Integrating WordNet and Wiktionary with lemon

John McCrae[1], Elena Montiel-Ponsoda[2] and Philipp Cimiano[1]

[1] Cognitive Interaction Technology Exzellenzcluster, Universität Bielefeld

[2] Ontology Engineering Group, Universidad Politécnica de Madrid

Monnet is supported by the European Union under Grant No. 248458

Outline

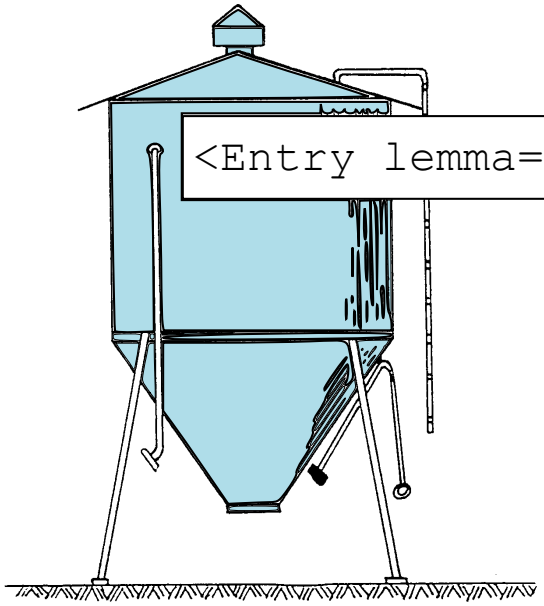
Outline

The need for lexical linked data

- Much lexical data is in “data silos”
  - Proprietary formats
  - Restricted access
- The *Linking Open Data* project fosters:
  - Publication using RDF
  - Linking between resources
- We need **open** and **RDF-native** formats for language resources
  - **lemon** - **Le**xicon **M**odel for **O**ntologies
  - Development under W3C OntoLex community group

