Apertium RDF

Linking bilingual dictionaries on the Web of Data

Jorge Gracia

Ontology Engineering Group (OEG)
Artificial Intelligence Department
Universidad Politécnica de Madrid (UPM)

jgracia@fi.upm.es

Talk at OEG
09/04/2015

Outline

- The context
- Motivation
- The Apertium platform
- Representing translations in RDF
- Building the Apertium RDF graph
- Traversing the graph
- Linking with external sources
- Conclusions

The context

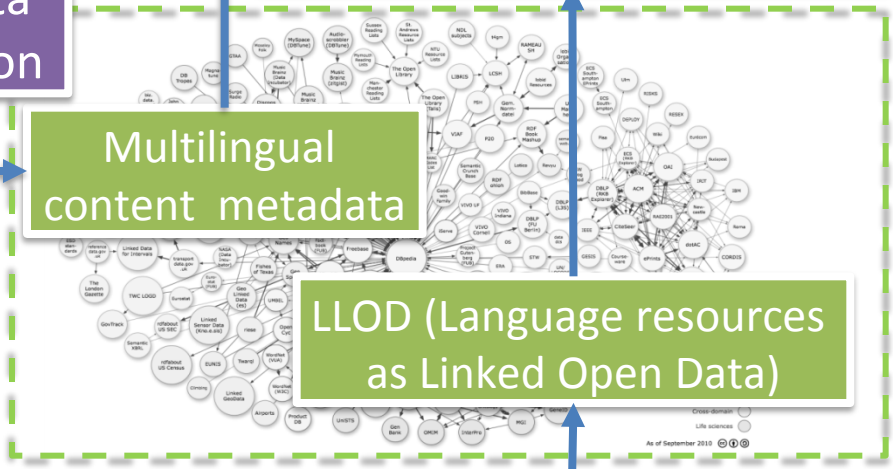
Linked Data as an enabler of cross-media and multilingual content analytics for enterprises across Europe

Content providers

Metadata generation

Content Analytics

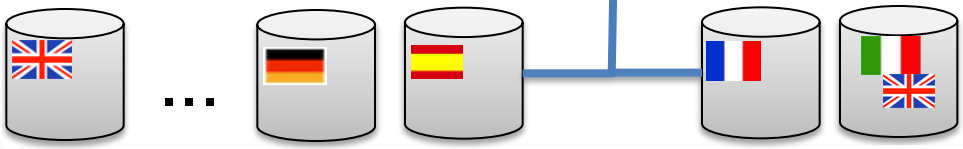
LOD-aware NLP services



Multilingual content metadata

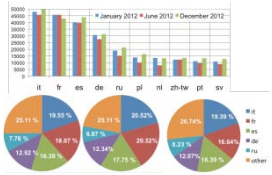
LLOD (Language resources as Linked Open Data)

LLOD generation



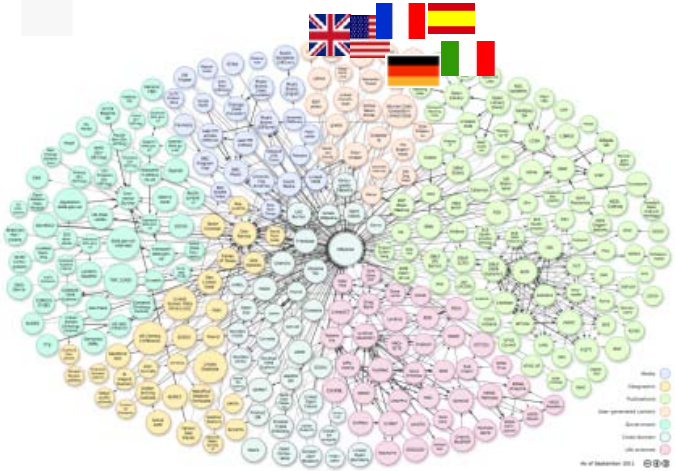
Language resources (lexicon, corpora,...)

