



# KYOTO: Open platform for mining facts

*Asian-European project funded by the EU, Taiwan and NICT (Japan)*

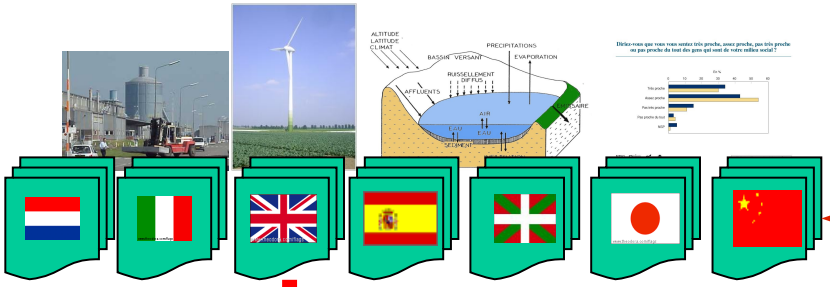
*Piek Vossen, VU University Amsterdam*

## 2nd KYOTO Workshop, 25-28th January 2011, Gifu


# Project goals and target groups

- 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 knowledge transition and information search across different target groups, transgressing linguistic, cultural and geographic boundaries.
  - Enables deep semantic search for facts and knowledge

**Distributed, diverse & dynamic data**



**Social communities:**
**Environmental organizations**

WWF ECNC

**Process text:**
"Sudden increase of CO2 emissions in 2008 in Europe"

KYOTO Knowledge Cycle

**Cross-lingual semantic search**

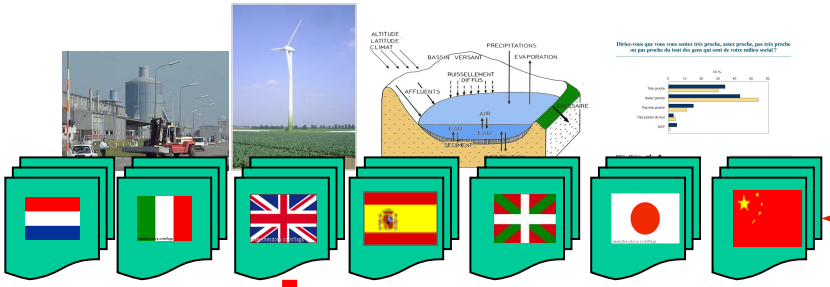
Show me a list of emissions?

| | | | |
|---|---|---|---|
| *emission* | *co2* | *2008* | *Europe* |
| *release* | *toxic gas* | *2005* | *Spain* |
| *emit* | *carbondioxide* | | *China* |
| ..... | | | |

**Index facts:**

| | |
|---|---|
| Process: | Emission |
| Involves: | CO2 |
| Property: | increase, sudden |
| When: | 2008 |
| Where: | Europe |

## Distributed, diverse & dynamic data



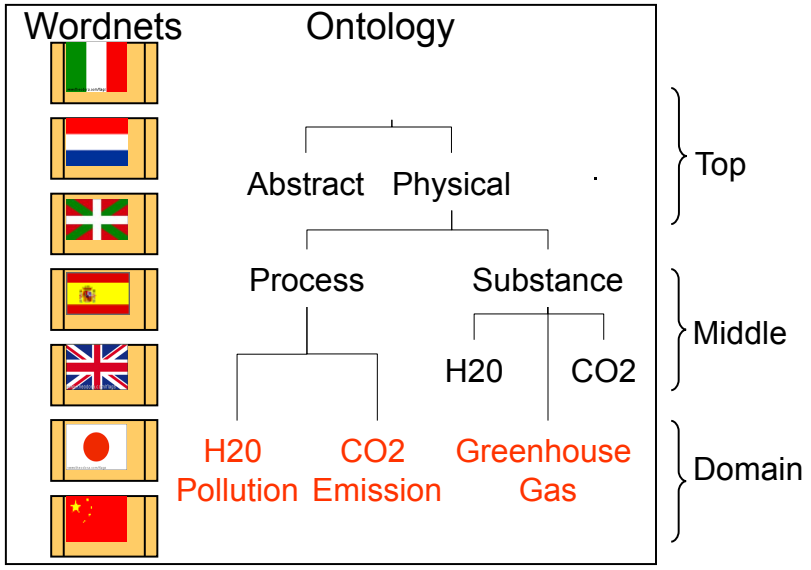
## Social communities: Environmental organizations