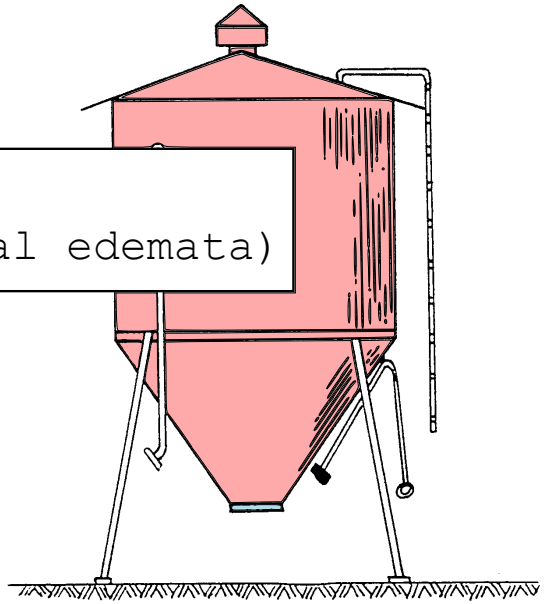
Outline

Stage 0: Data silos



```
<Entry lemma="edema" pos="NP"/>
```

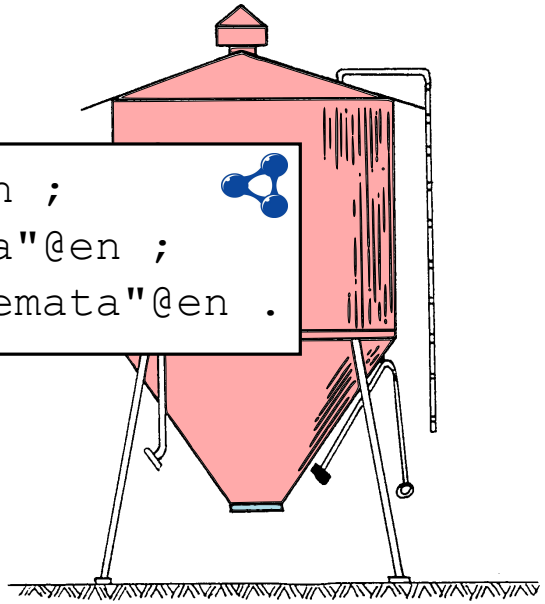
```
Noun:
  edema (plural edemata)
```



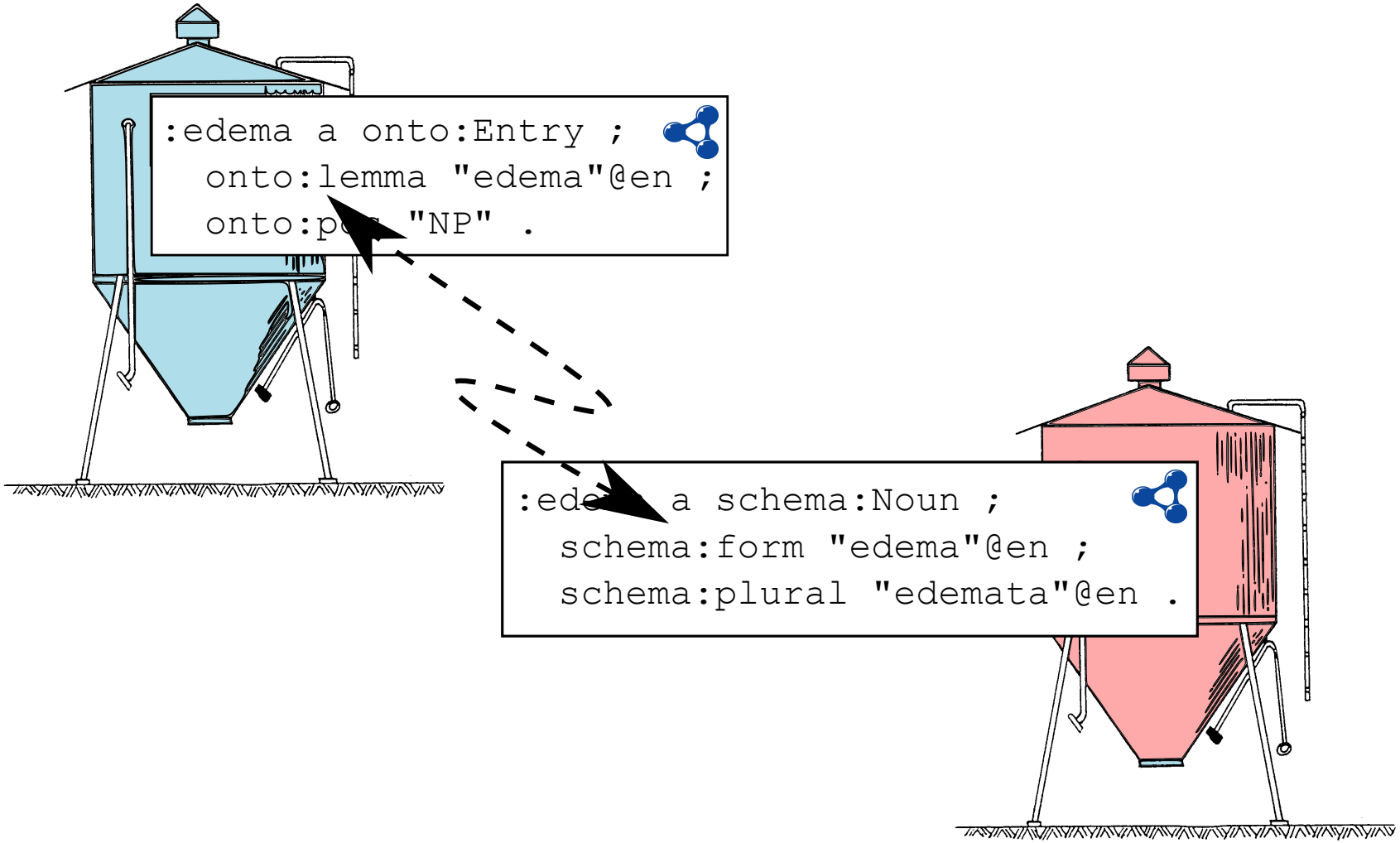
Stage 1: Syntactically interoperable

```
:edema a onto:Entry ;
  onto:lemma "edema"@en ;
  onto:pos "NP" .
```

```
:edema a schema:Noun ;
  schema:form "edema"@en ;
  schema:plural "edemata"@en .
```



