KYOTO
a platform for anchoring textual meaning across languages

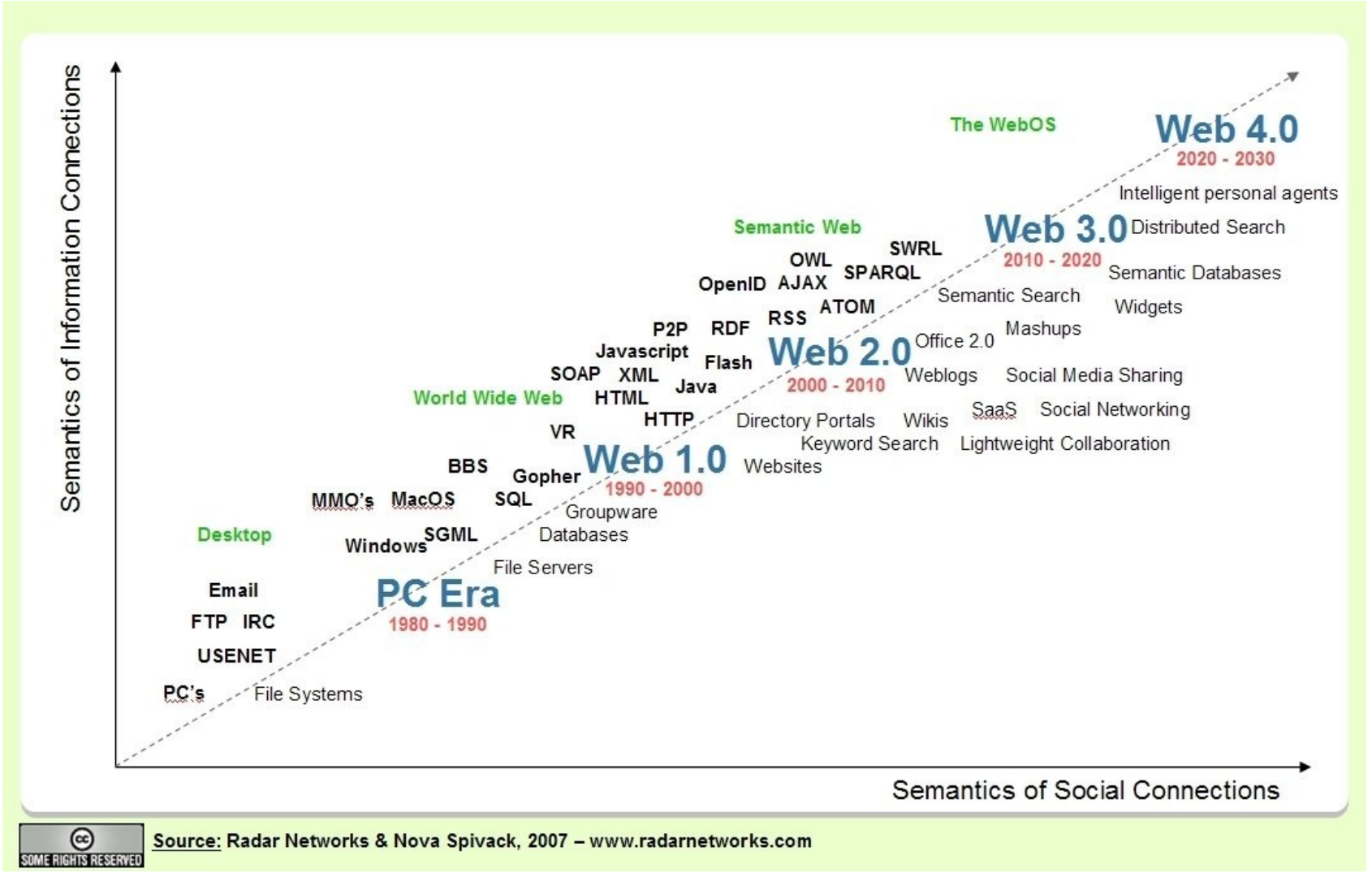
Piek Vossen
VU University Amsterdam
p.vossen@let.vu.nl
www.kyoto-project.nl

W3C Workshop:
The Multilingual Web - Where Are We?
26-27 October 2010, Madrid

Why translate text if you can mine text and represent the knowledge and information in a language neutral form?

Warning: older versions of the web are not going to disappear!

