Bibliographic databases, an ontological perspective

Javier Lacasta, Javier Nogueras-Iso, Gilles Falquet, Jacques Teller, F. Javier Zarazaga-Soria
20 June 2012, OAI8, Genève
Advanced Information Systems Laboratory (IAAA)


- 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 and 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
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
An example of the sense disambiguation problem

Thesaurus

Educational building

BT/NT

Concept noun

Primary school

High school

School

WordNet

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

School: an educational institution

School: the process of being formally educated at a school

School: a large group of fish

4 more senses
If we find a monosemic matching in Wordnet, we use it to decide matching for related thesaurus concepts.

Thesaurus context based disambiguation:

- **Sports**
- **Activity**
- **Physical object (living thing)**

- **Water sport**
- **Winter sport**
- **Equestrian sport**
- **Sport**: an active diversion requiring physical exertion and competition
- **Sport**: an organism that has characteristics resulting from chromosomal alteration
- **Water sport**: sports that involve bodies of water

5 more senses
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.

\[
\text{Sim}(s, c) = \frac{\sum_{n_i \in SN(s) \cap AN(c)} \left( \text{ocurr}(n_i, SN(s)) \times \text{ocurr}(n_i, AN(c)) \right)}{\sqrt{\sum_{n_i \in SN(s)} \left( \text{ocurr}(n_i, SN(s))^2 \right)} \times \sqrt{\sum_{n_i \in AN(c)} \left( \text{ocurr}(n_i, AN(c))^2 \right)}}
\]
Resource collection based disambiguation (II)

Resource collection

Collection of abstracts

Landscape abstracts

Thesaurus

Industrial landscape

Landscape

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

WordNet

Landscape: an expance 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

Matching weight calculation

0.66

0.18

0.16

0

140 articles
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.
## Relations refinement

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

<table>
<thead>
<tr>
<th>Pairs of DOLCE classes identified as superclasses of two concepts holding a BT/NT relation</th>
<th>Inferred relation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(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)</td>
<td>has-quality</td>
</tr>
<tr>
<td>(activity → rational-agent) (activity → information/physical-object) (activity → regulation) (activity → principle) (phenomenon → geographic-object)</td>
<td>participant</td>
</tr>
<tr>
<td>(abstract-quality → abstract-quality) (activity → plan) (phenomenon → activity) (geographic-object → geographic-object) (regulation → plan)</td>
<td>part</td>
</tr>
<tr>
<td>(plan → activity) (rational-agent → information-object) (rational-agent → physical-object) (rational-agent → plan) (norm → system-design)</td>
<td>generic-dependent</td>
</tr>
<tr>
<td>(geographical-object → physical-object) (rational-agent → rational-agent) (regulation → regulation) (information-object → information-object)</td>
<td>subclass-of</td>
</tr>
<tr>
<td>(physical-object → activity) (physical-object → plan)</td>
<td>instrument-of</td>
</tr>
<tr>
<td>(activity → activity)</td>
<td>result-of</td>
</tr>
</tbody>
</table>
An example of relations refinement

Environmental sustainability

- Green public procurement
  - Greenhouse gas emissions
  - Water quality
    - Renewable energy supplies
      - Noise
  - Air quality
    - Soil contamination
  - Biodiversity
- Waste management and recycling
- Environmental education

Environmental sustainability

- Activity
  - Result-of
  - Has-quality
    - Greenhouse gas emissions
    - Water quality
    - Environmental education
    - Waste management and recycling
    - Biodiversity

Physical quality

- Noise
- Renewable energy supplies
- Soil contamination
- Air quality
- Water quality
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

<table>
<thead>
<tr>
<th></th>
<th>Concepts</th>
<th>PrefLab(en)</th>
<th>AltLab(en)</th>
<th>BT/NT</th>
<th>RT</th>
<th>Defs</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUKN</td>
<td>263</td>
<td>263</td>
<td>0</td>
<td>262</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Urbamet</td>
<td>3844</td>
<td>3844</td>
<td>504</td>
<td>3821</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