Stage 2: Linked



Stage 3: Structurally interoperable

```
:edema a lemon:LexicalEntry ;
  lemon:canonicalForm [
    lemon:writtenRep "edema"@en
  ] ;
  onto:pos "NP" .
```

lemon

```
:edema a lemon:LexicalEntry , onto:Noun ,
  lemon:canonicalForm [
    lemon:writtenRep "edema"@en
  ] ;
  lemon:otherForm [
    lemon:writtenRep "edemata"@en ;
    schema:number schema:plural
  ].
```

lemon

Stage 4: Semantically interoperable

```
:edema a lemon:LexicalEntry ;
  lemon:canonicalForm [
    lemon:writtenRep "edema"@en
  ] ;
  onto:pos "NP" .
```

lemon

penn-syntax.owl

OLiA

ISOcat

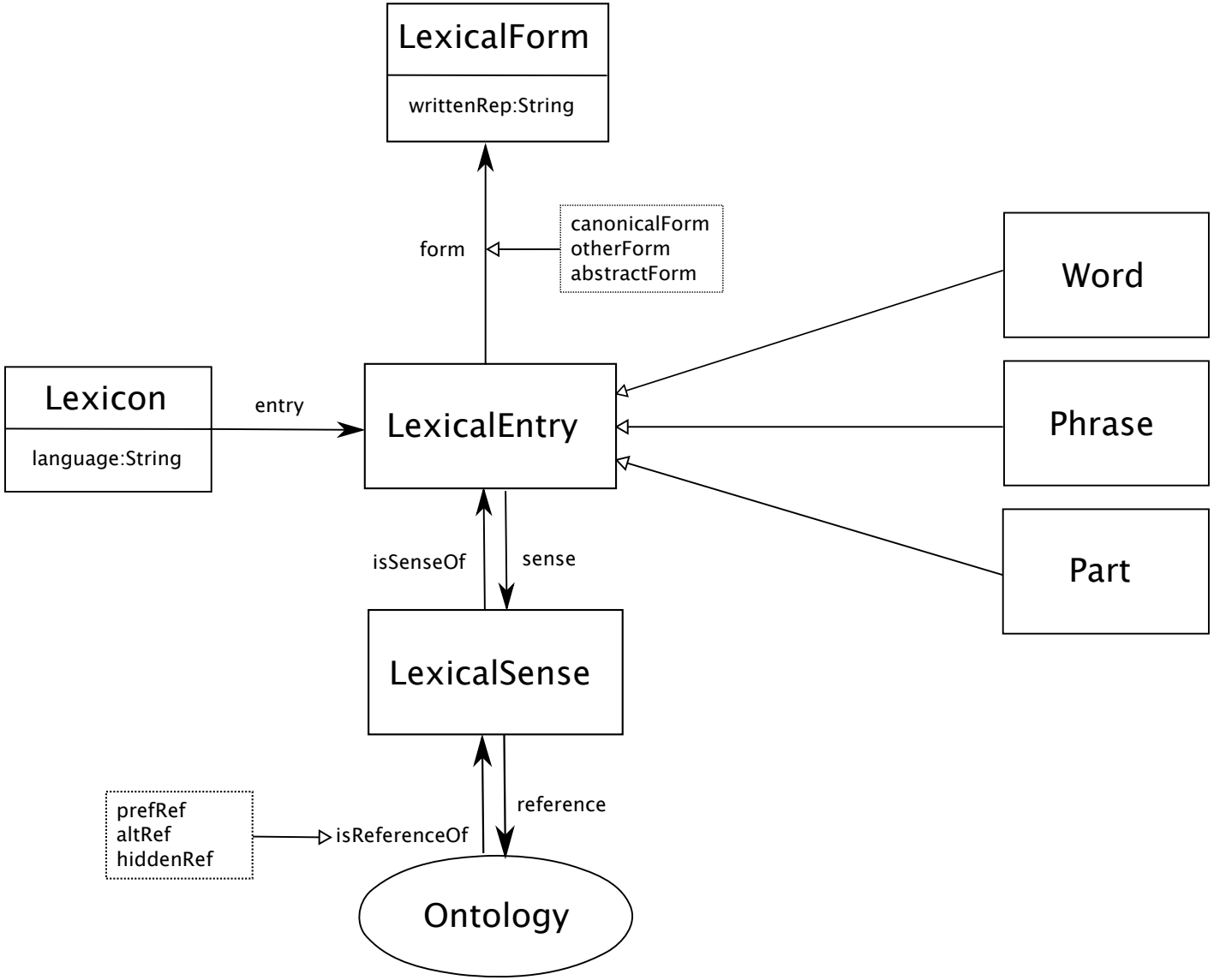
```
:edema a lemon:LexicalEntry , onto:Noun ;
  lemon:canonicalForm [
    lemon:writtenRep "edema"@en
  ] ;
  lemon:otherForm [
    lemon:writtenRep "edemata"@en ;
    schema:number schema:plural
  ].
```

