

Overview

- General information Kyoto
- System architecture
- Concept mining
- Knowledge editing environment
- Event mining
- Evaluation
- Semantic search

KYOTO (ICT-211423) Overview

- **Title:** Knowledge Yielding Ontologies for Transition-Based Organization
- **Funded:**
 - 7th Framework Program-ICT of the European Union: Intelligent Content and Semantics
 - Taiwan and Japan funded by national grants
- **Goal:**
 - Open & free platform for knowledge sharing across languages and cultures
 - Wiki environment for people in the field to maintain their knowledge and agree on meaning without knowledge engineering skills
 - Bootstrap through concept learning & open text mining
 - Enables knowledge transition and information search across different target groups, transgressing linguistic, cultural and geographic boundaries → deep semantic search for facts and knowledge
- **Languages:** English, Dutch, Italian, Spanish, Basque, Chinese, Japanese
- **Application domain:** environment
- **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

Information need in the environment

High-level targets & low-level questions

- High level target (about 300 questions collected) that simulate problem solving tasks:
 - Are there huge negative effects with regard to ecological networks and alien invasive species?
- Low level facts that support answering the high level targets:
 - cases of alien invasion
 - amount of species
 - causal relations associated with these (increments of) invasions
 - causes related to ecological networks
 - limit in the same time and location boundary

Baseline retrieval results

6 persons, 30 high-level questions,

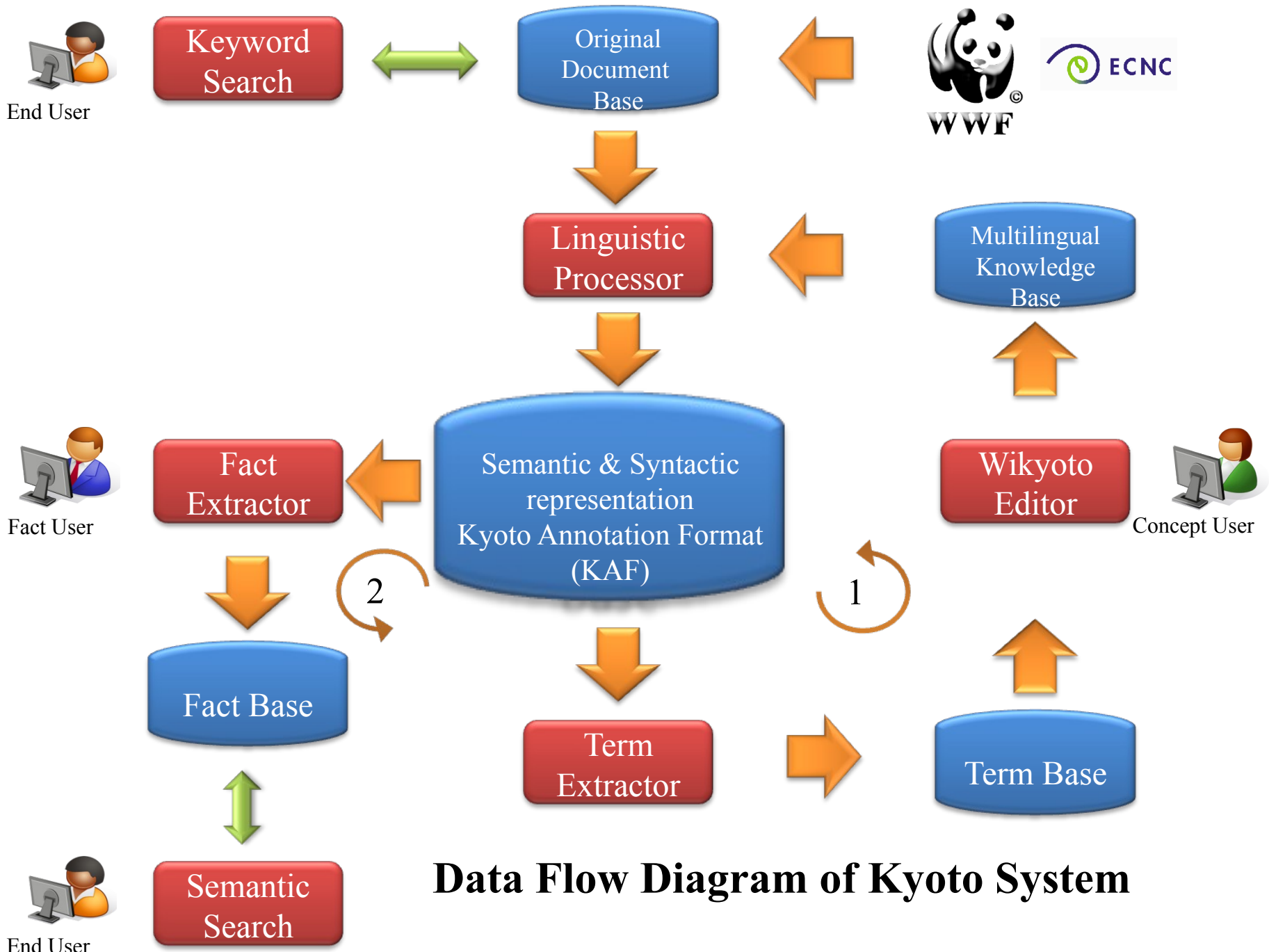
Result Rank	CONFIRMED		DISAPPROVED		UNDECIDED		Total	
0	13	20.31%	27	20.30%	10	15.87%	50	19.23%
1	6	9.38%	9	6.77%	9	14.29%	24	9.23%
2	8	12.50%	13	9.77%	7	11.11%	28	10.77%
3	5	7.81%	6	4.51%	3	4.76%	14	5.38%
4	8	12.50%	6	4.51%	2	3.17%	16	6.15%
5	2	3.13%	7	5.26%	3	4.76%	12	4.62%
6	2	3.13%	6	4.51%	4	6.35%	12	4.62%
7	2	3.13%	2	1.50%	1	1.59%	5	1.92%
8	4	6.25%	3	2.26%	1	1.59%	8	3.08%
9	1	1.56%	5	3.76%	0	0.00%	6	2.31%
-1	13	20.31%	49	36.84%	23	36.51%	85	32.69%
Total	64	24.62%	133	51.15%	63	24.23%	260	