### Table 2: Comparison of Urbamet and EUKN thesaurus

<table>
<thead>
<tr>
<th></th>
<th>Articles</th>
<th>% Thes Used</th>
<th>#Concepts/Article</th>
<th>#Articles/Concept</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUKN</td>
<td>3253</td>
<td>59.31%</td>
<td>1.10</td>
<td>7.95</td>
</tr>
<tr>
<td>Urbamet</td>
<td>9684</td>
<td>73.57%</td>
<td>8.74</td>
<td>4.30</td>
</tr>
</tbody>
</table>
Results: Thesaurus – Wordnet mapping

Table 3: Senses in WordNet of EUKN and Urbamet concepts

<table>
<thead>
<tr>
<th>Senses</th>
<th>EUKN # concepts</th>
<th>EUKN % concepts</th>
<th>Urbamet # concepts</th>
<th>Urbamet % concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13</td>
<td>4.94</td>
<td>13</td>
<td>6.25</td>
</tr>
<tr>
<td>1</td>
<td>55</td>
<td>20.91</td>
<td>20</td>
<td>9.61</td>
</tr>
<tr>
<td>2</td>
<td>54</td>
<td>20.53</td>
<td>19</td>
<td>9.13</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>17.49</td>
<td>38</td>
<td>18.26</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>9.50</td>
<td>39</td>
<td>18.75</td>
</tr>
<tr>
<td>5</td>
<td>15</td>
<td>5.70</td>
<td>10</td>
<td>4.80</td>
</tr>
<tr>
<td>6</td>
<td>30</td>
<td>11.4</td>
<td>25</td>
<td>12.01</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>1.52</td>
<td>13</td>
<td>6.25</td>
</tr>
<tr>
<td>8</td>
<td>5</td>
<td>1.90</td>
<td>1</td>
<td>0.48</td>
</tr>
<tr>
<td>9</td>
<td>10</td>
<td>3.80</td>
<td>13</td>
<td>6.25</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>2.40</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>1.90</td>
<td>5</td>
<td>2.40</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>0.38</td>
<td>4</td>
<td>1.92</td>
</tr>
<tr>
<td>&gt;=13</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1.44</td>
</tr>
</tbody>
</table>

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

Table 4: Thesaurus-WordNet alignment results

<table>
<thead>
<tr>
<th></th>
<th>Conc</th>
<th>Conc Align</th>
<th>% Thes Align</th>
<th>Conc Corr Align</th>
<th>% Corr Align</th>
<th>% Thes CAAlign</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUKN</td>
<td>263</td>
<td>169</td>
<td>64.25%</td>
<td>141</td>
<td>83.43%</td>
<td>53.61%</td>
</tr>
<tr>
<td>Urbamet</td>
<td>208</td>
<td>185</td>
<td>88.94%</td>
<td>161</td>
<td>87.02%</td>
<td>77.40%</td>
</tr>
</tbody>
</table>
Results: WordNet-DOLCE alignment

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

Table 5: WordNet-DOLCE alignment results

<table>
<thead>
<tr>
<th></th>
<th>WN Align</th>
<th>DC Align</th>
<th>% Align</th>
<th>% T Corr</th>
<th>% T Incorr</th>
<th>% T not</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUKN</td>
<td>141</td>
<td>83</td>
<td>58.86%</td>
<td>31.55%</td>
<td>24.71%</td>
<td>43.72%</td>
</tr>
<tr>
<td>Urbamet</td>
<td>161</td>
<td>120</td>
<td>74.53%</td>
<td>57.69%</td>
<td>22.21%</td>
<td>20.19%</td>
</tr>
</tbody>
</table>
Relations refinement

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 fulfill these restrictions.
- The quality of the assignments is high.

