES

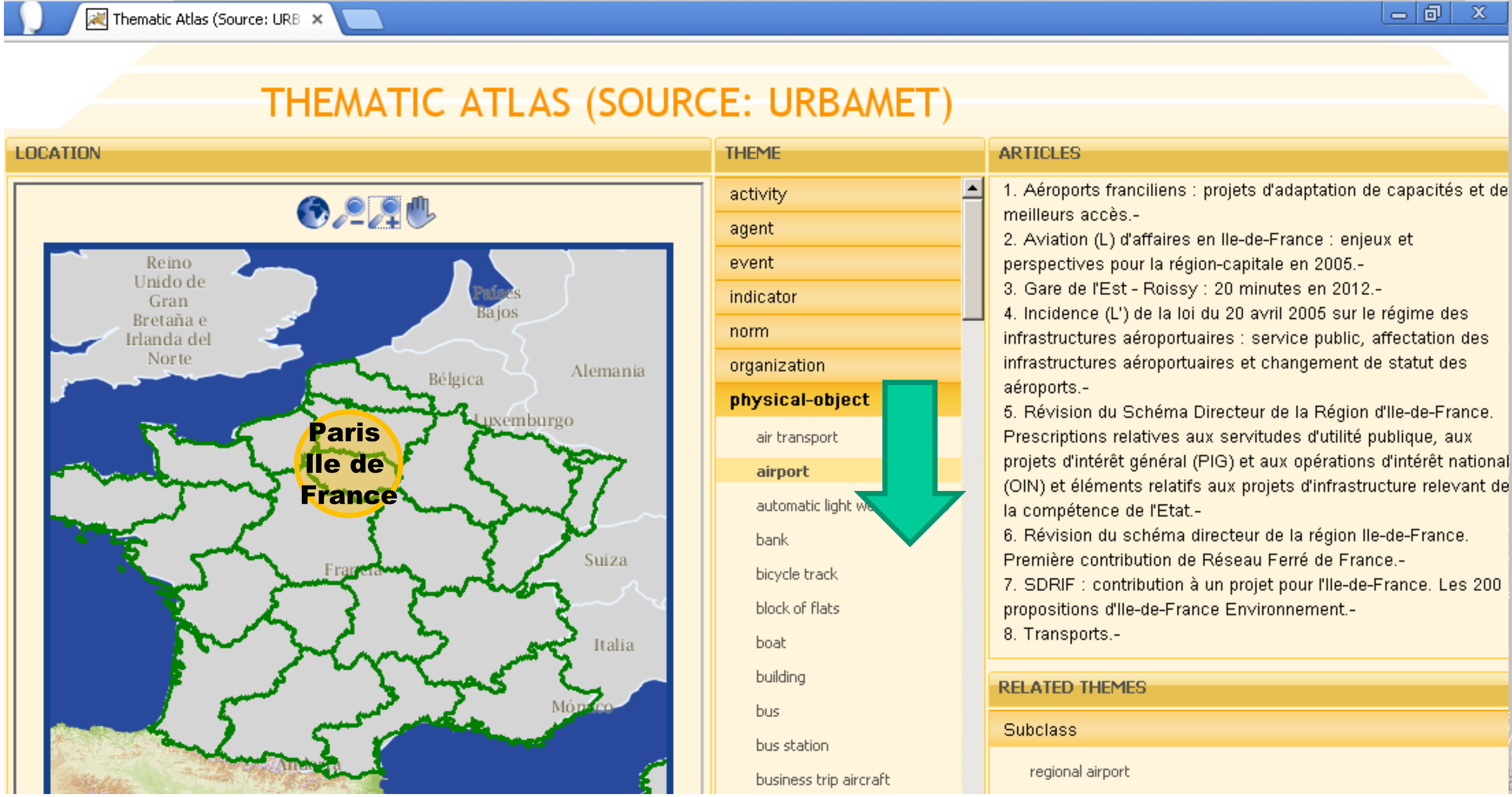
1. Aéroports franciliens : projets d'adaptation de capacités et de meilleurs accès.-
2. Aviation (L) d'affaires en Ile-de-France : enjeux et perspectives pour la région-capitale en 2005.-
3. Gare de l'Est - Roissy : 20 minutes en 2012.-
4. Incidence (L') de la loi du 20 avril 2005 sur le régime des infrastructures aéroportuaires : service public, affectation des infrastructures aéroportuaires et changement de statut des aéroports.-
5. Révision du Schéma Directeur de la Région d'Ile-de-France. Prescriptions relatives aux servitudes d'utilité publique, aux projets d'intérêt général (PIG) et aux opérations d'intérêt national (OIN) et éléments relatifs aux projets d'infrastructure relevant de la compétence de l'Etat.-
6. Révision du schéma directeur de la région Ile-de-France. Première contribution de Réseau Ferré de France.-
7. SDRIF : contribution à un projet pour l'Ile-de-France. Les 200 propositions d'Ile-de-France Environnement.-
8. Transports.-