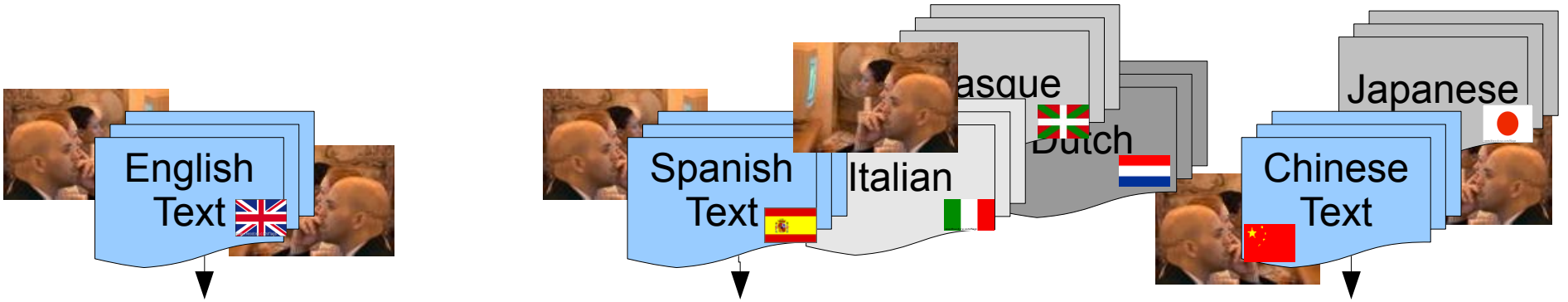
Evolution of the web

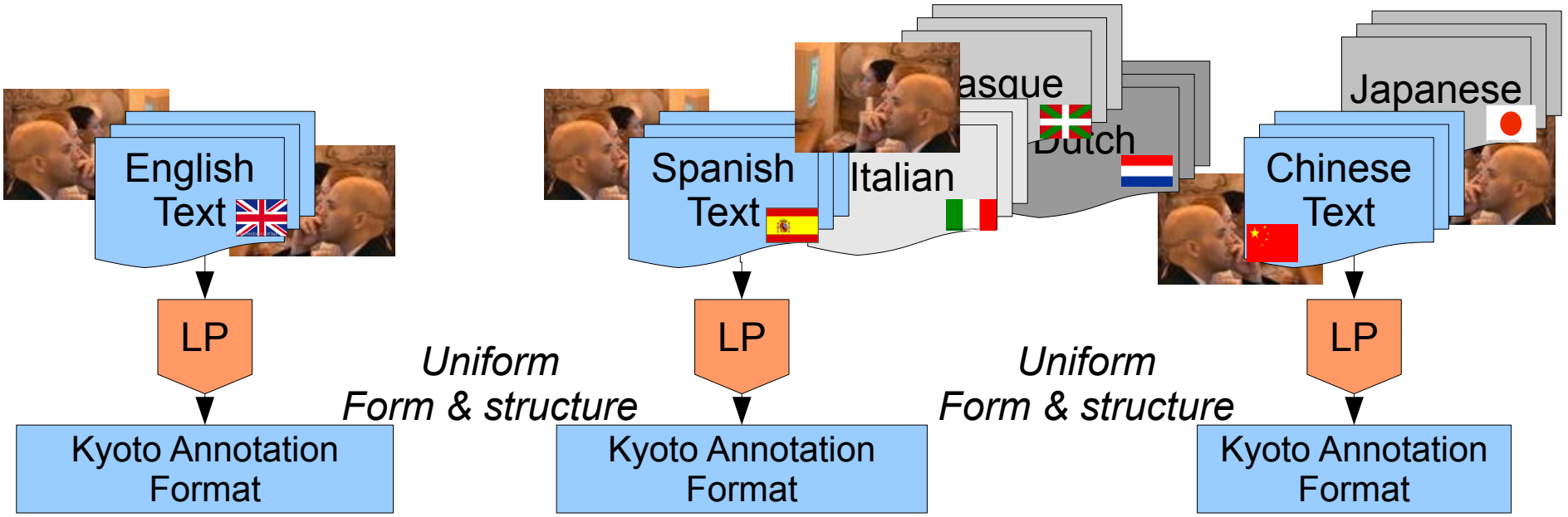


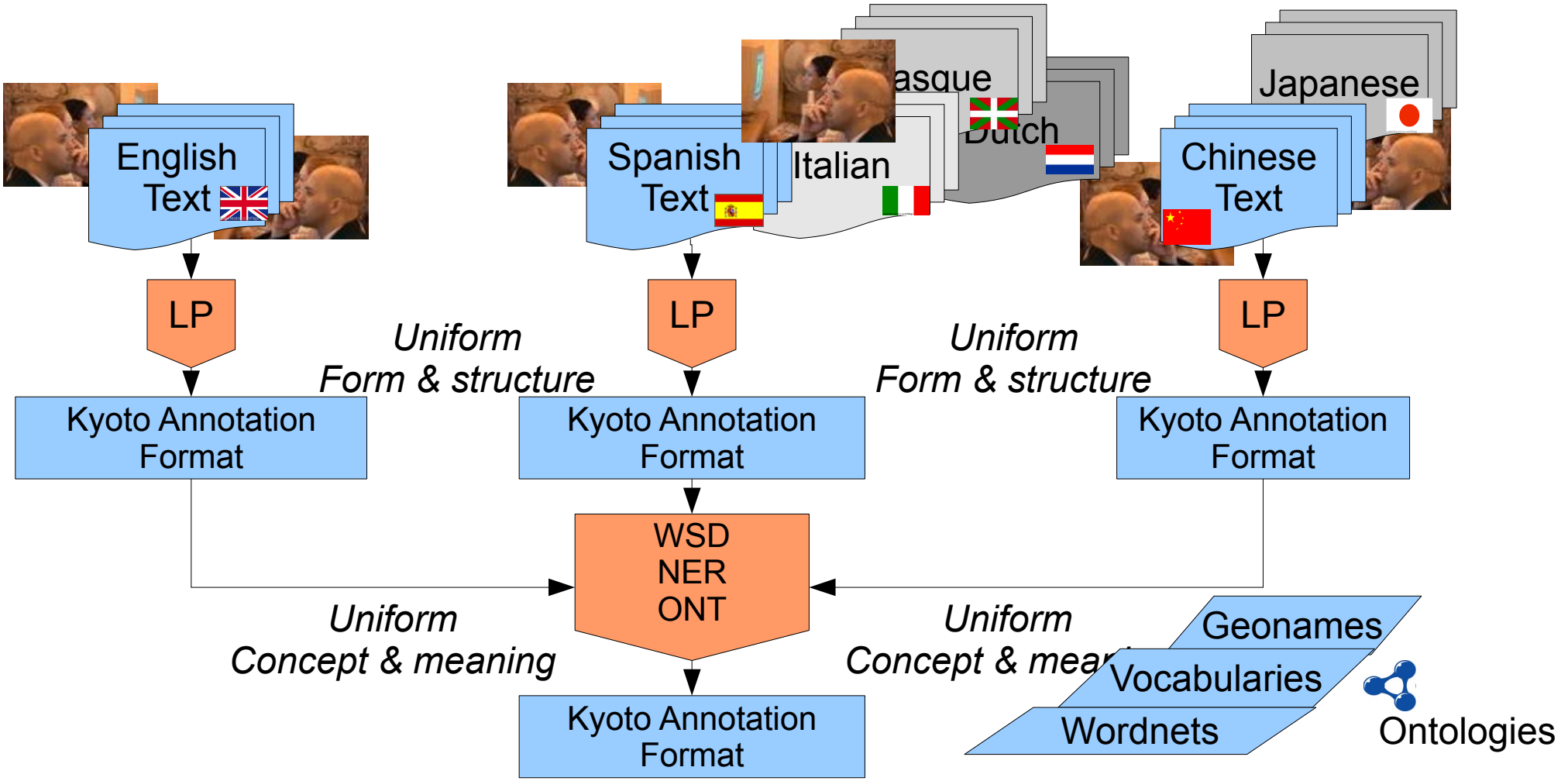
Source: Radar Networks & Nova Spivack, 2007 – www.radarnetworks.com

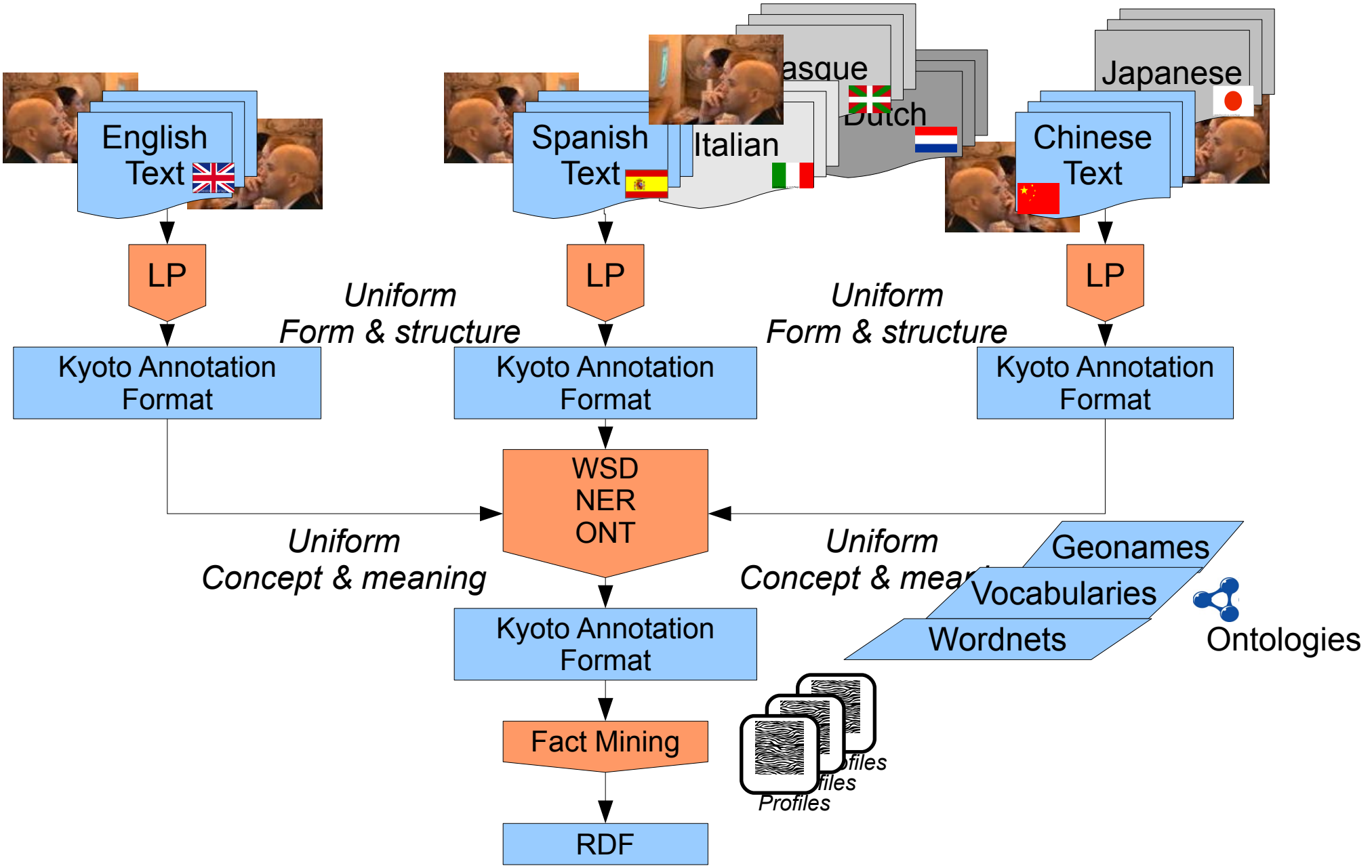
How to connect different versions of the web?

