KYOTO: Open platform for mining facts

Piek Vossen, VU University Amsterdam

German Rigau, Eneko Agirre, Aitor Soroa, University of the Basque Country

Monica Monachini, Roberto Bartolini, Istituto di Linguistica Computazionale, CNR

Ontolex workshop, COLING 2010, Beijing

Overview

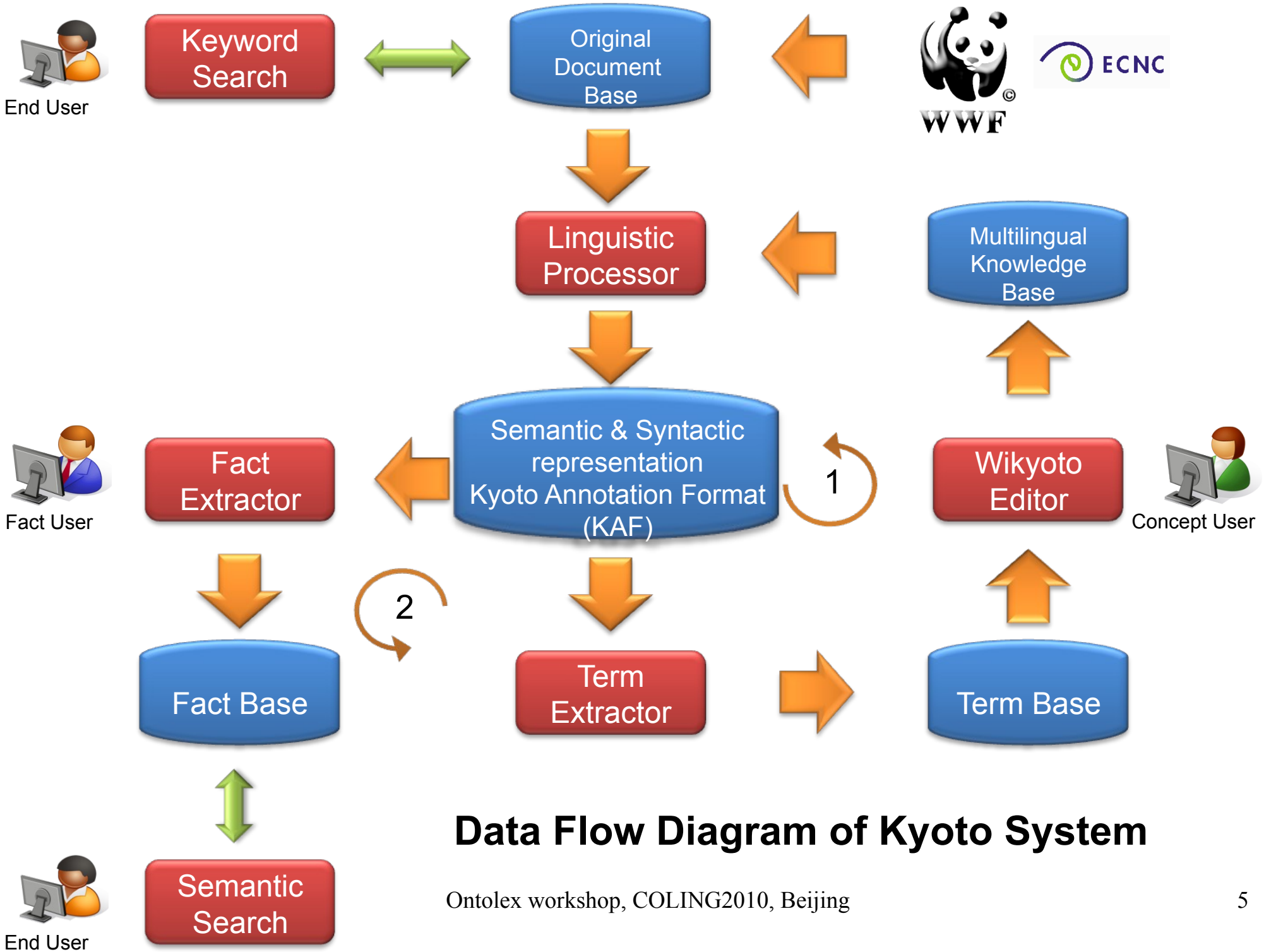
- Background on KYOTO project
- Knowledge architecture
- Ontotagging
- Text mining
- First results
- Conclusions