System architecture



XRCE Seminar, September 16th, 2010, Grenoble





Data Flow Diagram of Kyoto System

KyotoCore



Capture

↓ *Html*

Document base
Job dispatcher

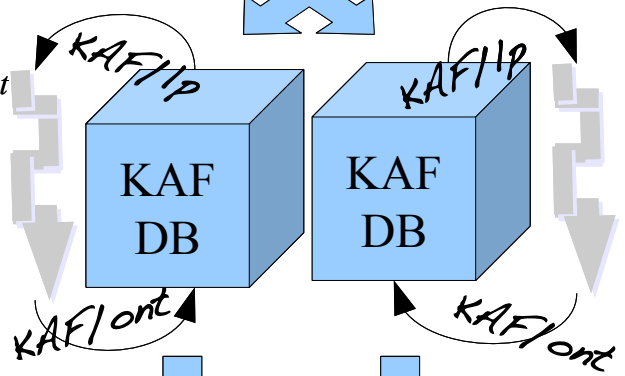
PipeT

Modules

- LP-client
- MW-tagger
- Sense-tagger^{UKB}
- NE-tagger
- ON-tagger
- Tybot
- Kybot

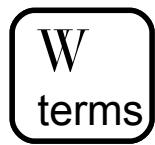
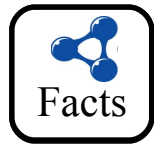
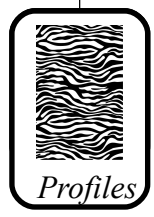
LP-server