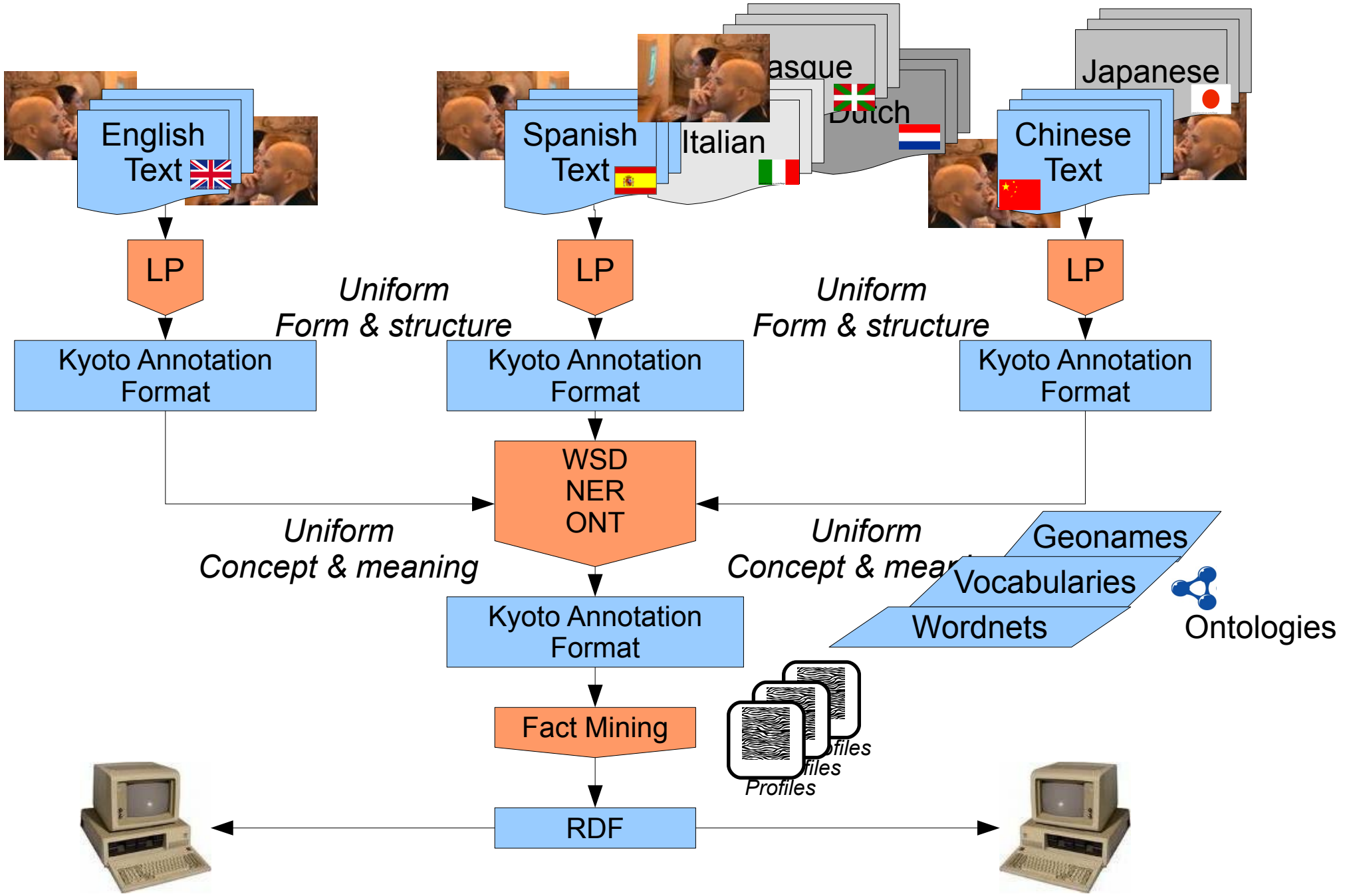
- Interoperable representation of the structure of language
- Interoperable representation of formal conceptual knowledge
- Methods to map natural language of Web1 and Web2 to the formal interoperable representations that can be used in Web3 and that allow agents to join Web2 in Web4