DC-1333

lemon

Outline

The core of lemon



lemon's origins

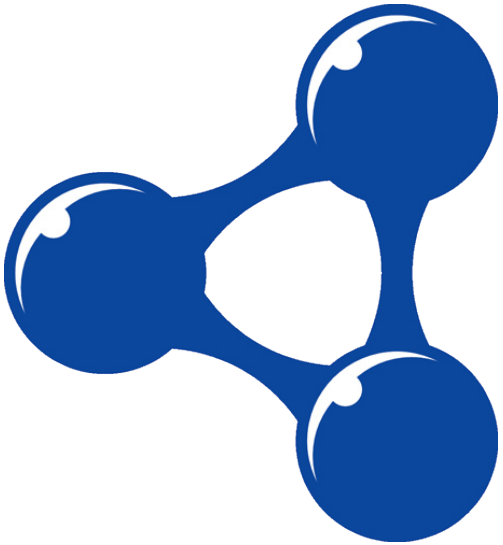
- Lexical Markup Framework (ISO 24613)
  - Standard for representing lexicons
  - XML, UML (primarily)
- LexInfo, LIR
  - Represent lexical information relative to an ontology
  - OWL
- SKOS (W3C Standard)
  - Designed for Taxonomy/Vocabulary representation
  - RDF

Design goals

- RDF(S)
- Conciseness
- Not prescriptive
  - i.e., uses data categories
- Semantics by reference
  - i.e., uses ontologies
- Extensible

Why lemon: RDF(S)

- RDF models are labelled directed graphs
  - Better representation
- Each entry has a URI
  - Queriable on the web using standards
  - Clear ownership of data
- Linking possible between different lexica
  - Reuse of lexicon data
- Some induction possible (subproperties, classes etc.)

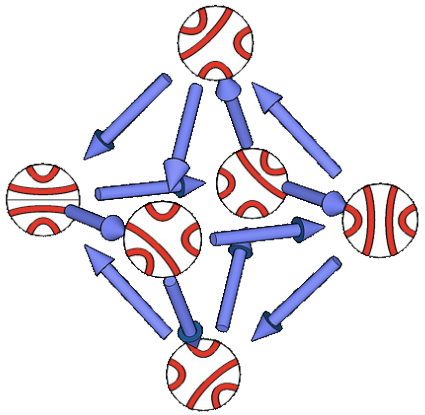


Why lemon: Conciseness

- Small models (i.e., fewer links, fewer kB)
- Easier to understand
- “Open-world”: Not necessary to state all facts
  - Multiple points of view

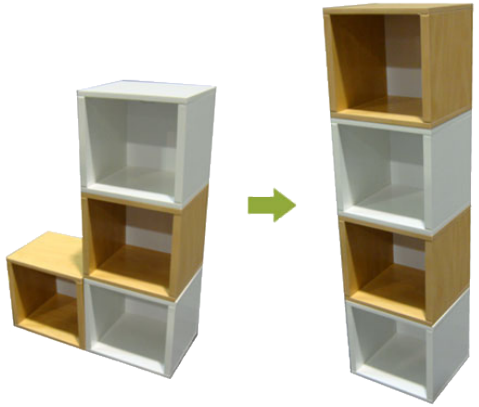
Why lemon: Semantics by Reference

- The web of data is full of ontologies in OWL, RDFS, RIF...
- Meaning of a word given by reference
- Reference (generally an ontology) capable of representing more complex semantic information
- Disambiguation is performed relative to the ontology
- No (traditional) word senses
  - No clashing of word senses in cross-lingual mappings



Why lemon: Modular and extensible

- RDF(S) extensibility allows representation of
  - Subtle differences
  - Unexpected data categories
- Modularity
  - Different modules for different user requirements
  - New modules can be added later without affecting core