- LP-client
- MW-tagger
- Sense-tagger
- NE-tagger
- ON-tagger
- Kybot



LP-server

- LP-client
- MW-tagger
- Sense-tagger
- NE-tagger
- ON-tagger
- Tybot



Knowledge Repository



Kyoto Annotation Format

KAF

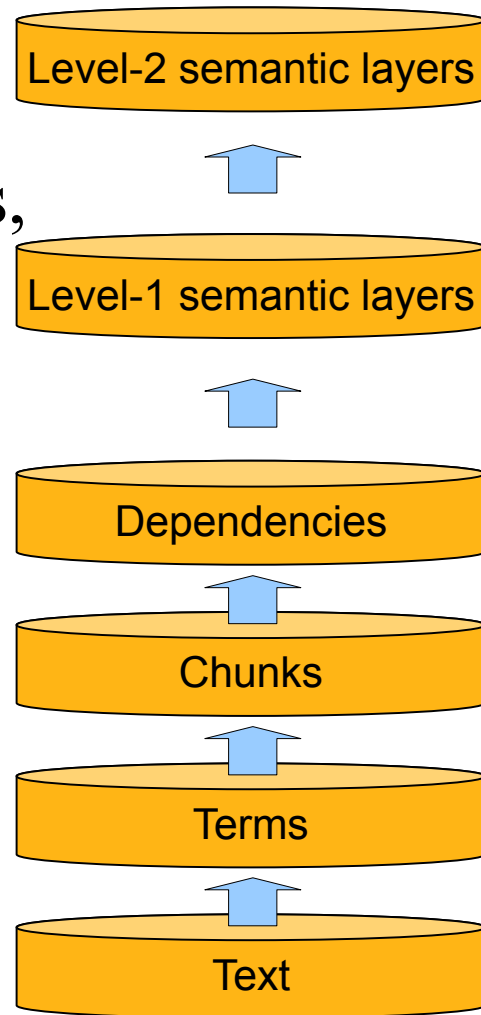
- Based on Layered Annotation Format or LAF (Ide and Romary 2002)
- Stand off annotation
- Uniform representation for 7 languages
- Sharing of semantic modules across different languages: multiword tagging, WSD, Named Entity recognition, Onto tagging and event/fact extraction
- Cross-lingual semantic search for 7 languages



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.
- **Chunks [Terms]:** constituents & phrases
- **Dependencies [Terms]:** dependency relations between terms



Structural KAF

```

<kaf>
  <text>
    <wf wid="w1" page="1" sent="1" para="1" f-offset="0,4">large</wf>
    <wf wid="w2" page="1" sent="1" para="1" f-offset="6,14">migratory</wf>
    <wf wid="w3" page="1" sent="1" para="1" f-offset="16,20">birds</wf>
  </text>
  <terms>
    <term tid="t1" type="open" lemma="large" pos="G">
      <span id="w1"/><!-- refers to "large" (w1) -->
    </term>
    <term tid="t2" type="open" lemma="migratory bird" pos="N">
      <span id="w2"/><span id="w3"/>
    </term>
  </terms>
</kaf>

```

Structural KAF

<kaf>

<text>...</text><!-- defines w1, w2, w3 -->

<terms>...</terms><!-- defines t1, t2 -->

<deps>

<!-- dependency: "large" (t1) → "migratory birds" (t2) -->

<dep from="t1" to="t2" rfunc="mod"/>

</deps>

<chunks>

<!-- two per cent -->

<chunk cid="c1" head="t2" phrase="NP">

<!-- refers to term: "large" -->

<!-- refers to term: "migratory bird" -->

</chunk>

</chunks>

</kaf>

KAF annotation: Semantic layers

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

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_n6	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

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

```
<externalReferences>
```

```
< externalRef resource="WN-1.7" reference="EN-17-00859568-n" confidence="0.80" />
```

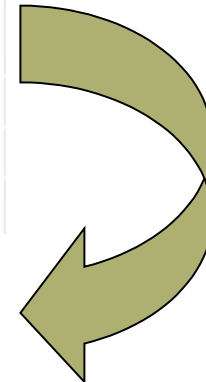
```
< externalRef resource="WN-1.7" reference="EN-17-00257849-n" confidence="0.13" />
```

```
< externalRef resource="WN-1.7" reference="EN-17-00962397-n" confidence="0.07" />
```

```
<externalRef resource="DOLCE" reference="Group" confidence="0.80"/>
```

```
</externalReferences>
```

```
</term>
```



KAF Named Entities

```
<date did="d4">
```

```
  <kafReferences>
```

```
    <kafReference pageId="1" id="t327"/>
```

```
    <kafReference pageId="6" id="t1578"/>
```

```
    <kafReference pageId="6" id="t1645"/>
```

```
  </kafReferences>
```

```
  <dateInfo dateISO="2010" lemma="2010"/>
```

```
</date>
```