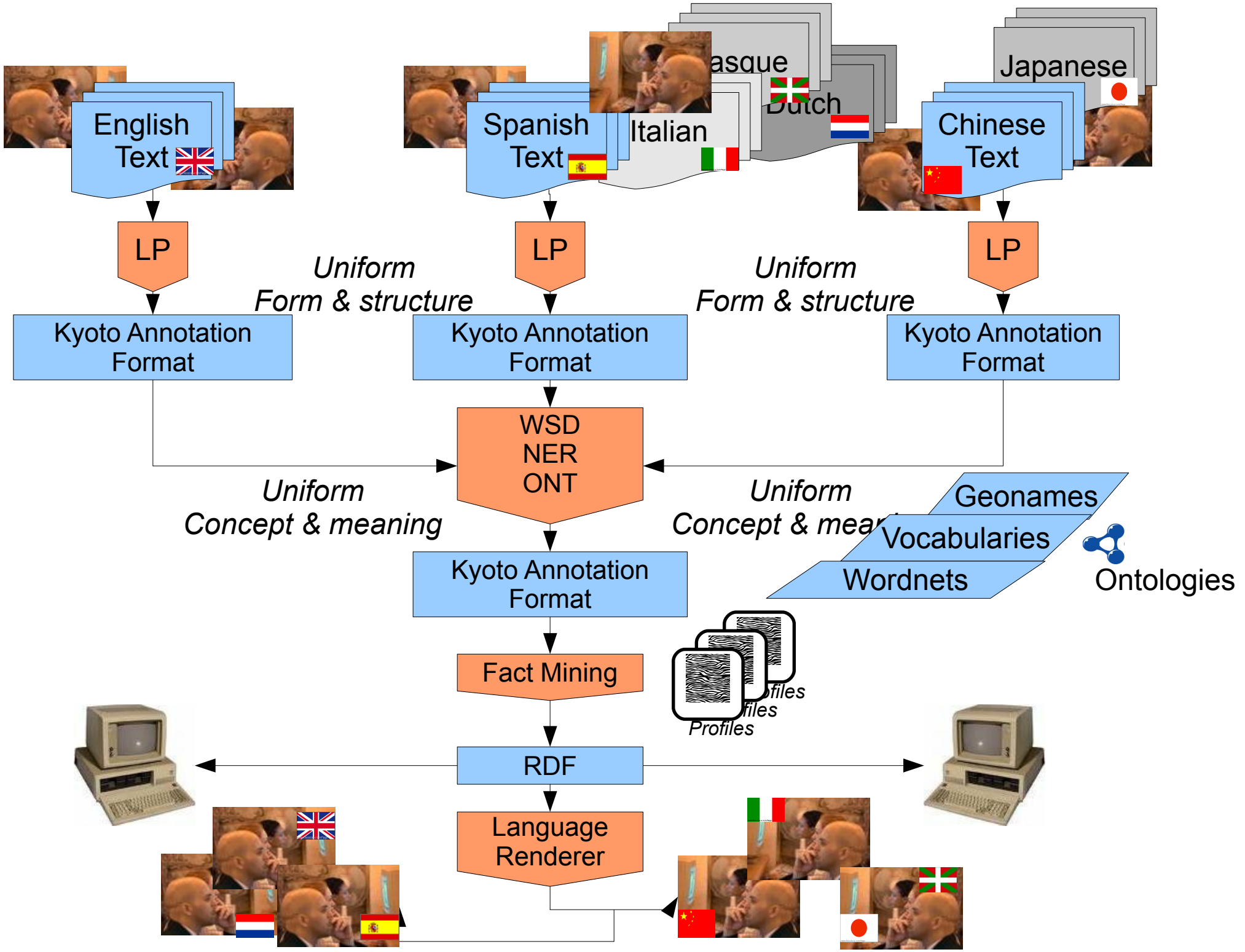

English
Text
Spanish
Text
Italian
Dutch
Japanese
Chinese
Text


English
Text
LP
Kyoto Annotation
Format
Uniform
Form & structure
Spanish
Text
Italian
asque
Dutch
LP
Kyoto Annotation
Format
Uniform
Form & structure
Japanese
Chinese
Text
LP
Kyoto Annotation
Format


English Text
Spanish Text
Italian
Dutch
asque
Japanese
Chinese Text
LP
LP
LP
Uniform
Form & structure
Uniform
Form & structure
Kyoto Annotation Format
Kyoto Annotation Format
Kyoto Annotation Format
WSD
NER
ONT
Uniform
Concept & meaning
Uniform
Concept & mean
Kyoto Annotation Format
Geonames
Vocabularies
Wordnets
Ontologies

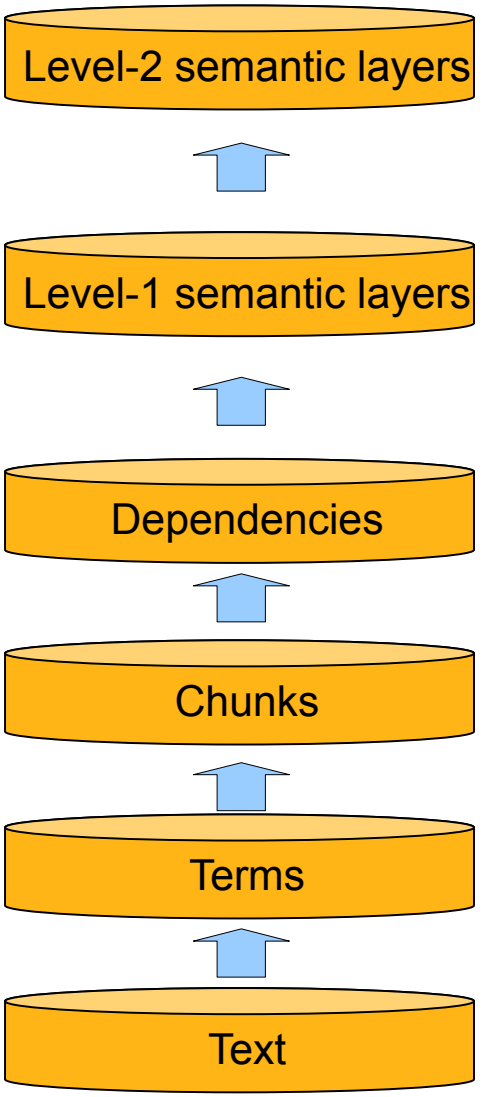

English Text
Spanish Text
Italian
Dutch
asque
Japanese
Chinese Text
LP
LP
LP
Uniform
Form & structure
Uniform
Form & structure
Kyoto Annotation Format
Kyoto Annotation Format
Kyoto Annotation Format
WSD
NER
ONT
Uniform
Concept & meaning
Uniform
Concept & mean
Geonames
Vocabularies
Wordnets
Ontologies
Kyoto Annotation Format
Fact Mining
Profiles
RDF


English Text
Spanish Text
Italian
Dutch
asque
Japanese
Chinese Text
LP
LP
LP
Uniform
Form & structure
Uniform
Form & structure
Kyoto Annotation Format
Kyoto Annotation Format
Kyoto Annotation Format
WSD
NER
ONT
Uniform
Concept & meaning
Uniform
Concept & mean
Geonames
Vocabularies
Wordnets
Ontologies
Kyoto Annotation Format
Fact Mining
Profiles
RDF


English Text
Spanish Text
Basque
Italian
Dutch
Chinese Text
Japanese
LP
LP
LP
Uniform
Form & structure
Uniform
Form & structure
Kyoto Annotation Format
Kyoto Annotation Format
Kyoto Annotation Format
WSD
NER
ONT
Uniform
Concept & meaning
Uniform
Concept & meaning
Geonames
Vocabularies
Wordnets
Ontologies
Kyoto Annotation Format
Fact Mining
Profiles
RDF
Language Renderer

Kyoto Annotation Format (KAF)

- Stands off annotation based on Layered Annotation Format or LAF (Ide and Romary 2002)
  - **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



Kyoto Annotation Format 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

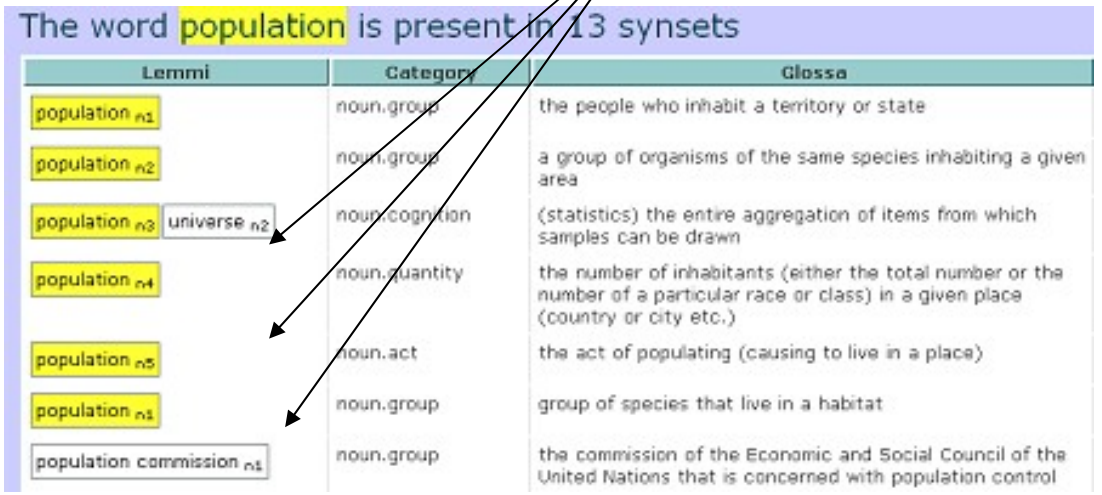
```
<kaf>
  <text>...</text><!-- defines w1, w2, w3 -->
  <terms>...</terms><!-- defines t1, t2 -->

  <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”>
      <span id=”t1”/><!-- refers to term: ”large” -->
      <span id=”t2”/><!-- refers to term: ”migratory bird” -->
    </chunk>
  </chunks>
</kaf>
```