Outline

Methodology

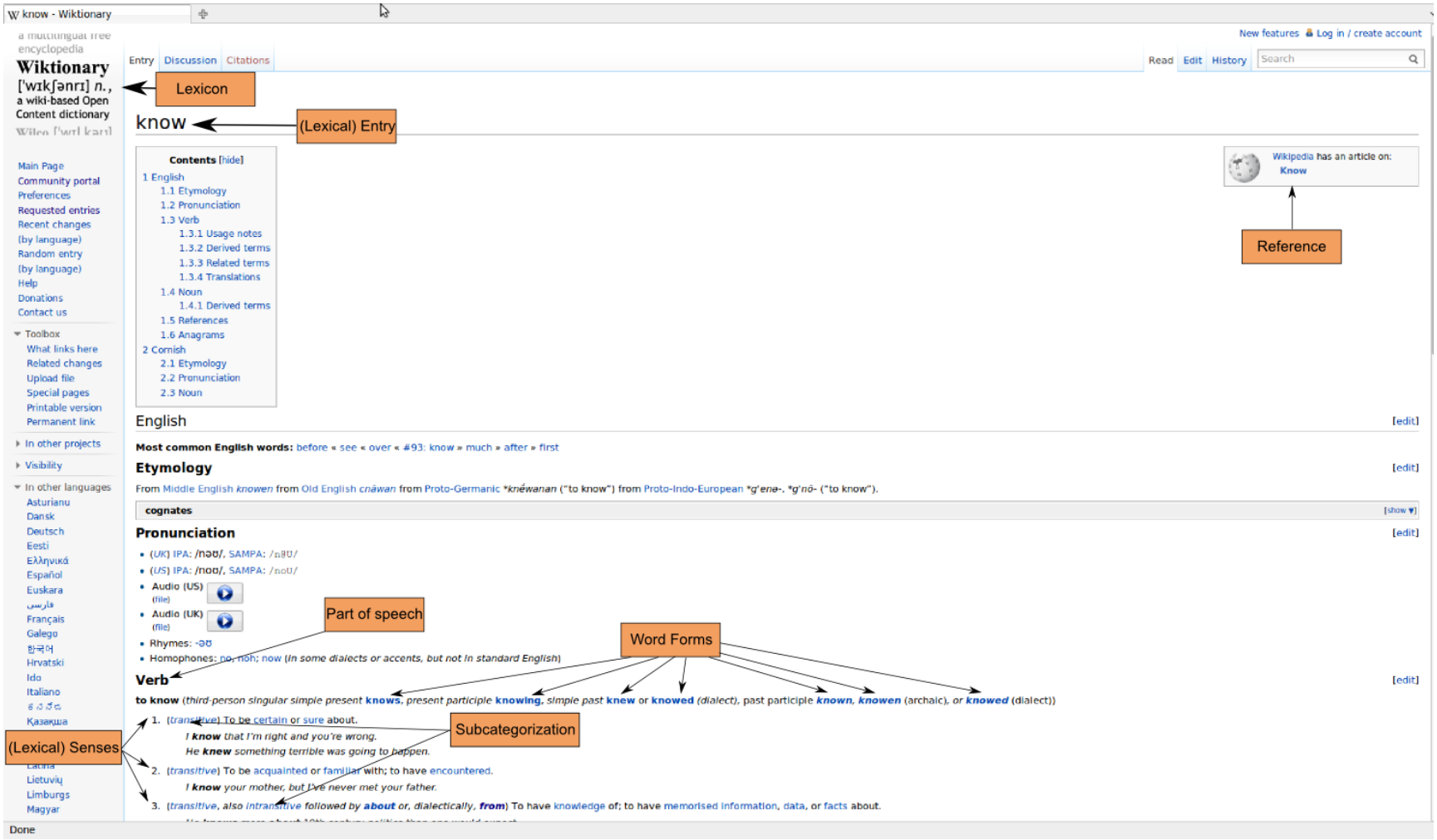
- Start with RDF-WordNet 2.0
- Mapped synsets to references
  - Hence synsets are treated as ontology classes
- Sense and Word correspond to **lemon**
- Canonical form introduced as new node, other forms extracted from WordNet files (not in RDF!)
- Part-of-Speech tags mapped to LexInfo