### Table 6: Relations refinement

<table>
<thead>
<tr>
<th></th>
<th>#BT/NT</th>
<th>#RToForm</th>
<th>%RToForm</th>
<th>#Corr</th>
<th>%Corr</th>
<th>%Incorr</th>
<th>%Not</th>
</tr>
</thead>
<tbody>
<tr>
<td>EUKN</td>
<td>262</td>
<td>37</td>
<td>14.1%</td>
<td>37</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
</tr>
<tr>
<td>Urbamen</td>
<td>207</td>
<td>71</td>
<td>34.3%</td>
<td>46</td>
<td>65%</td>
<td>4.2%</td>
<td>30.8%</td>
</tr>
</tbody>
</table>

An ontological perspective ... 20-jun-13 18
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
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

**Diagram:**
- Resource Collection
  - Generation of RDF resource descriptions
  - Identification of possible mappings
  - DOLCE Ontology
- Thesaurus
- Wordnet
- Semantic Descriptions
  - Association to equivalent concepts in selected thesaurus/list
  - JENA RDF triple store SPARQL end point
  - Thematic atlas
- Generation in RDF of the mapped model
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: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"/>
   <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
How to build the thematic atlas?

Take advantage of SPARQL and inference

THETMATIC ATLAS (SOURCE: URBAMET)

Select distinct ?urbTheme where {
  ?urbTheme rdfs:subClassOf <http://www.loa-cnr.it/ontologies/DOLCE-Lite.owl#physical-object>}

1. Aéroports franciliens : projets d'adaptation de capacités et de meilleure accès.
5. Révision du Schéma Directeur de la Région Ile-de-France. Prescriptions relatives aux services d'utilité publique, aux projets d'intérêt général (PIG) et aux opérations d'intérêt national (DIN) et éléments relatifs aux projets d'infrastructure relevant de la compétence de l'État.
7. SDRIF : contribution à un projet pour l'Ile-de-France. Les 200 propositions d'Ile-de-France Environnement.
8. Transports.
How to build the thematic atlas?

- Take advantage of SPARQL and inference

**THEMATICAL ATLAS (SOURCE: URBAMET)**

<table>
<thead>
<tr>
<th>LOCATION</th>
<th>THEME</th>
<th>ARTICLES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>activity</td>
<td>1. Aéroports franciliens : projets d'adaptation de capacités et de meilleur accès.-</td>
</tr>
<tr>
<td></td>
<td>agent</td>
<td>2. Aviation (A) affaires en Île-de-France : enjeux et perspectives pour la région-capitale en 2005.-</td>
</tr>
<tr>
<td></td>
<td>event</td>
<td>3. Gare de l'Est - RER : 20 minutes en 2012.-</td>
</tr>
<tr>
<td></td>
<td>indicator</td>
<td>4. Incidence (I) 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.-</td>
</tr>
<tr>
<td></td>
<td>norm</td>
<td>5. Révision du Schéma Directeur de la Région d'Île-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.-</td>
</tr>
<tr>
<td></td>
<td>organization</td>
<td>6. Plan du Schéma directeur de la région Île-de-France. Contribution à un projet pour Île-de-France. Les 200 propositions d'Île-de-France Environnement.-</td>
</tr>
<tr>
<td></td>
<td>physical-object</td>
<td>7. Contribution à un projet pour Île-de-France. Les 200 propositions d'Île-de-France Environnement.-</td>
</tr>
<tr>
<td></td>
<td>airport</td>
<td>8. Transports.-</td>
</tr>
<tr>
<td></td>
<td>car</td>
<td>Carriage</td>
</tr>
<tr>
<td></td>
<td>carriage</td>
<td>Cinema</td>
</tr>
<tr>
<td></td>
<td>chain store</td>
<td>Chamber of commerce</td>
</tr>
<tr>
<td></td>
<td>chamber of commerce</td>
<td>Cinema</td>
</tr>
<tr>
<td></td>
<td>cinema</td>
<td>Cinema</td>
</tr>
</tbody>
</table>

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)