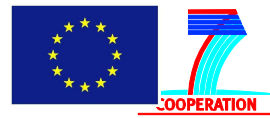
```
</dates>
```

KAF Named Entities

```

<location lid="110">
  <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 mining



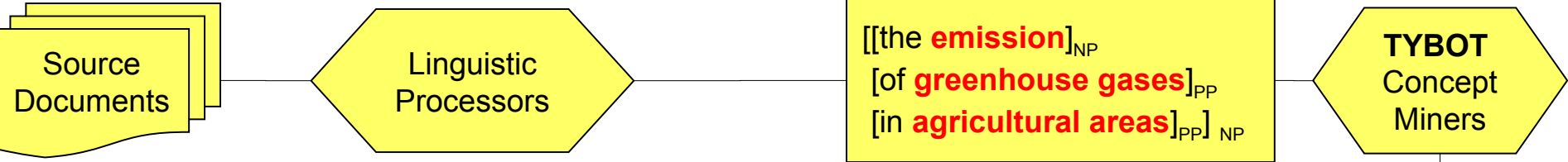
XRCE Seminar, September 16th, 2010, Grenoble



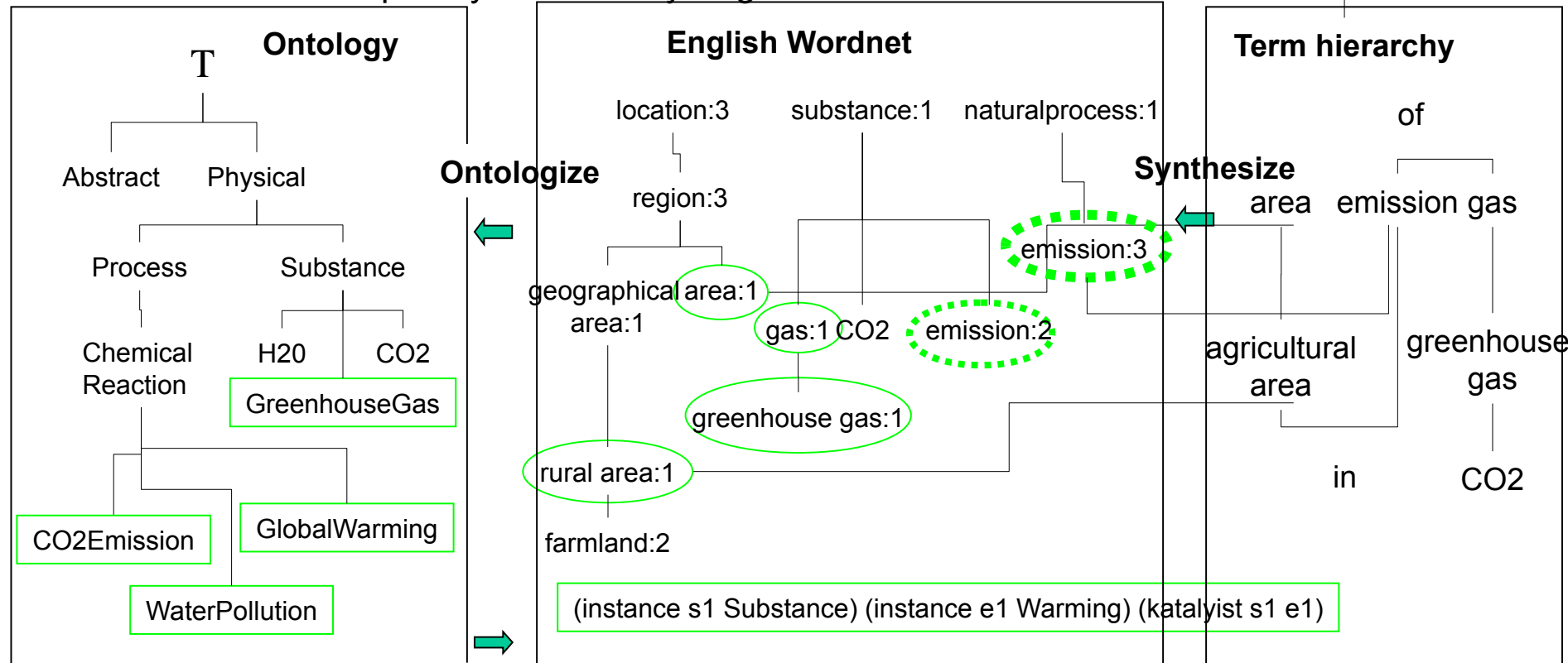
Concept mining

- Tybots = Term yielding robots
- Input are text documents
- Linguistic processors generate KAF annotation (sequential):
 - morpho-syntactic analysis: lemmas, constituents
 - wordnet mappings through WSD
 - semantic role detection
 - named entity detection
- Output are term hierarchies in TMF (generic):
 - relations based on term structure and Hearts patterns
 - statistical data
 - quantified structural and semantic relations