Example

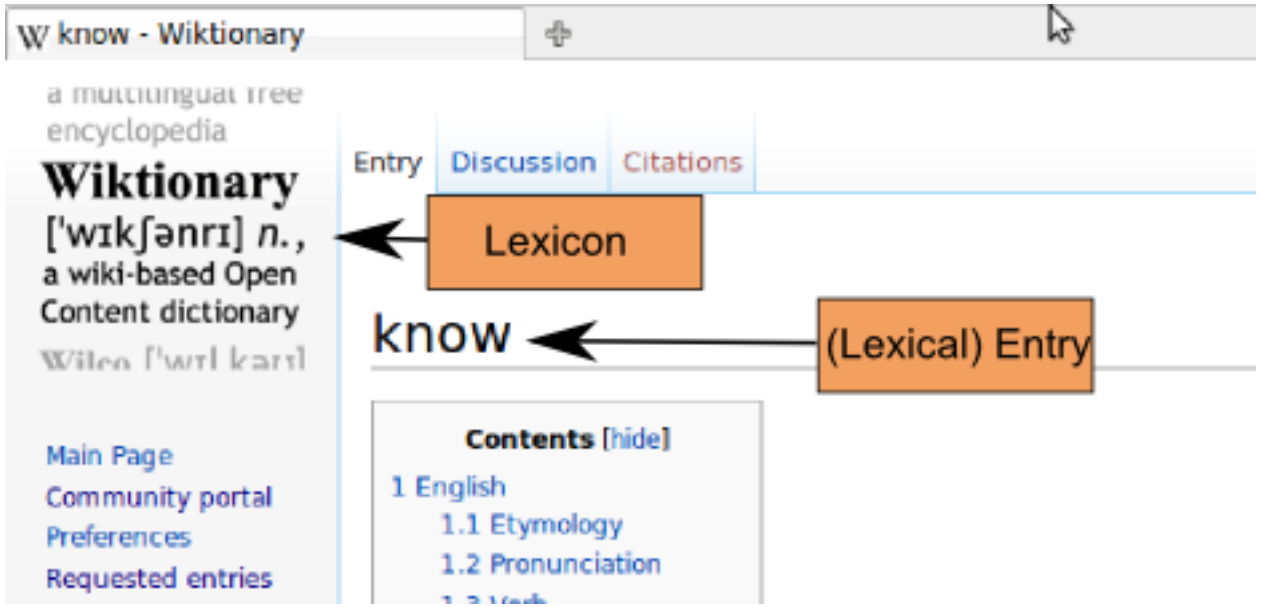
```
lwn:marmoset-noun-entry rdf:type lemon:LexicalEntry ;
  lexinfo:partOfSpeech lexinfo:noun ;
  lemon:sense lwn:sense-marmoset-noun-1 ;
 lemon:canonicalForm lwn:word-marmoset-canonicalForm .

lwn:sense-marmoset-noun-1
  lemon:reference wn20:synset-marmoset-noun-1 .

lwn:word-marmoset-canonicalForm
  lemon:writtenRep "Marmoset"@en .
```