KYOTO (ICT-211423) Overview

- **Title: K**nowledge **Y**ielding **O**ntologies for **T**ransition-Based **O**rganization
- **Funded:**
  - 7th Framework Program-ICT of the European Union: Intelligent Content and Semantics
  - Taiwan and Japan funded by national grants
- **Goal**:
  - Open and free platform for knowledge sharing across languages and cultures
  - Wiki environment that allows people in the field to maintain their knowledge and agree on meaning without knowledge engineering skills
  - Bootstrap through open text mining & concept learning
  - Enables deep semantic search for facts and knowledge transition across different target groups, transgressing linguistic, cultural and geographic boundaries.
- **Application domain:** environment
- **Languages**: English, Dutch, Italian, Spanish, Basque, Japanes and Chinese
- **URL**: **http://www.kyoto-project.eu/**
- **Duration**: March 2008 – March 2011
- **Effort**: 364 person months of work.

What makes KYOTO different?

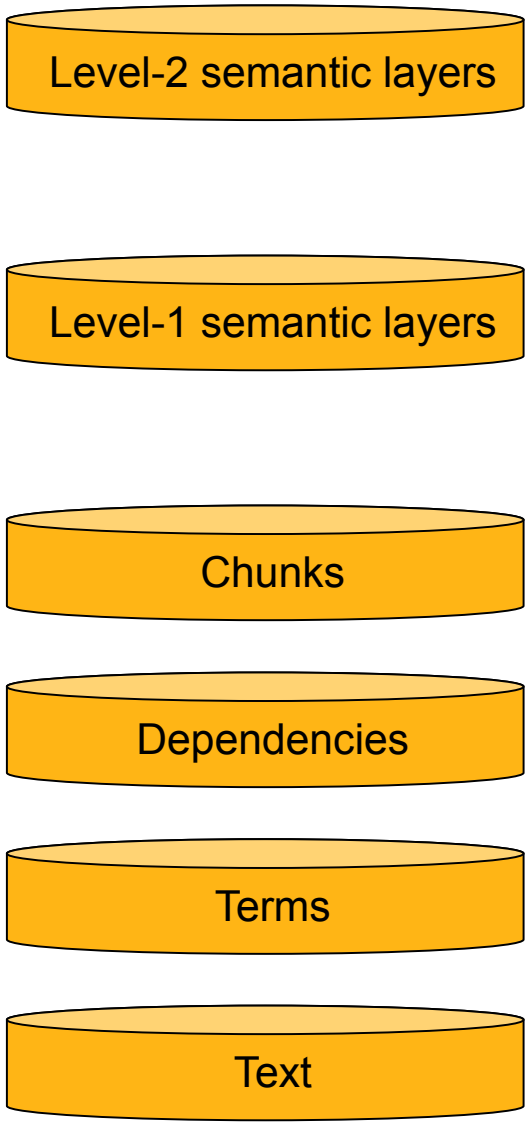
- Generic and open architecture that is tunable to any domain;
- Can be applied to any language and across languages;
- Represents a complete system for knowledge discovery:
  - Combines formal ontology with wordnets and domain background vocabularies
  - Text mining as ultimate application for exploiting rich knowledge system
  - Anchoring of knowledge done by domain experts and not by knowledge engineers
  - Mined facts represent return of investment for knowledge modeling



Data Flow Diagram of Kyoto System

Kyoto Annotation Format KAF

- **Text:** tokenization, sentences, paragraphs, with reference to the source
- **Terms** [Text]: words and multi-words, includes parts-of-speech, declension information, etc.
- **Dependencies** [Terms]: dependency relations between terms
- **Chunks** [Terms]: constituents & phrases



Structural KAF

```
<kaf>
  <text>
   <wf wid="w1" page="1" sent="1" para="1" fileoffset="0,3">most</wf>
   <wf wid="w2" page="1" sent="1" para="1" fileoffset="5,13">migratory</wf>
   <wf wid="w3" page="1" sent="1" para="1" fileoffset="15,19">birds</wf>
  </text>
  <terms>
    <term tid="t1" type="open" lemma="most" pos="Q">
      <span id="w1"/><!-- refers to "most" (w1) -->
    </term>
    <term tid="t2" type="open" lemma="migratory bird" pos="N">
      <span id="w2"/><span id="w3"/>
    </term>
  </terms>
</kaf>
```

