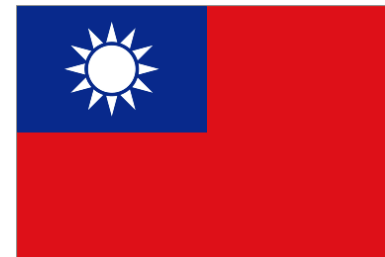
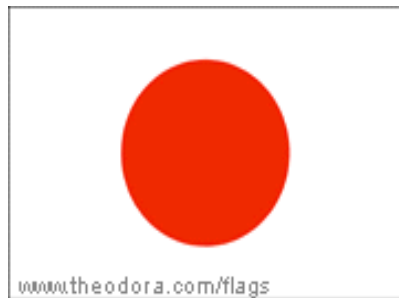




European-Asian project  
<http://www.kyoto-project.eu/>



# Knowledge Mining in Kyoto

- Generic concept-based fact mining system
- Try to represent all factual information in text:
  - Fact = event+participants+role+time+place
- Using generic semantic models:
  - wordnets+shared ontology
- In a uniform and interoperable way for **7 different languages**:
  - English, Dutch, Spanish, Basque, Italian
  - Japanese and Chinese
- Tested on the 10,000 documents;
- **KAF** (Kyoto Annotation Format) is the input of both:
  - Tybot: term extraction
  - Kybot: fact extraction



# KAF

- Based on current ISO proposals
- Language-neutral annotation of text,
  - concepts, facts,...
  - Multilingual
- Interoperable across linguistic processors
  - KAF is the basis for integration
- Incremental
  - Modules reads KAF and produces KAF
- Flexible and extendible

# Linguistic Processors

- KAF (Kyoto Annotation Format)
  - English: **Synthema**
  - Dutch: **VUA**
  - Italian: **Synthema**
  - Basque: **EHU**
  - Spanish: **EHU**
  - Chinese: **AS**
  - Japanese: **NICT**
  
- Pdf2html: **Irion**
- MW tagger: **VUA**
- WSD module (UKB): **EHU**
- NE Tagger: **Irion**
- OntoTagger: **CNR-ILC, EHU**
  
- PipeT: **VUA**



# KAF

- **KAF** XML files include sections for:
  - Word forms
  - Terms / Items
  - Chunks: grouping of sequences of terms
  - Dependencies: syntactic relations between terms
  - WSD: WN senses of the term
  - Ontological references of the term:
    - Base Concepts
    - Explicit ontology
  - Events
  - Locations, Time expressions
  - ...

# KAF: Layers for text and terms

<kaf>

<text>

<wf wid="w1" page="1" sent="1" para="1" fileoffset="0,4">large</wf>

<wf wid="w2" page="1" sent="1" para="1" fileoffset="6,14">migratory</wf>

<wf wid="w3" page="1" sent="1" para="1" fileoffset="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"/>

<!--refers to "migratory"(w2)+"birds"(w3)-->

</term>

</terms>

</kaf>



# KAF: Layers for chunks and dependencies

<kaf>

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

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

<chunks>

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

<span id="t1"/><!-- refers to term: "large" -->

<span id="t2"/><!-- refers to term: "migratory bird" -->

</chunk>

</chunks>

<deps>

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

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

</deps>

</kaf>



# KAF: Layers for semantic annotation

```
<term tid="t4" type="open" lemma="population" pos="N">
```

```
  <span><target id="w4"/></span>
```

```
<externalReferences>
```

```
  < externalRef resource="WN-3.0" ref="EN-30-00859568-n" conf="0.80 "/>
```

```
  < externalRef resource="WN-3.0" ref="EN-30-00257849-n" conf="0.13 />
```

```
  < externalRef resource="WN-3.0" ref="EN-30-00962397-n" conf="0.07 />
```

```
  <externalRef resource="DOLCE-1.0" ref="dolce#group" conf="0.80"/>
```

```
</externalReferences>
```

```
</term>
```





# KYOTO Integrated System

- Document Base & Term Base
- Capture
- Pipeline
- Linguistic Processors
- Multiword Recognition
- Word Sense Disambiguation
- Named Entity Recognition
- OntoTagger
  
- KAF annotator
- KAF SAX parser
- KAF Debugger
- 
- <http://ixa2.si.ehu.es/demokaf/demokaf.pl>



# From Cyc

- Fred saw the plane flying over Zurich.
- Fred saw the train flying over Zurich.





European-Asian project  
<http://www.kyoto-project.eu/>