WWF ECNC

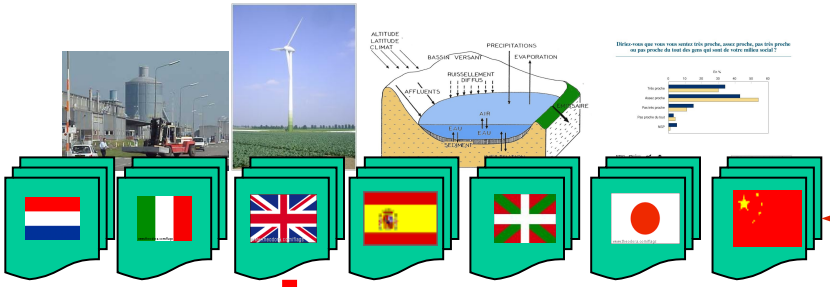
**Process text:**
"Sudden increase of CO2 emissions in 2008 in Europe"

**Tybot: term yielding robot**

CO2 emission



## Distributed, diverse & dynamic data



## Social communities: Environmental organizations

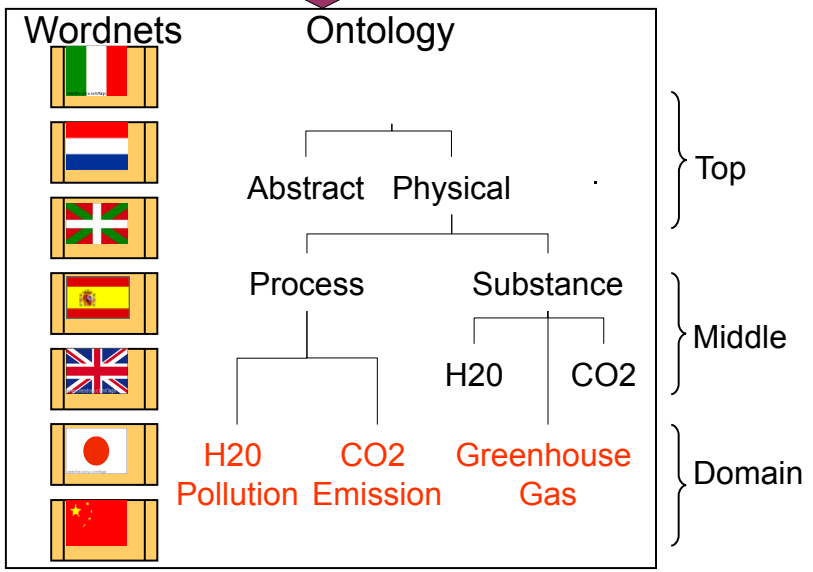
WWF ECNC

Wikyoto Knowledge Editor maintain terms & concepts

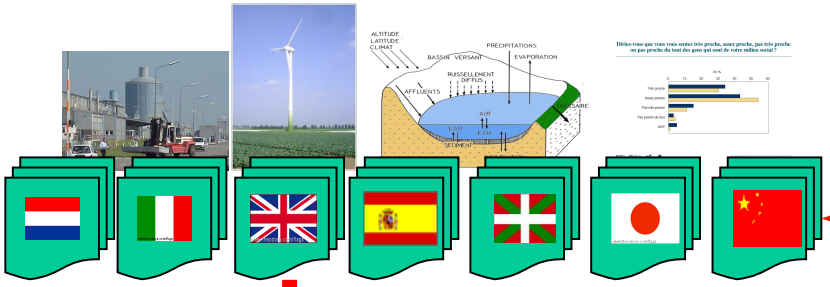
**Process text:**
"Sudden increase of CO2 emissions in 2008 in Europe"