KAF annotation: Semantic layers

```
<term tid="t4" type="open" lemma="population" pos="N">
    <span> <target id="w4"/>
    </span></term>
```

The word population is present in 13 synsets

Lemmi	Category	Glossa
population n1	noun.group	the people who inhabit a territory or state
population n2	noun.group	a group of organisms of the same species inhabiting a given area
population n3, universe n2	noun.cognition	(statistics) the entire aggregation of items from which samples can be drawn
population n4	noun.quantity	the number of inhabitants (either the total number or the number of a particular race or class) in a given place (country or city etc.)
population n5	noun.act	the act of populating (causing to live in a place)
population n1	noun.group	group of species that live in a habitat
population commission n1	noun.group	the commission of the Economic and Social Council of the United Nations that is concerned with population control

Word-Sense-Disambiguation

```
<term tid="t4" type="open" lemma="population" pos="N">
    <span>        <target id="w4"/>        </span>
  <externalReferences>
        < externalRef resource="WN-1.7" reference="ENG-3.0-00859568-n" confidence="0.80 "/>
        < externalRef resource="WN-1.7" reference="ENG-3.0-00257849-n" confidence="0.13 />
        < externalRef resource="WN-1.7" reference="ENG-3.0-00962397-n" confidence="0.07 />
        <externalRef resource="DolceLite-Kyoto" reference="physical plurality" confidence="0.80"/>
</externalReferences>
</term>
```

KAF Named Entities: locations

```
<location lid="l10">
  <kafReferences><kafReference pageId="7" id="t1753"/></kafReferences>
  <externalReferences>
    <externalRef confidence="0.9" reference="2648147" resource="GeoNames"/>
    <externalRef reference="eng-30-09316454-n" resource="wn30g">
    <externalRef confidence="1.0" reference="eng-30-00002684-n"  reftype="baseConcept"/>
    <externalRef confidence="1.0" reference="Kyoto#island-eng-3.0-09316454-n"
                    reftype="sc_equivalentOf" resource="ontology"/>
  </externalReferences>
  <geoInfo>
    <place countryCode="GB" countryName="United Kingdom" fname="island"
            latitude="54" longitude="-2" name="Great Britain" timezone="Europe/London"/>
  </geoInfo>
</location>
```

Knowledge architecture

Available repositories in KYOTO 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: 3 million species
- **Wordnets** for 7 languages: about 50,000 to 120,000 synsets per language
- **Ontologies**: SUMO, DOLCE, SIMPLE, etc.

What knowledge should the ontology represent?

- Vocabularies are too large for full inferencing with current reasoners
- Vocabularies are linguistically very diverse: include both rigid concepts (e.g. all *species*) and non-rigid concepts (e.g. *endangered species*), cf Guarino and Welty (2002)
- Inferencing capabilities of formal ontologies is not needed for all levels of knowledge
- Relevance for fact mining:
  - Variation in naming and reference → non-rigid
  - Encyclopedias define rigid concepts, text about the world mostly consist of non-rigid statements: *what is the case at a certain moment of time at a certain place for certain individuals*

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 reason with semantics and do useful stuff with textual data when it is capable of detecting the processes and states in which instances of such *frogs* are involved.
- Therefore:
  - the ontology should focus on modelling sufficient processes and states for entities.
  - Vocabularies should differentiate rigid from non-rigid terms and map the latter to the relevant processes and states
  - Vocabularies map all relevant words to processes and states

Wordnet-ontology-relations

- Rigid synset relations to ontology (e.g. *water*):
  - sc_equivalenceOf (= in WN-SUMO) or sc_subclassOf (+ in WN-SUMO)
- Non-rigid synset relations to ontology (e.g. *polluted water*):
  - sc_domainOf: range of ontology types that restricts a role
  - sc_playRole: role that is being played
  - sc_participantOf: the process in wich the role is played
- Rigidity can be detected automatically (Rudify, 80% precision, IAG 80%) and is stored in wordnets as attributes to synsets
- Domain processes (e.g. *water pollution)*:
  - sc_hasRole: role that is being played
  - sc_hasParticipant: the participant that plays the role played

Global Wordnet Grid Model

English Wordnet in WN-LMF

bird_1_N — *rigid* — sc_equivalentOf bird

migratory_bird_1_N — *non-rigid* — sc_domainOf bird; sc_playRole done-by; sc_participantOf migration