RELATED THEMES

```
Select distinct ?dolceClass where {
   ?resUri dc:coverage <http://jdo/france/ile-de-france>.
   ?resUri dc:subject ?urbTheme.
   ?urbTheme rdfs:subClassOf ?dolceClass}
```

How to build the thematic atlas?

- Take advantage of SPARQL and inference

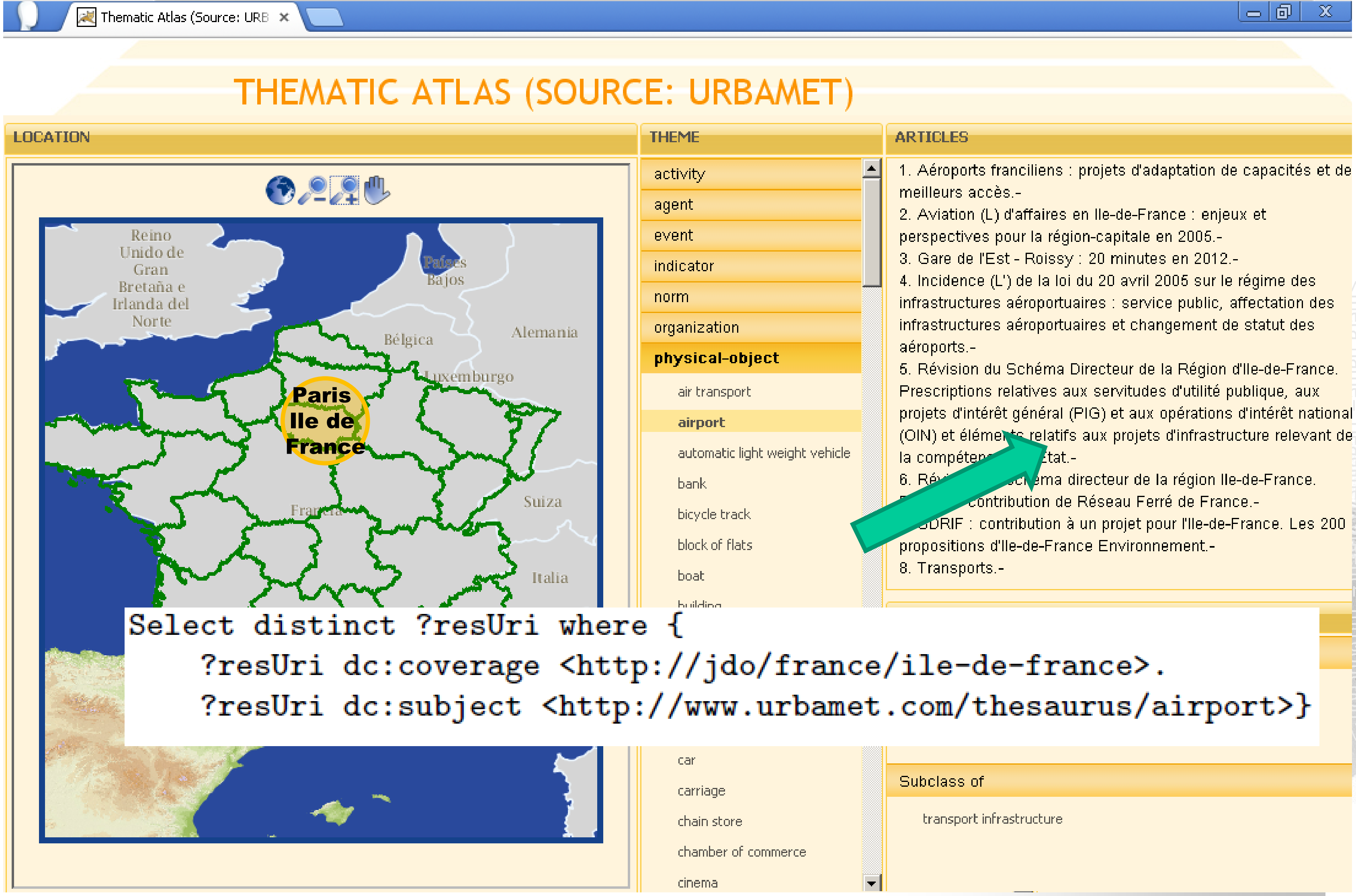


1. Aéroports franciliens : projets d'adaptation de capacités et de meilleurs accès.-
2. Aviation (L) d'affaires en Ile-de-France : enjeux et perspectives pour la région-capitale en 2005.-
3. Gare de l'Est - Roissy : 20 minutes en 2012.-
4. Incidence (L') de la loi du 20 avril 2005 sur le régime des infrastructures aéroportuaires : service public, affectation des infrastructures aéroportuaires et changement de statut des aéroports.-
5. Révision du Schéma Directeur de la Région d'Ile-de-France. Prescriptions relatives aux servitudes d'utilité publique, aux projets d'intérêt général (PIG) et aux opérations d'intérêt national (OIN) et éléments relatifs aux projets d'infrastructure relevant de la compétence de l'Etat.-
6. Révision du schéma directeur de la région Ile-de-France. Première contribution de Réseau Ferré de France.-
7. SDRIF : contribution à un projet pour l'Ile-de-France. Les 200 propositions d'Ile-de-France Environnement.-
8. Transports.-

```
Select distinct ?urbTheme where {
   ?resUri dc:coverage <http://jdo/france/ile-de-france>.
   ?resUri dc:subject ?urbTheme.
   ?urbTheme rdfs:subClassOf <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#physical-object>}
```

How to build the thematic atlas?

- ❑ Take advantage of SPARQL and inference



Thematic Atlas (Source: URB

THEMATIC ATLAS (SOURCE: URBAMET)

LOCATION

Reino Unido de Gran Bretaña e Irlanda del Norte
Países Bajos
Bélgica
Alemania
Luxemburgo
Paris Ile de France
Francia
Suiza
Italia

THEME

activity
agent
event
indicator
norm
organization
physical-object
air transport
airport
automatic light weight vehicle
bank
bicycle track
block of flats
boat
building
car
carriage
chain store
chamber of commerce
cinema

ARTICLES

1. Aéroports franciliens : projets d'adaptation de capacités et de meilleurs accès.-