**Tybot: term yielding robot**

CO2 emission



## Distributed, diverse & dynamic data



## Social communities: Environmental organizations

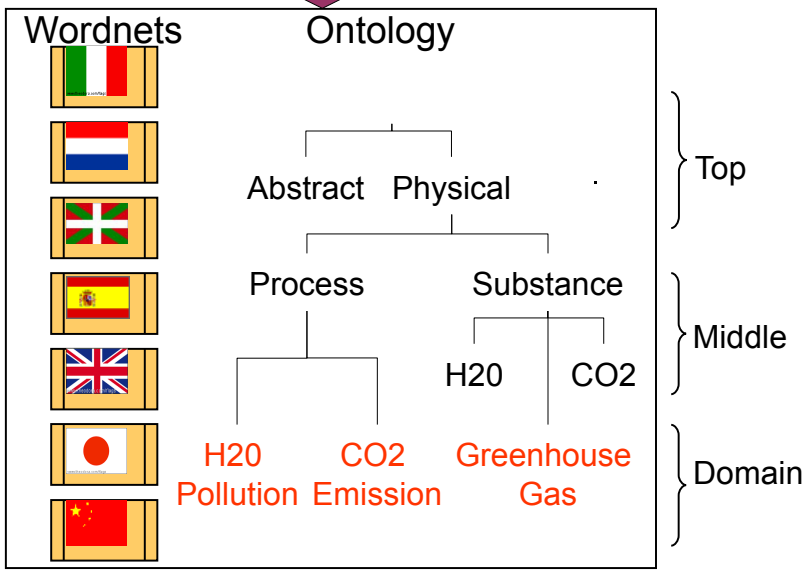
WWF ECNC

Wikyoto Knowledge Editor maintain terms & concepts

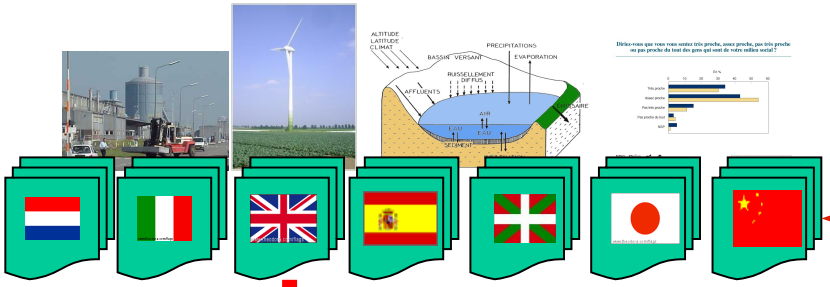
**Process text:**
"Sudden increase of CO2 emissions in 2008 in Europe"