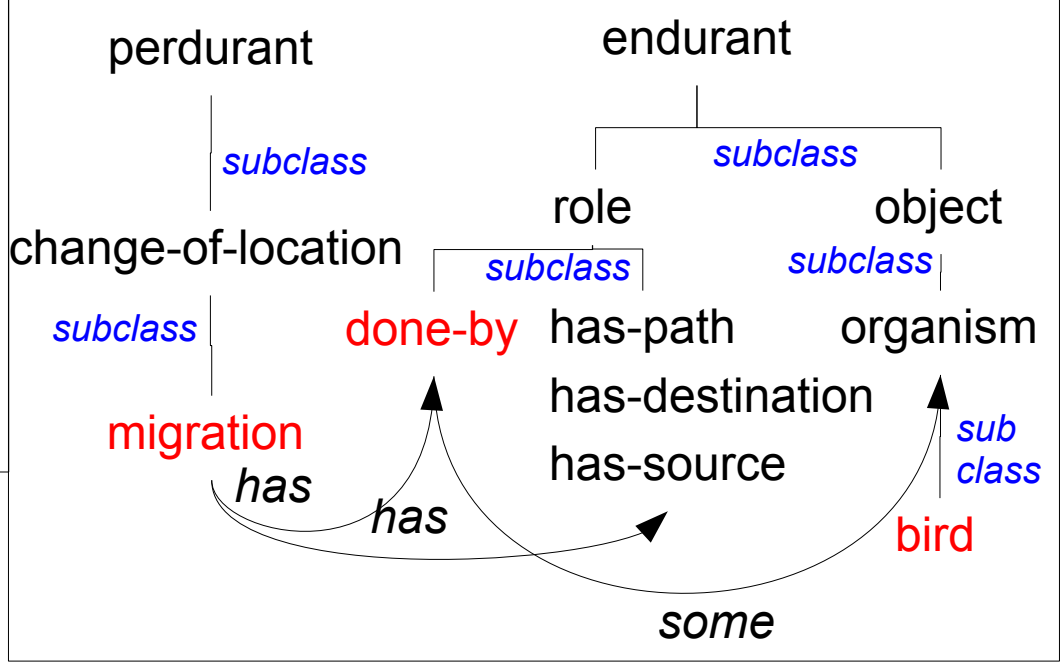
hyponym

duck_1_N, *rigid*

migration_4_N — sc_equivalentOf migration

migrate_1 _V — sc_equivalentOf migration

KYOTO Ontology in OWL-DL
(Extension of DOLCE LT)