Conceptual modeling



Morpho-syntactic analysis generates KAF



Axiomatize



Term database demo



XRCE Seminar, September 16th, 2010, Grenoble



Wikyoto knowledge editor

http://xmlgroup.iit.cnr.it/demos/Wikyoto_Knowledge_Editor/

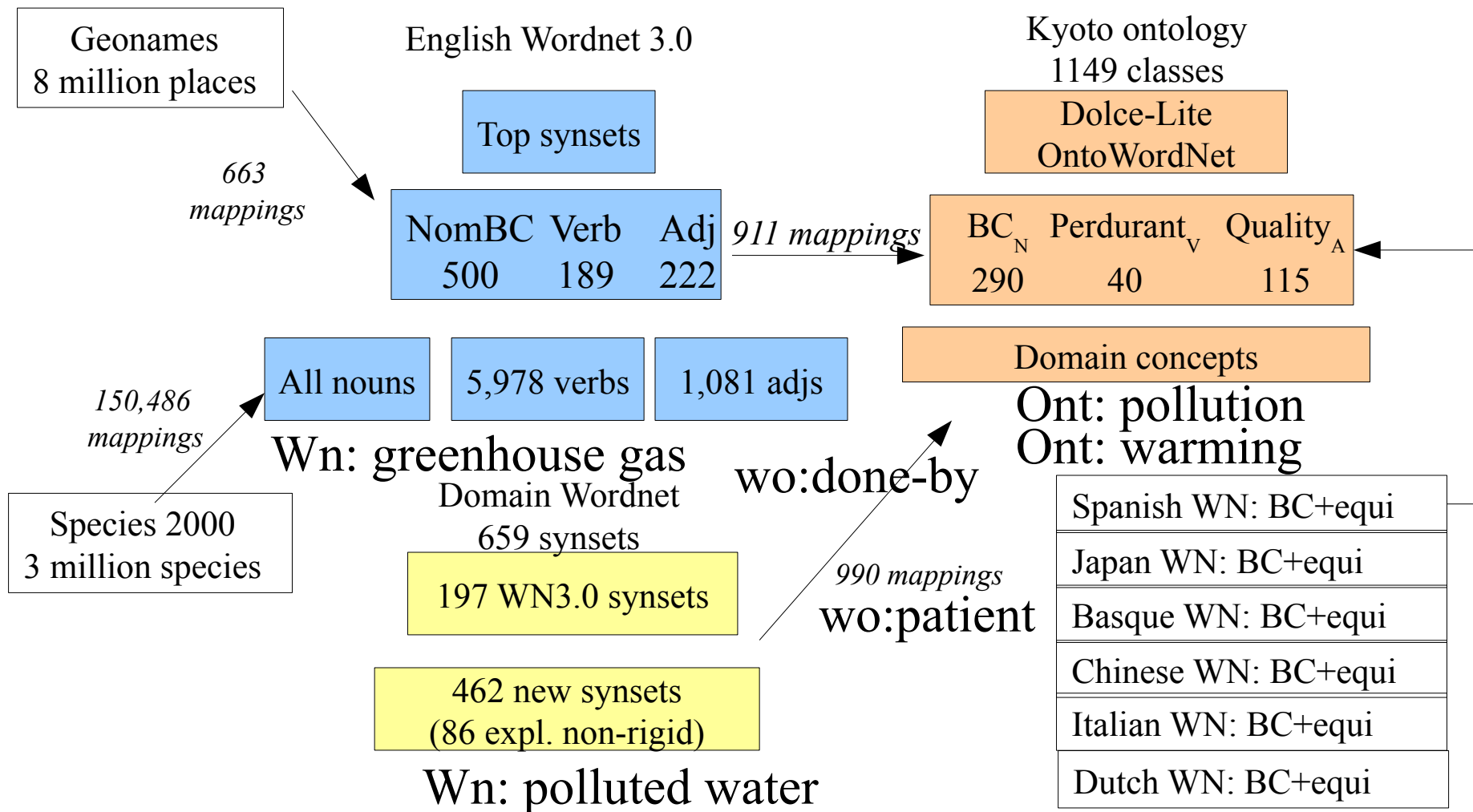


XRCE Seminar, September 16th, 2010, Grenoble



3-layered knowledge model

Division of labor (Putnam 1975)



Fact mining

- KYBOT = Knowledge Yielding Robot
- Engine that applies any set of kybot 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" pos="n" lemma="impact"/>
    <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>
  <facts>
    <fact id='impact-of'>
      <target id='$Z/@tid' lemma='$Z/@lemma'/>
    </fact></facts></Kybot>
```