Outline

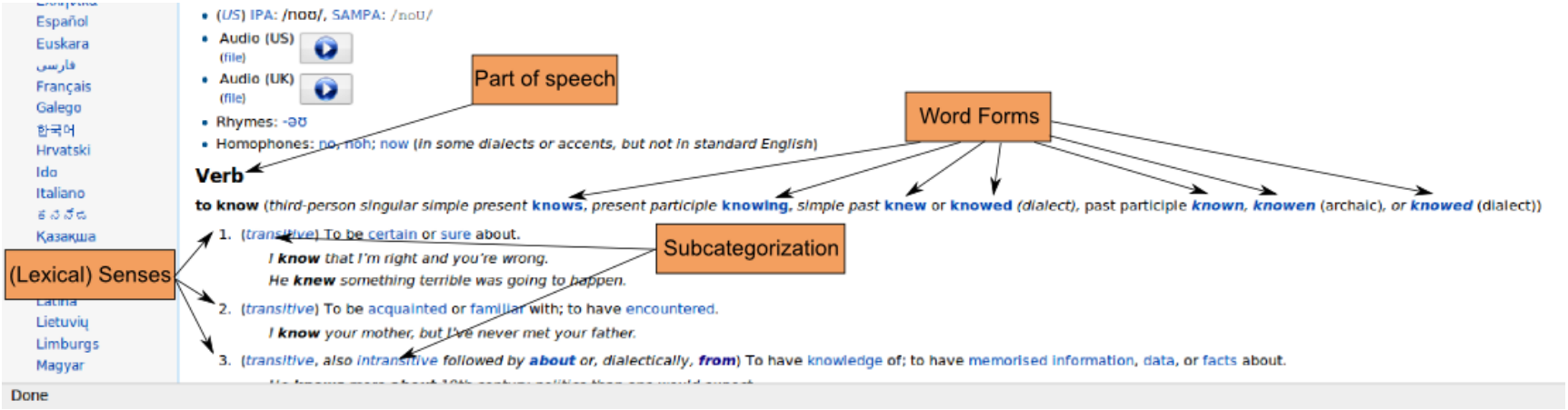
Mapping strategy



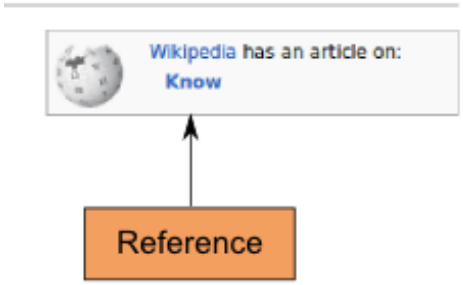
Mapping strategy



Mapping strategy



Mapping strategy



Example

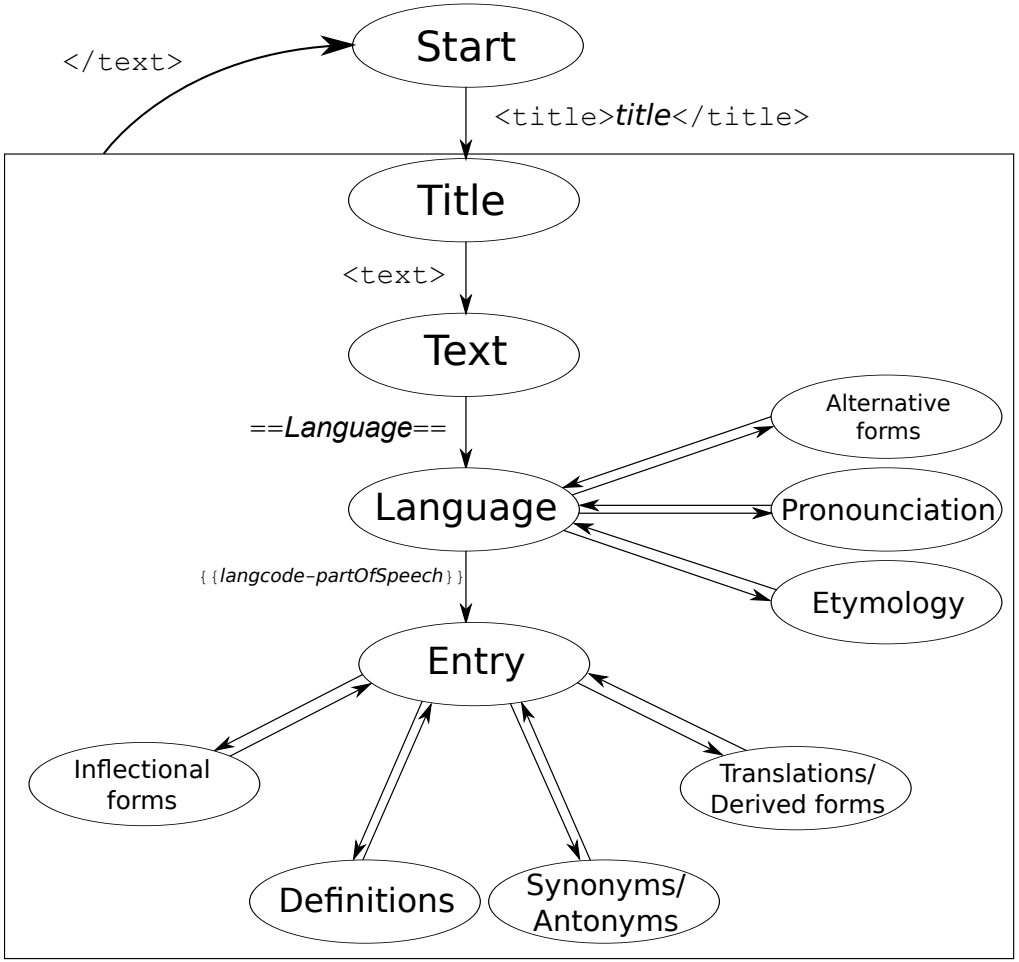
Wiktionary:

```
<page>
<title>free</title>
<text>
==English==
===Adjective===
{{en-adj}}

# Not [[imprisoned]] or [[enslaved]].
# Obtainable without any [[payment]].

====Synonyms====
* {{sense|obtainable without payment}}:
    [[free of charge]], [[gratis]]

====Translations====
{{trans-top|not imprisoned}}
* German: {{t+|de|frei}}
{{trans-bot}}
</text>
</page>
```