Kyoto Annotation Format Semantic layers

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



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

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 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 mining 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 mining 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">

  </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>
```

Kybot mining 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>
```

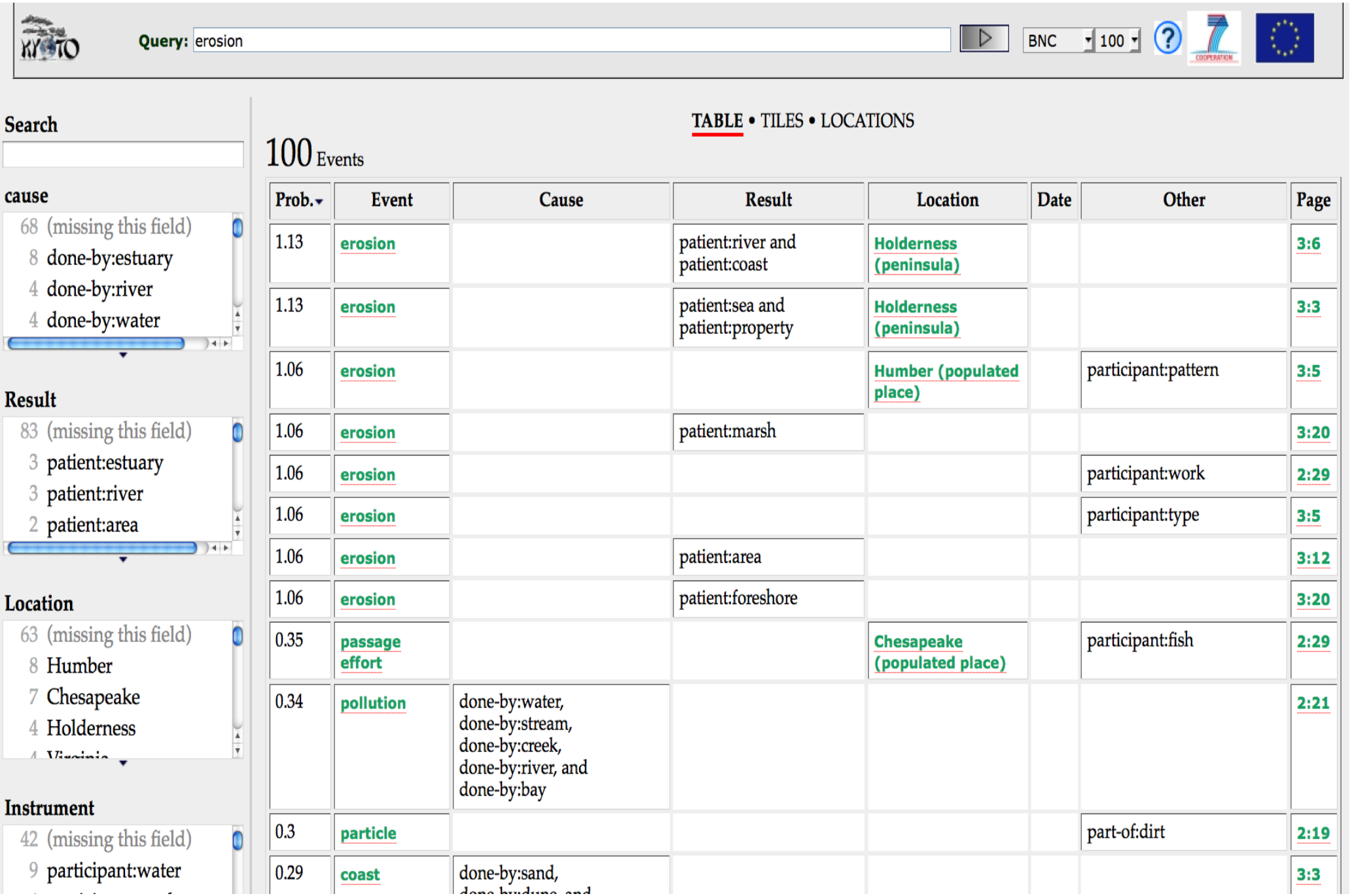
Evaluation: triplet example

"…. in 2008 (w12221). 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 →

First results for English October-27th-2010

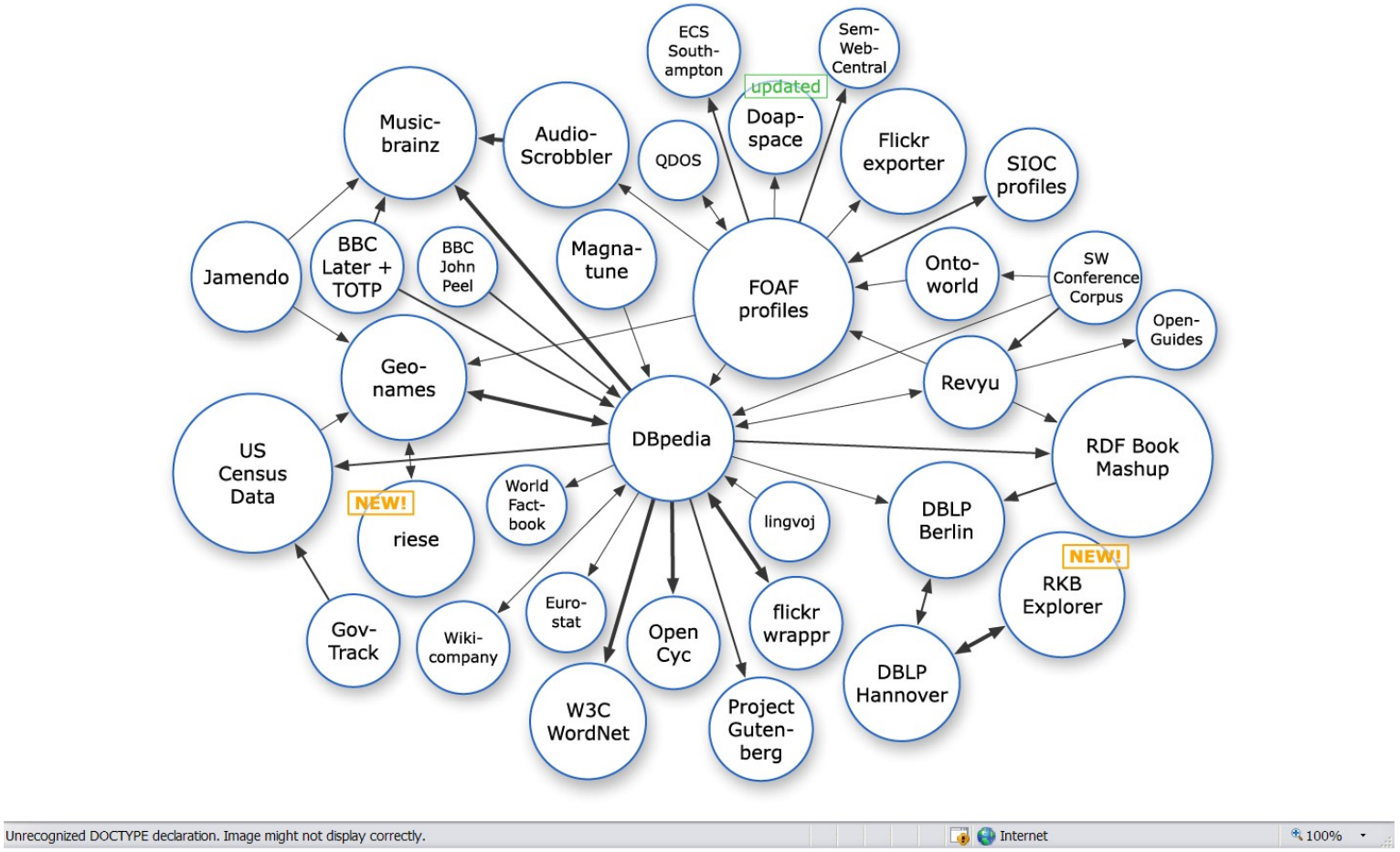
- Single document on Chesapeake Bay: 16,145 words
- Gold standard 348 event triplets
- System output: 968 event triplets
- Totally 9453 event triplets using 235 generic profiles
- Precision 31%, recall 71%