Ontotagger

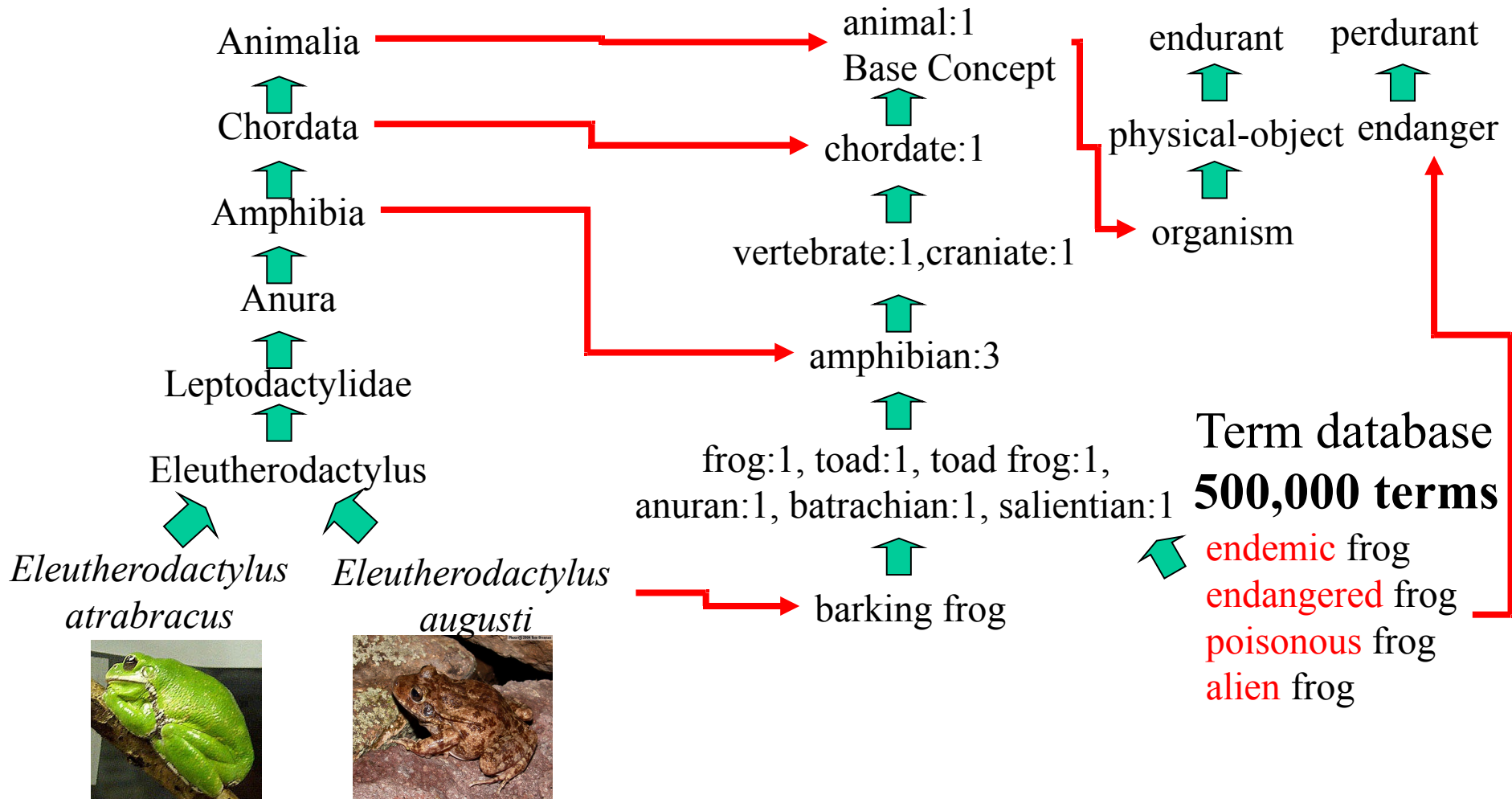
- Insert ontological statements into KAF with wordnet synset annotations
- Allows to specify high-level semantic constraints in Kybot profiles instead of word or synset constraints
- Offline reasoning using so-called explicit ontology
- Collect all implied ontological statements for all ontology labels:
 - Subclass relations: $\text{bird} \Rightarrow \text{animal} \Rightarrow \dots \Rightarrow \text{endurant}$
 - Axioms:
 - $\text{migration} \Rightarrow \text{motion}(\text{done-by}, \text{has-source}, \text{has-target}, \text{has-path}) \dots \Rightarrow \text{perdurant}(\text{has-temporal-quality})$
- Match all wordnet synsets to ontological classes