**Tybot: term yielding robot**

CO2 emission



**Distributed, diverse & dynamic data**



**Social communities: Environmental organizations**

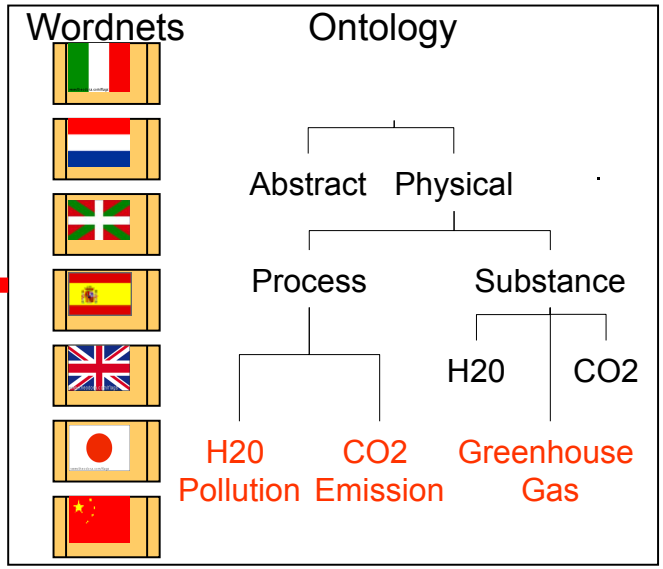
WWF ECNC

**Process text:**
"Sudden increase of CO2 emissions in 2008 in Europe"

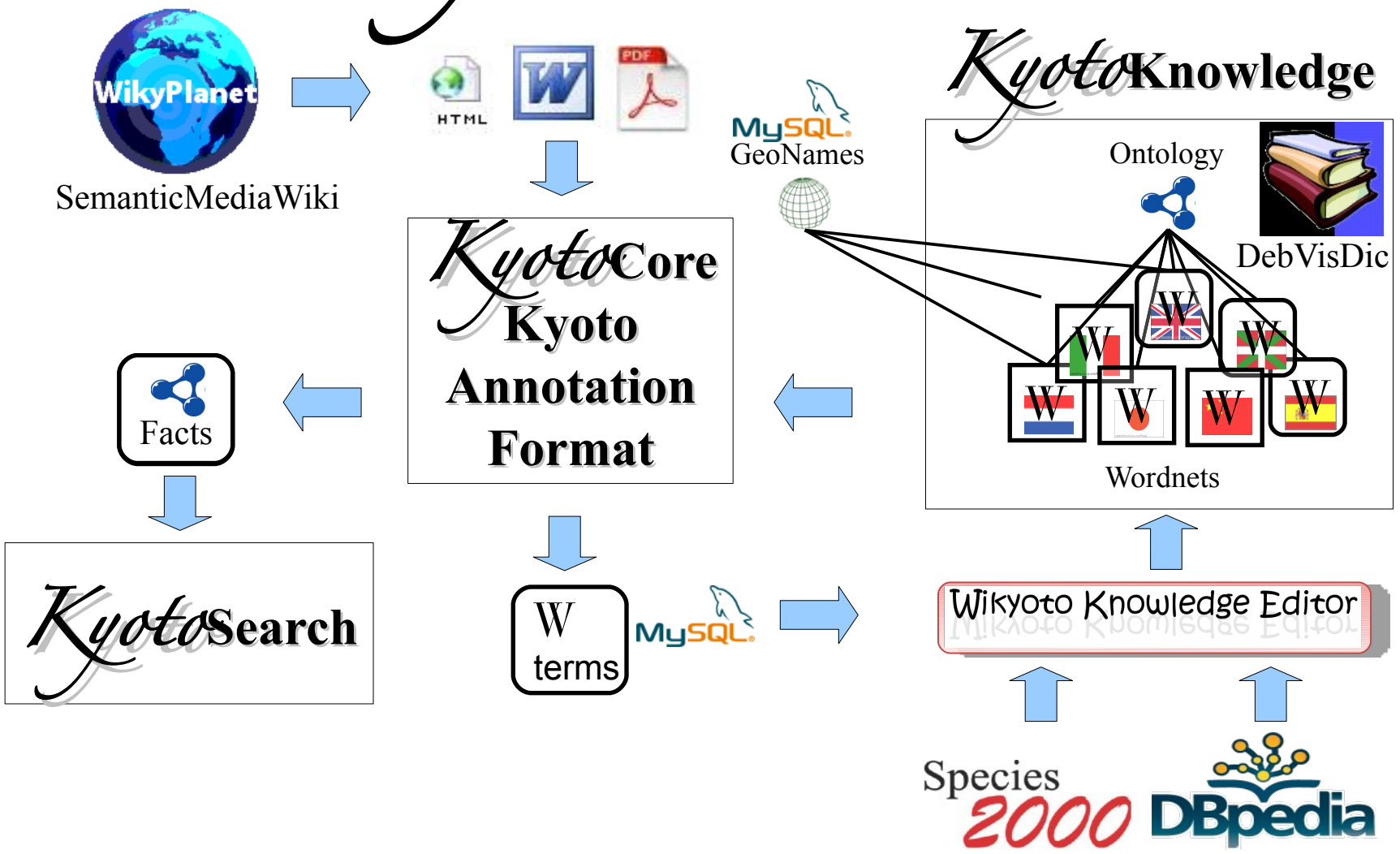
**Kybot: knowledge yielding robot**

**Index facts:**

| | |
|---|---|
| Process: | Emission |
| Involves: | CO2 |
| Property: | increase, sudden |
| When: | 2008 |
| Where: | Europe |