Wordnet to ontology mappings

{create, produce, make}Verb, English

-> sc_ equivalenceOf ***construction***

{artifact, artefact}Noun, English

-> sc_domainOf ***physical_object***

-> sc_playRole ***result-existence***

-> sc_participantOf ***construction***

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

-> sc_domainOf ***amount_of_matter***

-> sc_playRole ***result-existence***

-> sc_participantOf ***construction***

{meat}Noun, English

-> sc_domainOf ***cow, sheep, pig***

-> sc_playRole ***patient***

-> sc_participantOf ***eat***

{ 名 肉 , 食物 , 餐 }Noun, Chinese

-> sc_domainOf ***animal***

-> sc_playRole ***patient***

-> sc_participantOf ***eat***

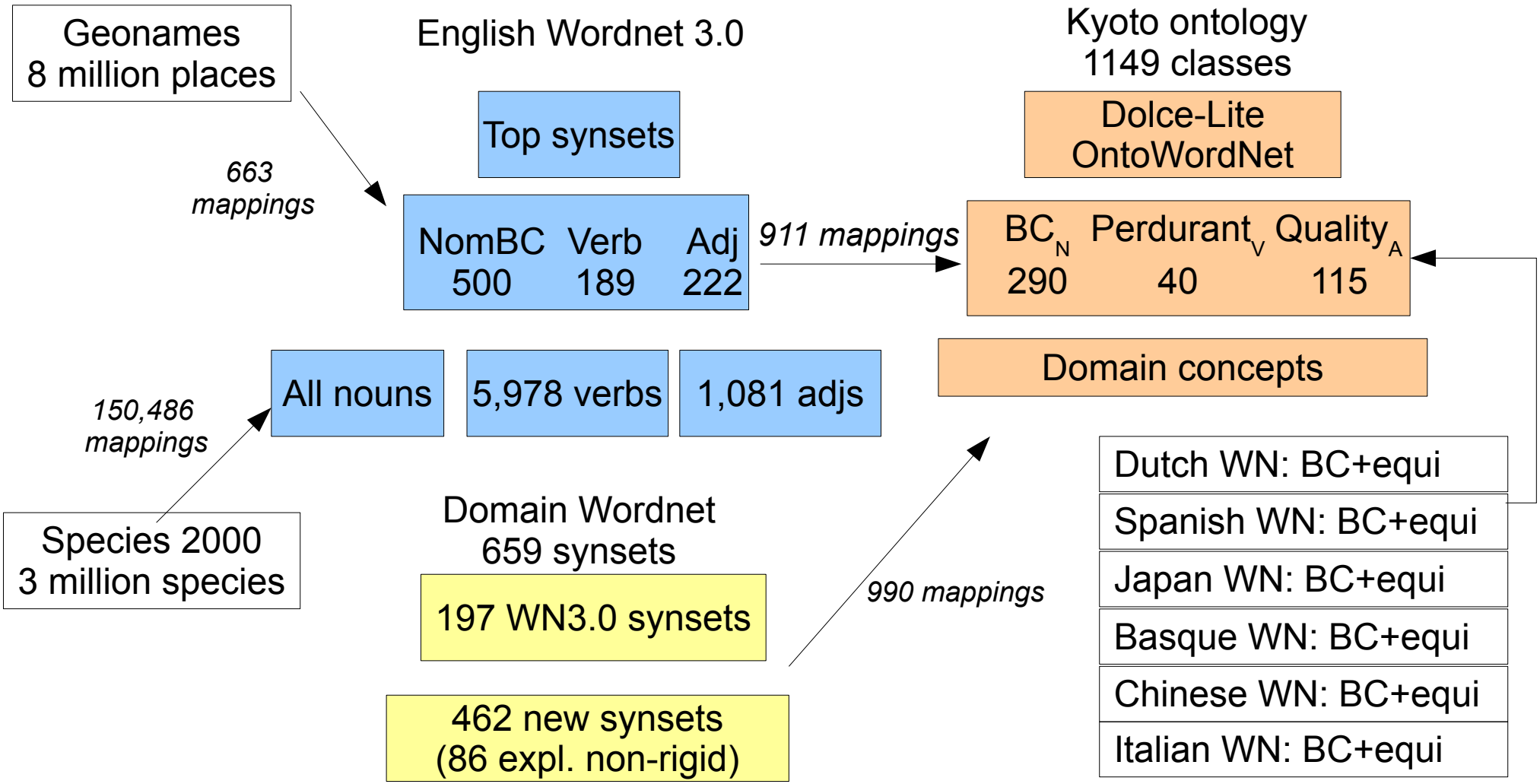
{ طعام ,لحم ,غذاء}Noun, Arabic

-> sc_domainOf ***cow, sheep***

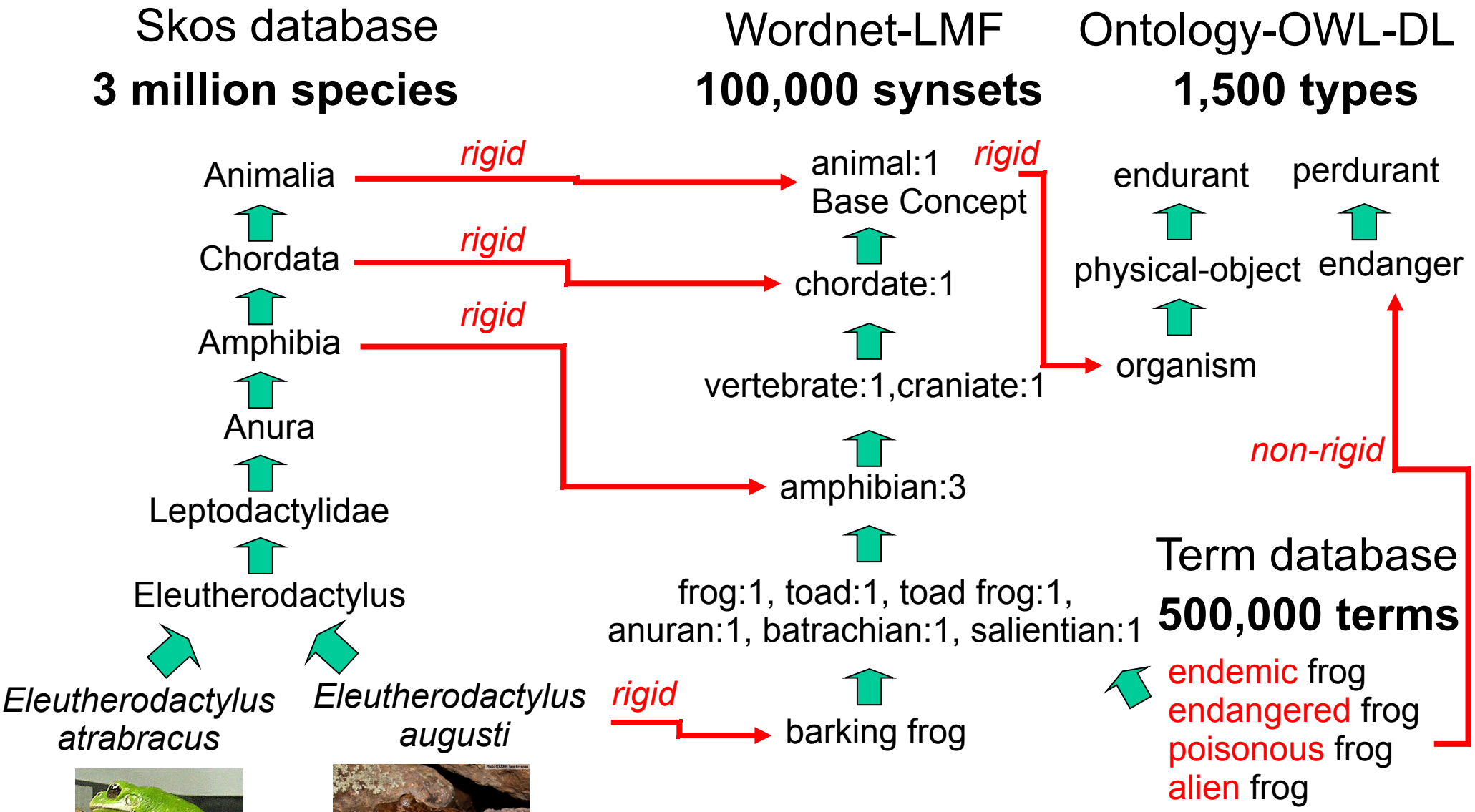
-> sc_playRole ***patient***

-> sc_participantOf ***eat***

3-layered knowledge model
Division of labor (Putnam 1975)



Division of labor in knowledge sources



Ontological tagging

Why OntoTagger?

- Abstract from words and synsets when we match patterns in KAF
- Find relations that are implied in different ways of expressions
  - *fish migration*
  - *migratory fish*
  - *migration of fish*
  - *fishes that migrate*

    … share the same ontological implications
- Do inferencing for ontological implications offline and not during the matching of patterns

Ontotagger

- Offline reasoning using so-called explicit ontology
- Explicit ontology: collects all implied ontological statements for all ontology labels:
  - Subclass relations: bird=>animal=>...=>endurant
  - Axioms:
    - migration=>motion(done-by, has-source, has-target, has-path).....=>perdurant (has-temporal-quality)
  - 20363 explicit ontology statements
- Match all wordnet synsets to ontological classes (directly or indirectly)
- Insert ontological statements into KAF for all wordnet synset annotations of terms

Ontotagged KAF

```
<term lemma="pollution" pos="N" tid="t13444" type="open">
  <externalReferences>
   <externalRef ref="eng-30-276987-n" confidence="0.8" res="wn30g"/> <!-- WSD output-->
   <externalRef reftype="baseConcept" ref="eng-30-191142-n" res="wn30g"/>
   <externalRef reftype="sc_subClassOf" ref="Kyoto#change_of_state-eng-3.0-19130-n">
      <externalRef reftype="SubClassOf" ref="Kyoto#change-eng-3.0-191142-n" status="implied"/>
      <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#accomplishment" status="implied"/>
      <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#event" status="implied"/>
      <externalRef reftype="SubClassOf" ref="DOLCE-Lite.owl#perdurant" status="implied"/>
    </externalRef>
  </externalReferences>
</term>
```

Text mining

Event and Fact extraction

- **KYBOT** = Knowledge Yielding Robot
- **Engine** that applies any set of **profiles** to KAF annotated documents and generates output layers when constraints are matched
- **Kybot profiles** are simple Xml structures that specify:
  - A set of variables, with for each variable a set of constraints
  - Relations between variables
  - Some kind of output format
- Profiles can combine logical schemas with linguistic patterns
- Kybots can run sequentially taking the output of one kybot as the input for the next