Division of labor in knowledge sources

Skos database
2.1 million species

Wordnet-LMF
100,000 synsets

Ontology-OWL-DL
2,000 types



Ontotagged KAF

```
<term lemma="water pollution" pos="N" tid="t13444" type="open">
  <externalReferences>
    <externalRef reference="eng-30-14516743-n" confidence="0.8" resource="wn30g"/> <!-- WSD output -->
    <externalRef reference="eng-30-00191142-n" reftype="baseConcept" resource="wn30g"/>
    <externalRef reftype="sc_hasParticipant" reference="Kyoto#water">
    <externalRef reftype="sc_hasRole" reference="DOLCE-Lite.owl#patient">
    <externalRef reftype="sc_subClassOf" reference="DOLCE-Lite.owl#contamination_pollution">
      <externalRef reftype="SubClassOf" reference="Kyoto#change-eng-3.0-00191142-n" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#accomplishment" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#event" status="implied"/>
      <externalRef reftype="SubClassOf" reference="DOLCE-Lite.owl#perdurant" status="implied"/>
    <externalRef>
  </externalReferences>
</term>
```

Kybot profile

```
<kprofile>
  <variables>
    <var name="x" type="term" pos="N" ref="DOLCE-Lite.owl#physical-object"/>
    <var name="y" type="term" ref="Kyoto#creation" lemma="! make"/>
    <var name="z" type="term" ref="DOLCE-Lite.owl#accomplishment"
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 admin
  division" latitude="40.27" longitude="-76.90"
  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>
```



Fact database demo



XRCE Seminar, September 16th, 2010, Grenoble



Full knowledge cycle

- Document base databases on Estuaries from English PDFs and web pages (almost 5,000 sources)
- Processed by LP^{KAF}, Multiword tagger, WSD^{Eval} and NER^{Eval}
- 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 with **more than half a million statements**
- Kybots using 95 profiles generated 3,693 events and 2,481 roles for 3 benchmark documents
- Semantic search accesses the output of the Kybots

Comparing estuary databases

	bench mark documents (3)		estuary documents (4742)	
	No Domain	Domain	No Domain	Domain
terms	22,204	22,204	2,419,839	2,419,839
multiwords	145	600	4,389	6,671
synsets	12,526	12,910	1,021,598	1,023,017
ne location	158	126	41,681	40,714
ne date	67	66	10,288	10,233

- Detected more multiwords in domain versions: 400% & 52% respectively
- More or less same number synsets and named entities

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

Text mining evaluation

- Reproducible evaluation that can be compared to other systems
- Conversion of Kybot events to triplets
 - Relation
 - List of word tokens representing the event
 - List of word tokens representing the participant
- Measure precision & recall by matching triplets in terms:
 - Tokens representing events and participant
 - Relation

Evaluation

“Research continued on the disease (w12239) mycobacteriosis (w12240). Modeling results provided the first evidence of mycobacteriosis (w12249) mortality (w12250) in the striped (w12253) bass (w12254) population (w12255) in the Bay (w12258).”