Prob.	Event	Cause	Result	Location	Date	Other	Page
1.13	erosion		patient:river and patient:coast	Holderness (peninsula)			3:6
1.13	erosion		patient:sea and patient:property	Holderness (peninsula)			3:3
1.06	erosion			Humber (populated place)		participant:pattern	3:5
1.06	erosion		patient:marsh				3:20
1.06	erosion					participant:work	2:29
1.06	erosion					participant:type	3:5
1.06	erosion		patient:area				3:12
1.06	erosion		patient:foreshore				3:20
0.35	passage effort			Chesapeake (populated place)		participant:fish	2:29
0.34	pollution	done-by:water, done-by:stream, done-by:creek, done-by:river, and done-by:bay					2:21
0.3	particle					part-of:dirt	2:19
0.29	coast	done-by:sand,					3:3

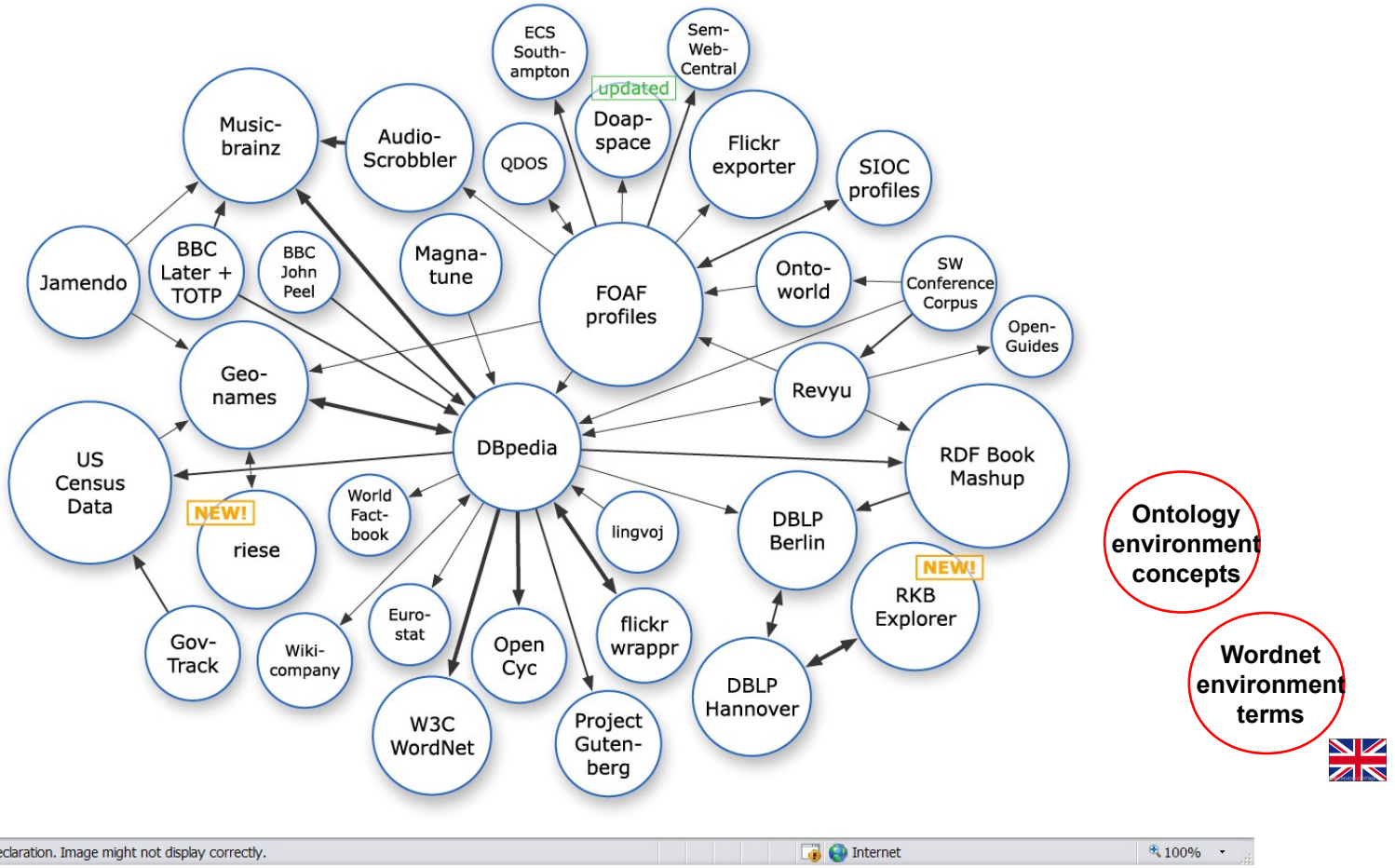
Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