Kybot profile

```
<?xml version="1.0" encoding="utf-8"?>
<Kybot id="impact-of">
 <variables>
  <var name="X" type="term" lemma="impact ! effect"/>
  <var name="Y" type="term" lemma="of"/>
  <var name="Z" type="term" pos="n"/>
 </variables>
 <relations><!-- X Y Z -->
  <root span="X"/>
  <rel span="Y" pivot="X" direction="following" immediate="true" />
  <rel span="Z" pivot="Y" direction="following" immediate="true" />
 </relations>
 <events>
  <event id='impact-of'>
   <target id='$Z/@tid' lemma='$Z/@lemma'/>
  </event>
 </events>
</Kybot>
```

Kybot profile

```
<kprofile>
 <variables>
  <var name="x" type="term" pos="N" ref="DOLCE-Lite.owl#physical-object"
        reftype="SubClassOf"/>
  <var name="y" type="term" ref="Kyoto#creation__creative_activity-eng-3.0-908492-n"
        reftype="SubClassOf"/>
  <var name="z" type="term" ref="DOLCE-Lite.owl#contamination_pollution"
        reftype="SubClassOf"/>
 </variables>
 <relations>
  <root span="y"/>
  <rel span="x" pivot="y" direction="preceding" immediate="true"/>
  <rel span="z" pivot="y" direction="following"/>
 </relations>
 <events>
  <event target="$y/@tid" lemma="$y/@lemma" pos="$y/@pos"/>
  <role target="$x/@tid" rtype="done-by" lemma="$x/@lemma"/>
  <role target="$z/@tid" rtype="result"lemma="$z/@lemma"/>$
 </events>
</kprofile>
```

Kybot output


```
<kybotOut>
 <doc name="11767.mw.wsd.ne.onto.kaf">
  <event eid="e1" lemma="generate" pos="V" target="t3504"
          synset="eng-30-01621555-v" score="0.16">
   <place countryCode="US" countryName="United States" fname="first-order administrative
          division" latitude="40.2724534393745" longitude="-76.9056701660156"
          name="Pennsylvania" population="12440621" timezone="America/New_York"/>
   <dateInfo dateISO="1950" lemma="1950"/>
  </event>
  <role rid="r1" lemma="sceptic system" rtype="done-by" target="t3493" pos="N" event="e1"
          synset="dw-eng-30-113-n" score="1.0"/>
  <role rid="r2" lemma="pollution" rtype="result" target="t3495" pos="N" event="e1"
          synset="eng-30-14516743-n" score="0.85"/>
 </doc>
</kybotOut>
```

First results

- Document base databases on Estuaries from English PDFs and web pages (almost 5,000 sources)
- Processed by into KAF applying multiword detection, WSD and NER
- Term databases derived by Tybots with almost 4,000 candidate terms from 3 benchmark documents
- Wikyoto:
  - Domain ontology (940 domain classes added to middle & top)
  - Domain wordnet (650 domain terms)
- Reprocessed through Multiword tagger, WSD using domain extension
- Ontotagger uses explicit ontology and wordnet-ontology mappings to enrich KAF for 3 documents with **more than half a million statements**
- Kybots using 95 profiles generated 3,693 events and 2,481 roles for 3 benchmark documents

Comparing estuary databases

	bench mark documents (3)				estuary documents (4742)			
	No Domain		Domain		No Domain		Domain	
ontology references	555,677		576,432				48,708,300	
implied ontology ref	457,332	82.30%	474,916	82.39%			40,523,452	83.20%
direct ontology ref.	53,178	9.57%	54,769	9.50%			4,377,814	8.99%
domain synset to ontology map.	45,167	8.13%	46,747	8.11%			3,807,034	7.82%
Total	555,677		576,432				48,708,300	