(TIME, w12250, w12221)

<!-- mortality, 2008 →

(DONE-BY, w12250, w12239;w12240)

<!-- mortality, disease
mycobacteriosis →

(PATIENT, w12250, w12253;w12254;w12255)

<!-- mortality, striped bass
population →

(LOCATION, w12250, w12258,)

<!-- mortality, Bay →

Evaluation

- Gold standard using KAF annotation tool:
 - Manual annotation of word tokens in KAF
 - Events, participants, time and location
- Kybot profiles:
 - 68 profiles
 - Basic English structures, N-N, A-N, N-V-N
 - Semantic constraints using ontological labels

Evaluation

- Single document on Chesapeake Bay: 16,145 words
- Gold standard:
 - Nr. of triplets 348
 - Average nr. of event tokens 1
 - Average nr. of participant tokens 2
- System output
 - Nr. of triplets 622
 - Average nr. of event tokens 1
 - Average nr. of participant tokens 1

Evaluation

| | | Total correct | Precision | Recall |
|-----------------------|------------------|---------------|-----------|--------|
| identical identifiers | same relation | 50 | 0,08 | 0,14 |
| | ignored relation | 146 | 0,23 | 0,42 |
| partial identifiers | same relation | 92 | 0,15 | 0,26 |
| | ignored relation | 212 | 0,34 | 0,61 |

- Identical identifiers = exactly the same word tokens need to be specified
- Partial identifiers = at least one word token needs to overlap

| | Gold-standard | % of events | System | % of events |
|-----------------|---------------|-------------|--------|-------------|
| Events | 169 | | 771 | |
| Roles | 348 | 2,06 | 1035 | 1,34 |
| Participant | 0 | 0 | 323 | 0,42 |
| Done-by | 46 | 0,27 | 285 | 0,37 |
| Patient | 115 | 0,68 | 173 | 0,22 |
| Simple-cause-of | 49 | 0,29 | 16 | 0,02 |
| Destination-of | 24 | 0,14 | 0 | 0 |
| Has-state | 33 | 0,2 | 153 | 0,2 |
| Others | 27 | 0,16 | 85 | 0,11 |
| TIME | 20 | 0,12 | 0 | 0 |
| PLACE | 5 | 0,03 | 0 | 0 |

Some problems for Kybots

- Handle complex terms such as *crab fishery*, *migratory bird* that imply events and relations
- Prepositions can only be interpreted as more specific roles in combination with the verb or noun denoting the event: *run off*, *increase to*
- Ontology is not rich enough (yet) to make constraints more precise
- Word-sense-disambiguation is not good enough (50% precision) so that we cannot exclude irrelevant meanings: *crab* is also a process in wordnet

Future directions

- Extension of the domain wordnet and ontology
- Improve WSD using the domain wordnets
- Ranking of the results using the WSD scores of the event elements → down-ranks *crab* as an event
- Use of FrameNet-like resources for preposition-verb dependencies → SRL
- Use machine-learning for SRL and for creating more-specific Kybot profiles

Where do we stand now?

- Fully integrated system:
 - Build around a flexible, extendible representation format (KAF) tested for 7 languages
 - For which we build a new knowledge repository structure that combines background knowledge, wordnets and ontologies in a formal model
 - Through which we applied a full knowledge cycle for Estuary databases
- KYOTO is **NOT** another ad hoc Text Mining solution but a generic knowledge and information modeling platform that can be tuned conceptually and maps to many languages
- We will improve through further modeling and evaluation cycles, resulting in version releases in the 3rd year.



The core Kyoto system is distributed
under the free open source license
(GPL)

Thank you for your attention