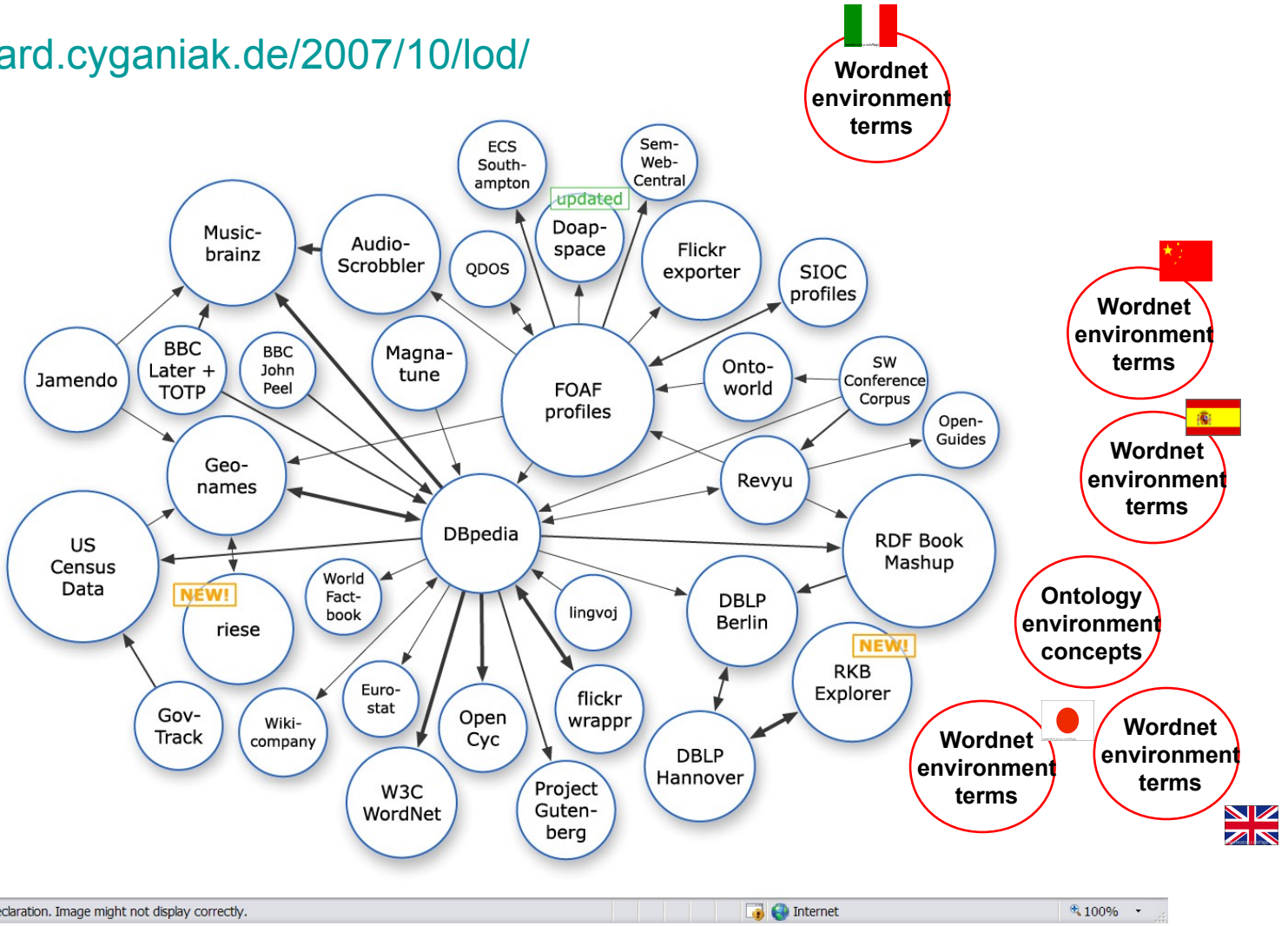
Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



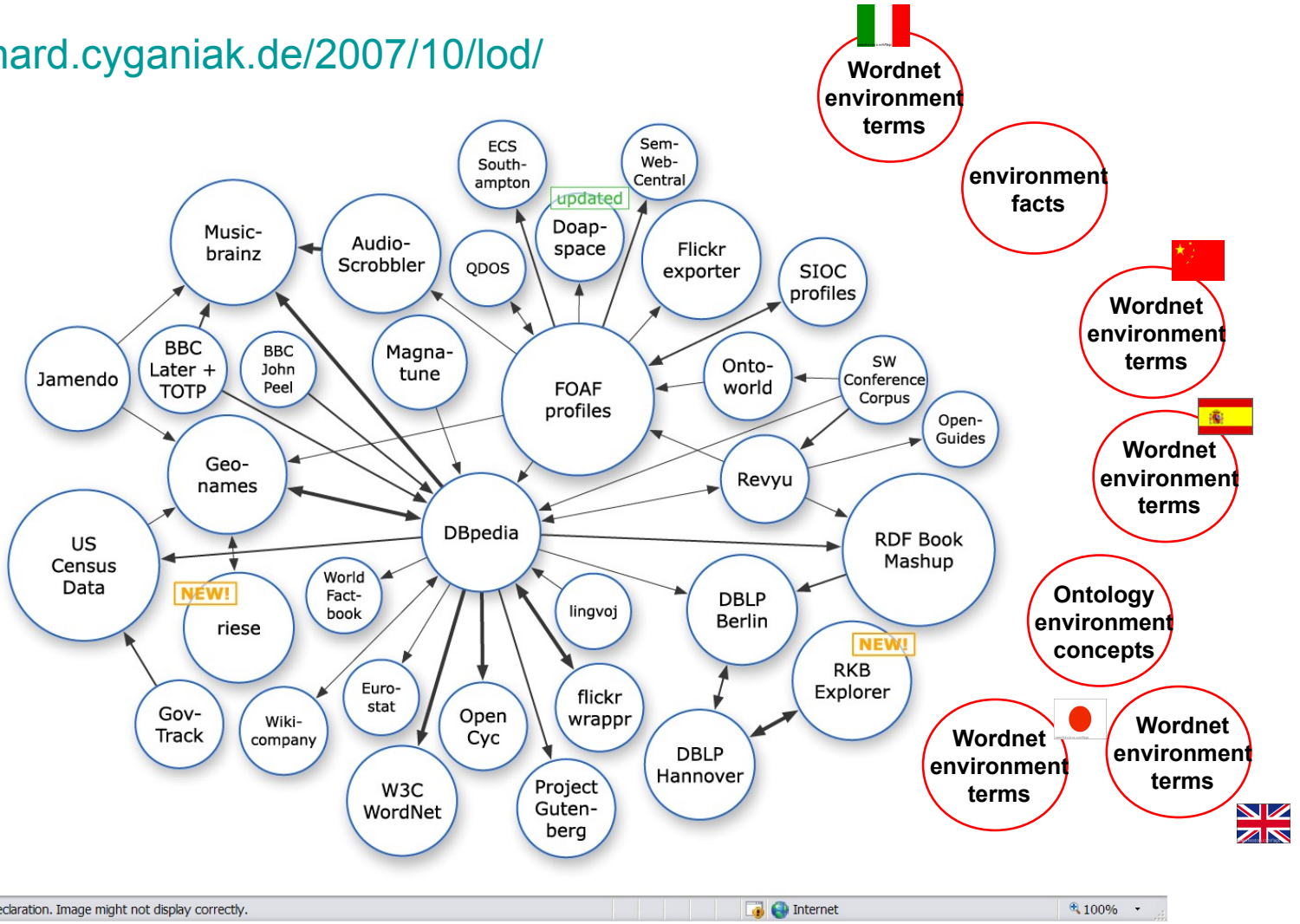
Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