lemon:

```
:free_en_adj lemon:canonicalForm [
  lemon:writtenRep "free"@en ] ;
  lexinfo:partOfSpeech lexinfo:adjective ;
  lemon:sense :free_en_adj_sense0 ;
  lemon:sense :free_en_adj_sense1 ;
  lemon:sense :free_en_sense_def .

:free_en_adj_sense0 lemon:definition [
  lemon:value "Not imprisoned or enslaved"@en ] ;
  lemon:reference
    <http://en.wiktionary.org/wiki/free> ;
  lexinfo:translation :frei_de_sense_def .

:free_en_adj_sense1 lemon:definition [
 lemon:value "Obtainable without any payment"@en ] ;
  lemon:reference
    <http://en.wiktionary.org/wiki/free> ;
  lexinfo:synonym :free_of_charge_en_sense_def .
```

Mapping algorithm



Sense mapping

- (English) Wiktionary uses different glosses to link pages
  - "Not imprisoned or enslaved" vs. "Not imprisoned"
  - "Obtainable without any payment" vs. "Obtainable without payment"
- We merge information on the same Wiktionary page

  IF The secondary gloss is a substring of the primary gloss

  OR The Levenshtein distance between the glosses exceeds some λ

  AND The Levenshtein distance is maximal among candidates

Sense mapping results

λ	Merged	Coverage	Precision	Harmonic Mean
Substring	36595	37.8%	99.5%	54.8%
0.9	6842	44.9%	100%	62.0%
0.8	3398	48.4%	99%	65.0%
0.7	2669	51.2%	99%	67.5%
0.6	3243	54.5%	97%	69.8%
0.5	7128	61.9%	97%	75.6%
0.4	4612	66.6%	98%	79.3%
0.3	6295	73.1%	91%	81.1%
0.2	7983	81.4%	92%	86.4%
0.1	6934	88.5%	73%	80.0%
0.0	3862	92.5%	71%	80.3%

Outline

Linking WordNet and Wiktionary

- We used the following criteria:
  - The canonical (lemma) form is equivalent
  - Part-of-speech is the same
  - Do not assert different values for the same property
  - Do not have a different non-canonical form with the same properties
    - e.g., German: “Banken” versus “Bänke”
- Results:

	#Entries	Percent (WN)	Percent (Wikt)
Linked	63,478	21.0%	26.9%
Not Linked (Wiktionary)	172,674	-	73.1%
Not Linked (WordNet)	238,408	79.0%	-
Ambiguous	1,741	0.6%	0.7%

Sample of failed links

(in Wiktionary not in WordNet)

- 28: In WordNet
- 9 (“polysemic”, “abaciscus” (pictured)): Omissions
- 10 (“false friend”, “apples and pears”): Idioms not covered by WordNet
- 2 (“raven” (adj), “to minute” (verb)): Not with same part-of-speech
- 1 (“wares”): Other

Outline

Conclusion

- Conversion of WordNet easy due to model interoperability (... even stage 1 helps!)
- Wiktionary much harder
- **lemon** is an adequate model for representing Wiktionary and WordNet
- Wiktionary's data model is flawed!
- Overlap between WordNet and Wiktionary quite low (~25%)
- Linking these resources can create a “virtual” resource with much higher coverage

Learn more

- `http://monnetproject.deri.ie/lemonsource`: Data sets from the presentation
- `http://www.lexinfo.net/lemon-cookbook.pdf`: The **lemon** cookbook (technical manual)
- `http://www.w3.org/community/ontolex`: OntoLex Community group
- `http://www.monnet-project.eu/lemon`: **lemon** Ontology