# Kyoto System

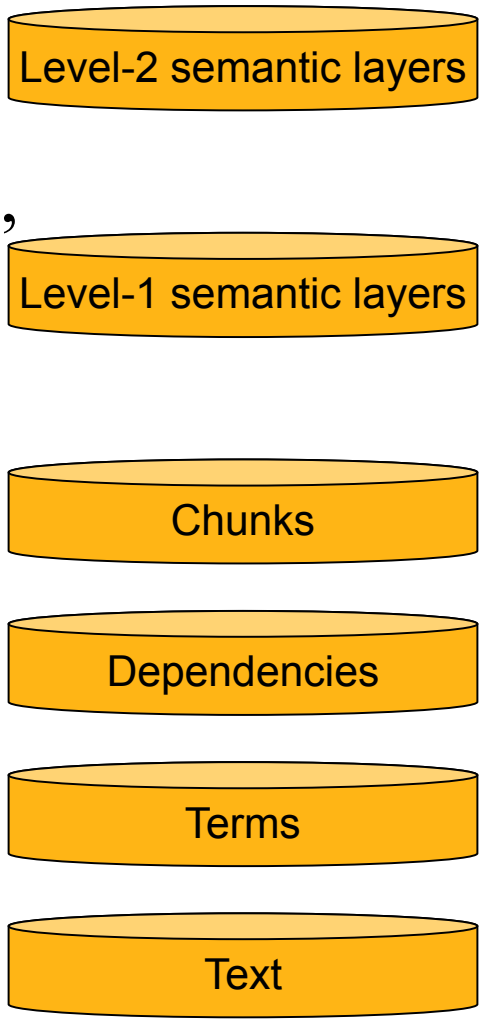


# Kyoto System

- **WikyPlanet**: a semantic media wiki for collecting and sharing textual information in a community;
- **KyotoCore**: pipeline architecture of modules for processing text documents for term and concept extraction and for text mining;
- **Wikyoto**: Wiki platform for editing domain terms and concepts across different languages and cultures;
- **DebVisDic** platform: database system for storing the wordnets and the central ontology;
- **KyotoSearch**: index and search module on events extracted through **KyotoCore**