- Slight increase in the number of ontological implications

Relations btw WN and central Onto

relation	11614	23452	11767
stateOf	3	13	43
hasPart	16	24	272
playCoRole		4	26
hasCoParticipant	1	4	27
partOf	5	11	26
equivalenceOf	358	954	2325
domainOf	2	8	118
playRole	3	21	68
hasRole	11	17	66
hasState	23	21	124
participantOf	3	23	79
subClassOf	4317	12407	25324
hasParticipant	12	19	71

```
<event lemma="pollution" pos="N" freq="89" synset="">
<role lemma="excess nitrogen" pos="N" rtype="source" ivalue="0.5" freq="32" synset=""/>
<role lemma="farmland" pos="N" rtype="source" ivalue="0.5" freq="64" synset=""/>
<role lemma="chemical fertilizer" pos="N" rtype="source" ivalue="0.5" freq="16" synset="eng-30-14520829-n""/>
<role lemma="septic system" pos="N" rtype="source" ivalue="0.33" freq="4" synset="eng-30-13474290-n""/>
<role lemma="oxygen" pos="N" rtype="source" ivalue="0.33" freq="16" synset="eng-30-14648100-n""/>
<role lemma="chemical" pos="N" rtype="source" ivalue="0.33" freq="16" synset="eng-30-00276987-n""/>
<role lemma="sediment" pos="N" rtype="source" ivalue="0.16" freq="16" synset="eng-30-09428967-n""/>
<role lemma="subdivision" pos="N" rtype="source" ivalue="0.25" freq="16" synset="eng-30-00276987-n""/>
<role lemma="stream" pos="N" rtype="source" ivalue="0.083" freq="128" synset="eng-30-09448361-n""/>
<role lemma="creek" pos="N" rtype="source" ivalue="0.071" freq="128" synset="eng-30-09229409-n""/>
<role lemma="power plant" pos="N" rtype="source" ivalue="0.125" freq="32" synset="eng-30-14516743-n""/>
<role lemma="vehicle" pos="N" rtype="source" ivalue="0.16" freq="8" synset="eng-30-00248977-n""/>
<role lemma="water" pos="N" rtype="source" ivalue="0.025" freq="8" synset="eng-30-14845743-n""/>
<role lemma="primary source" pos="N" rtype="source" ivalue="0.5" freq="32" synset="eng-30-00786195-n""/>
<role lemma="outside" pos="N" rtype="source" ivalue="0.5" freq="9" synset="eng-30-08613472-n""/>
<role lemma="farming operation" pos="N" rtype="source" ivalue="0.5" freq="8" synset=""/>
<role lemma="location" pos="N" rtype="source" ivalue="0.2" freq="4" synset="eng-30-14516743-n""/>
<role lemma="habitat" pos="N" rtype="source" ivalue="0.14" freq="4" synset="eng-30-14516743-n""/>
<role lemma="tributary" pos="N" rtype="source" ivalue="0.16" freq="4" synset="eng-30-14516743-n""/>
<role lemma="land" pos="N" rtype="source" ivalue="0.018" freq="5" synset="eng-30-00276987-n""/>

<role lemma="overexploitation" pos="N" rtype="participant" ivalue="1.0" freq="4" synset="dw-eng-30-258-n""/>
<role lemma="load" pos="N" rtype="participant" ivalue="0.11" freq="9" synset="eng-30-00276987-n""/>
<role lemma="diet" pos="N" rtype="participant" ivalue="1.0" freq="4" synset="eng-30-13085864-n""/>
<role lemma="control effort" pos="N" rtype="participant" ivalue="1.0" freq="4" synset="dw-eng-30-258-n""/>
<role lemma="cap" pos="N" rtype="participant" ivalue="0.5" freq="4" synset="eng-30-00201058-n""/>
<role lemma="area" pos="N" rtype="participant" ivalue="0.038" freq="16" synset="eng-30-00268557-n""/>
<role lemma="reduction effort" pos="N" rtype="participant" ivalue="0.5" freq="4" synset=""/>
<role lemma="water" pos="N" rtype="participant" ivalue="0.025" freq="4" synset="eng-30-05798863-n""/>

<role lemma="atmosphere" pos="N" rtype="in" ivalue="1.0" freq="4" synset=""/>
<role lemma="jurisdiction" pos="N" rtype="in" ivalue="1.0" freq="2" synset="eng-30-00453731-n""/>
<role lemma="waterway" pos="N" rtype="in" ivalue="0.2" freq="8" synset="eng-30-09476331-n""/>
<role lemma="influence" pos="N" rtype="in" ivalue="1.0" freq="4" synset="eng-30-00551215-n""/>
```

Conclusions

- We described an open and generic platform for knowledge sharing;
- Applied to 7 languages, semantically processed in the same way;
- Tailored to a domain to mine any type of conceptual relations;
- Handles semantic implications expressed in many different way;
- Future work: benchmarking the extracted knowledge and the semantic modules