2. Aviation (L) d'affaires en Ile-de-France : enjeux et perspectives pour la région-capitale en 2005.-
3. Gare de l'Est - Roissy : 20 minutes en 2012.-
4. Incidence (L') de la loi du 20 avril 2005 sur le régime des infrastructures aéroportuaires : service public, affectation des infrastructures aéroportuaires et changement de statut des aéroports.-
5. Révision du Schéma Directeur de la Région d'Ile-de-France. Prescriptions relatives aux servitudes d'utilité publique, aux projets d'intérêt général (PIG) et aux opérations d'intérêt national (OIN) et éléments relatifs aux projets d'infrastructure relevant de la compétence de l'État.-
6. Révision du schéma directeur de la région Ile-de-France. [illegible] contribution de Réseau Ferré de France.-
7. SDRIF : contribution à un projet pour l'Ile-de-France. Les 200 propositions d'Ile-de-France Environnement.-
8. Transports.-

Subclass of

transport infrastructure

```
Select distinct ?resUri where {
    ?resUri dc:coverage <http://jdo/france/ile-de-france>.
    ?resUri dc:subject <http://www.urbamet.com/thesaurus/airport>}
```

Conclusions and future work (I)

- We have presented a method to increase the formalism of thesauri
  - Experiments with URBAMET and EUKN
- Possible improvements
  - Thesaurus – WordNet alignment
    - WordNet is only available in English
      - Pb. with thesauri or bibliographic database in other languages
      - Consider EuroWordnet or other ontological resources
    - Needed of improvements in the disambiguation steps
  - WordNet – Dolce alignment
    - Improve coverage of the WordNet – Dolce alignment
    - Extend Dolce with additional relations

Conclusions and future work (II)

- We have shown that an ontology could help to create a semantic repository,
  - Allow the construction of better applications
  - Facilitate other perspectives: a thematic atlas
- Issues to improve in the semantic repository
  - Integrate other knowledge models such as
    - Temporal ontologies
    - Authority information (VIAF = International Virtual Authority File)

Advanced Information Systems Laboratory

http://iaaa.unizar.es