# 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"/> <!--refers to "migratory"(w2)+"birds"(w3)-->
  </term>
 </terms>
</kaf>
```

# KAF annotation: Semantic layers

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

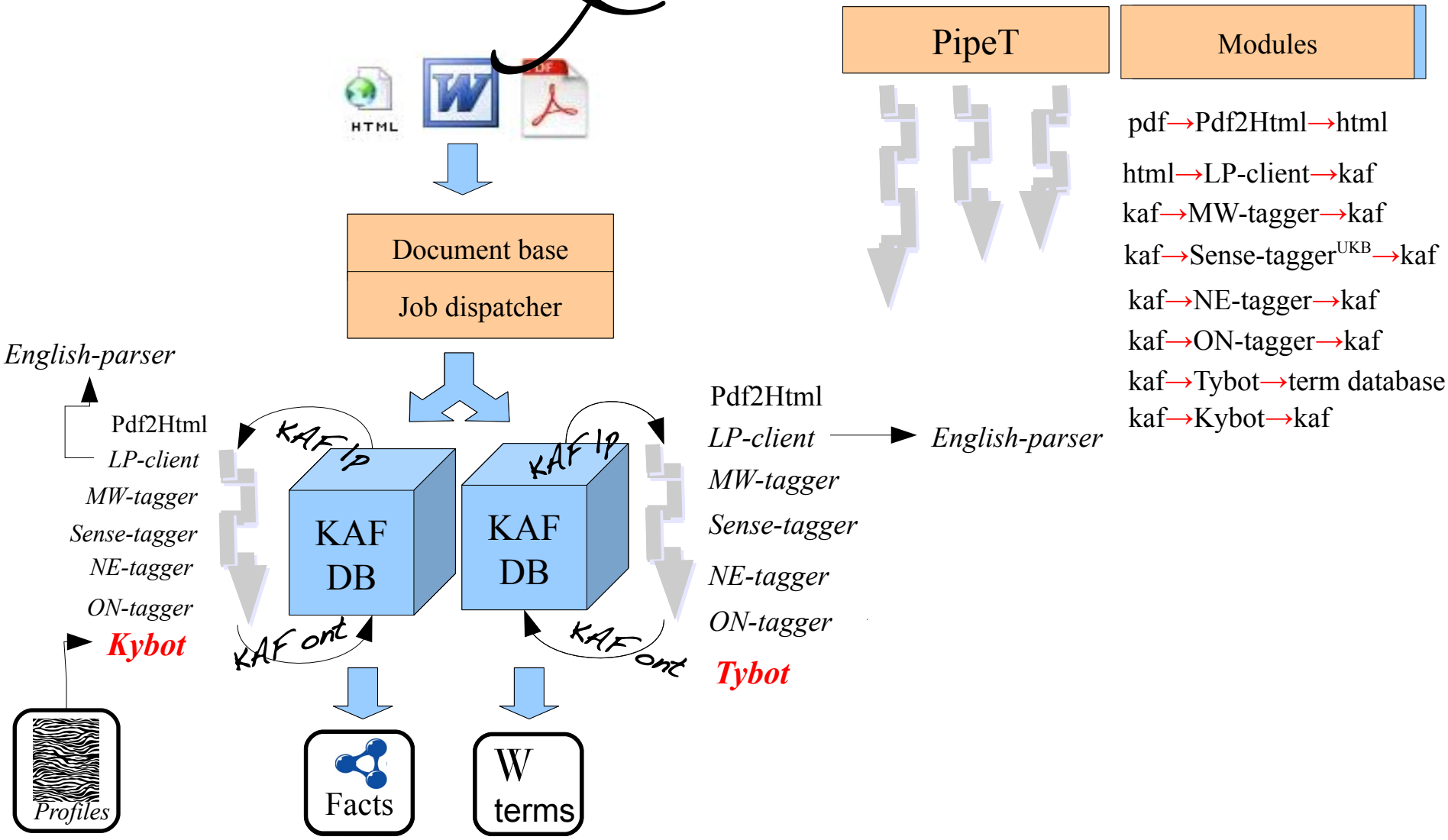
Word-Sense-Disambiguation

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

# KyotoCore


PipeT
Modules
pdf→Pdf2Html→html
html→LP-client→kaf
kaf→MW-tagger→kaf
kaf→Sense-tagger^UKB→kaf
kaf→NE-tagger→kaf
kaf→ON-tagger→kaf
kaf→Tybot→term database
kaf→Kybot→kaf
HTML
Document base
Job dispatcher
English-parser
Pdf2Html
LP-client
MW-tagger
Sense-tagger
NE-tagger
ON-tagger
Kybot
KAF lp
KAF DB
KAF DB
KAF lp
KAF ont
KAF ont
Pdf2Html
LP-client
English-parser
MW-tagger
Sense-tagger
NE-tagger
ON-tagger
Tybot
Profiles
Facts
W terms

# KyotoCore Features

- **PipeT**: a platform for creating pipelines of processing modules through input and output stream connections;
- **Document base:**
  - maintains, documents, databases, users and user privileges
  - stores meta data and multiple representations of the same document
  - assigns pipelines of processing modules to databases;
- **Job dispatcher:**
  - Applies processing  pipelines to databases
  - Continuously monitors the documents in databases, checks their processing status and starts next step in the pipelines;

# 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

# Full knowledge cycle

- Document base databases on Estuaries from English PDFs and web pages: 4,625 source documents, 3,091,842 words in size.
- Term database derived by Tybots with almost 100,000 candidate terms
- Knowledge repository:
  - Ontology extension of DOLCE-Lite with about 1,500 classes