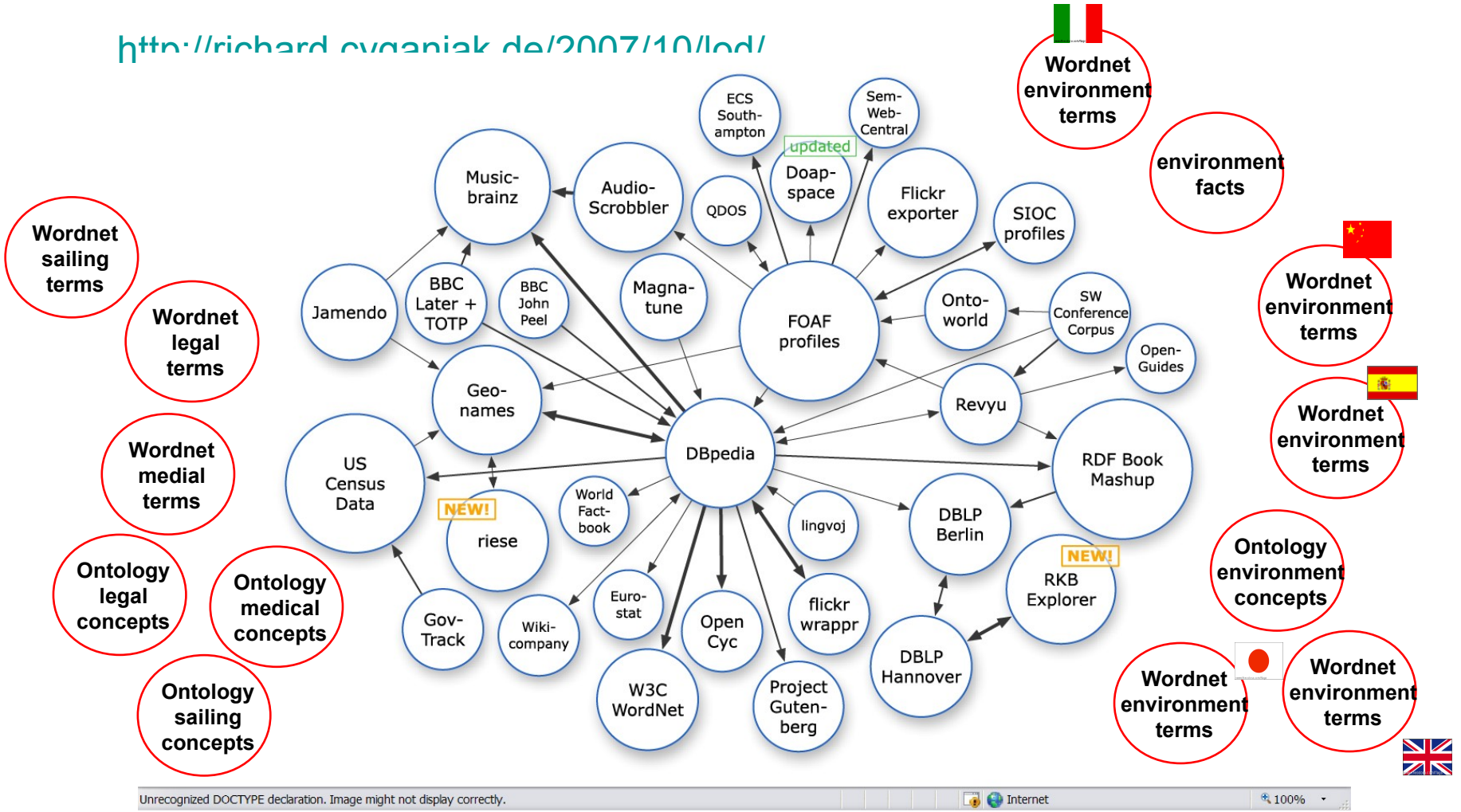
Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



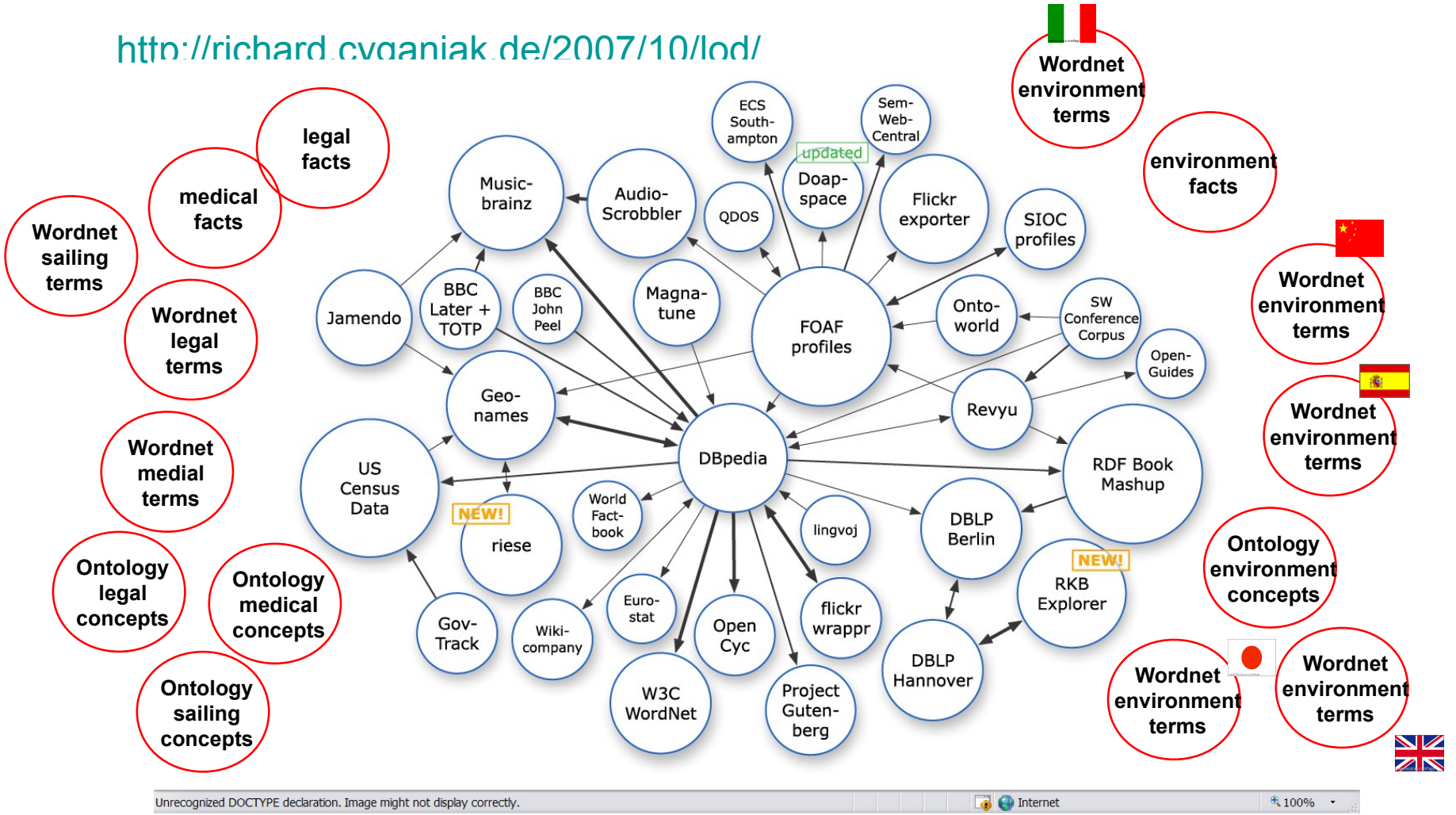
Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



Linking Open Data dataset cloud

http://richard.cyganiak.de/2007/10/lod/



Conclusions

- We should focus on mining textual data across language to convert web1 and web2 textual data to web3 RDF
- For this we need a uniform representation of text across different languages
- For this we need to anchor the vocabularies of all languages to a common conceptual backbone
- We need to focus on how to represent complex mined information in RDF
- We need to develop renderers of complex information in all languages