  - Wordnet completely mapped to the ontology: Base Concept mappings (96.328 records), synset to ontology mappings (179.797 records), and explicit ontology mappings (27.983 records)
- Wikyoto: Domain wordnet has 1259 words, 3,260 concepts, 991 mappings to the ontology

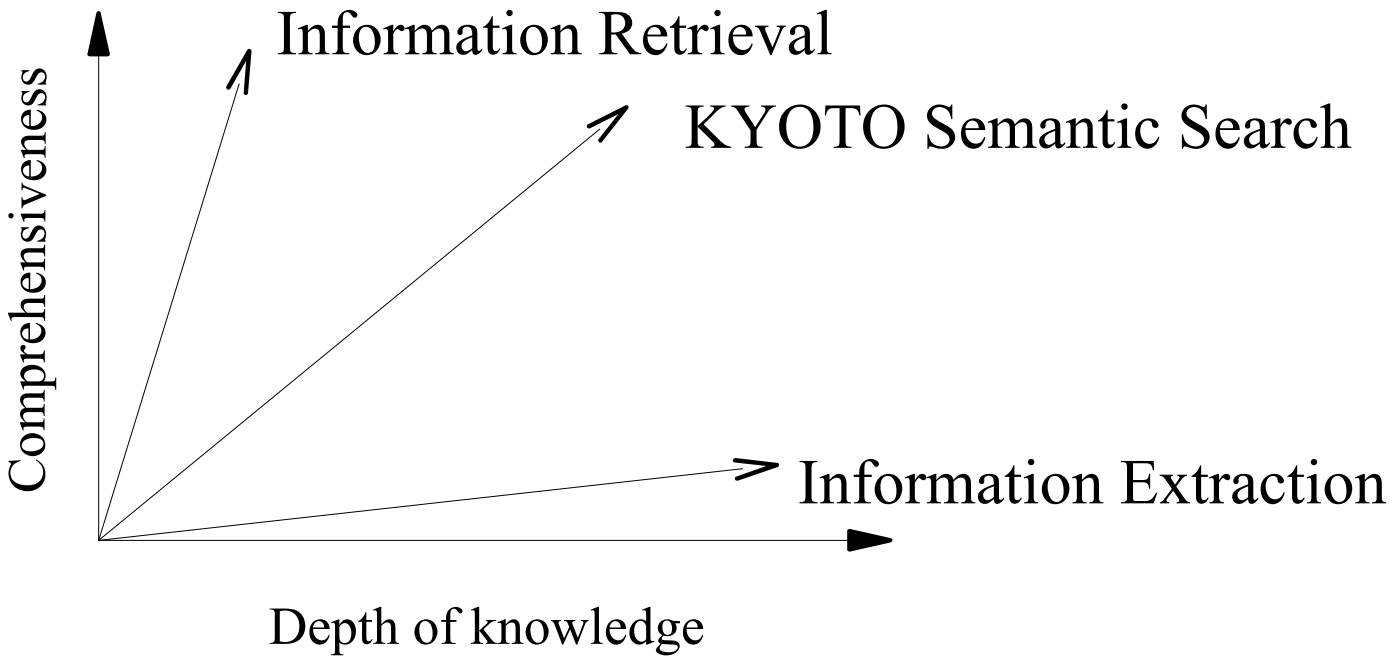
# Full knowledge cycle

- 260 generic Kybot profiles for English using ontology classes and basic patterns
- Kybots generated 1 million information triplets:
  - 118,255 events with 245,563 involved participants, 317,749 dates, 271,734 place relations and 64,604 mappings to countries.
  - Dates and places are entities mapped to ISO dates and GeoNames locations: 5,075 unique locations and 1,587 dates
- Semantic search on the output of the Kybots

# Relations extracted for Estuary database

| Relation | Nr. participants | Relation | Nr. participants | Relation | Nr. participants |
|---|---|---|---|---|---|
| destination-of | 11,033 | part-of | 2,464 | source-of | 5,185 |
| done-by | 37,096 | patient | 131,662 | state-of | 2,575 |
| generic-location | 15,883 | purpose-of | 8,570 | use-of | 2,093 |
| has-state | 5,278 | simple-cause-of | 23,724 | | |

# Application range KYOTO



# What happens after KYOTO

- Project results are available as open-source
- Extend to other languages
- Extend to other domains
- Collaborate with standardization efforts and language-technology infra-structure projects
- Improve scalability
- Improve precision and recall
- Extend types of knowledge

# Kyoto System

4,625 sources
3 million words


SemanticMediaWiki


MySQL
GeoNames

Kyoto Knowledge

Ontology

DebVisDic

Wordnets

Kyoto Core
Kyoto Annotation Format

1 million facts

Facts

Kyoto Search

W terms

MySQL

100,000 terms

Wikyoto Knowledge Editor

Species 2000

DBpedia