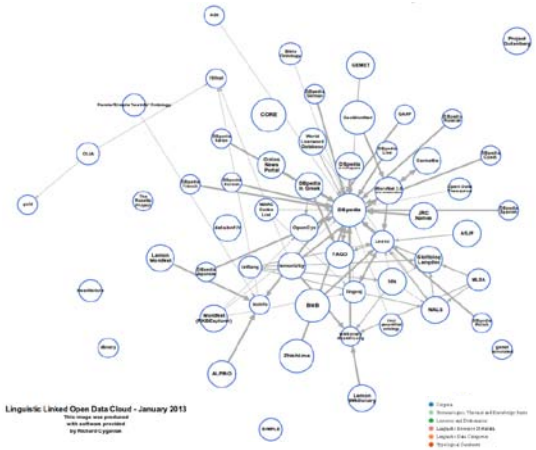
Content Analytics consumers



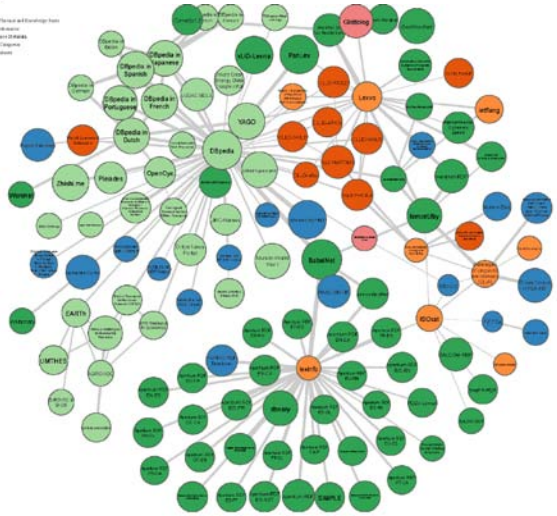
Linguistic LOD (LLOD) cloud

LOD cloud





January 2013



March 2015

Motivation

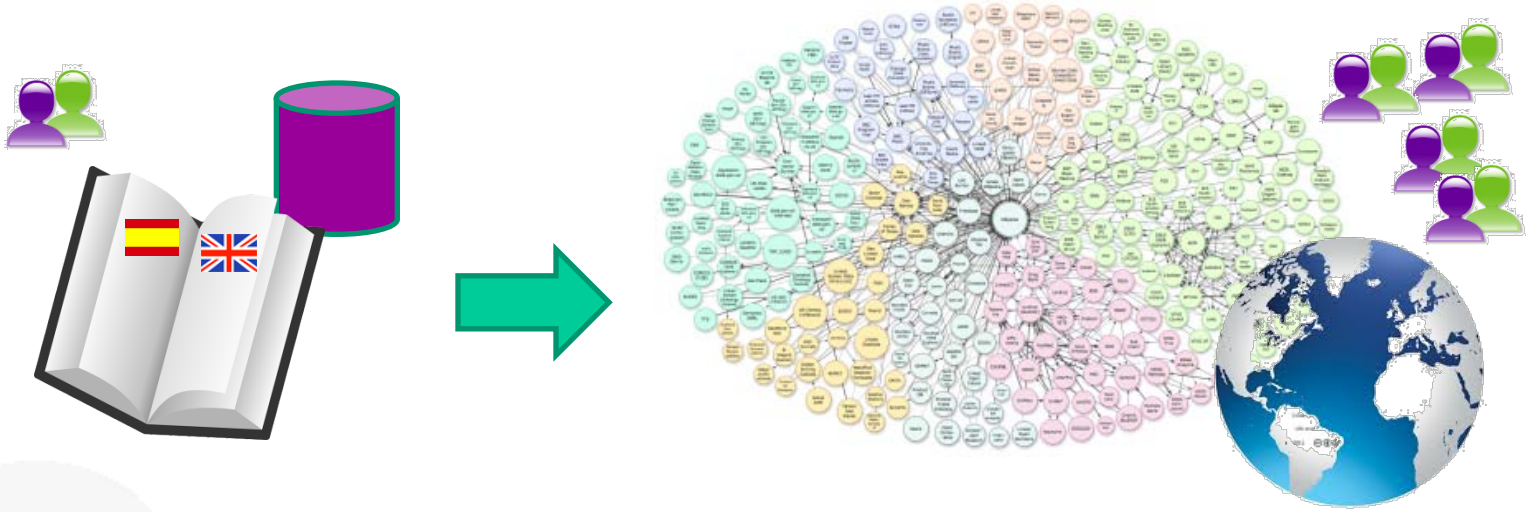
Motivation

Current multilingual lexica and electronic dictionaries

- Proprietary formats
- Non-standard APIs
- Disconnected from other resources

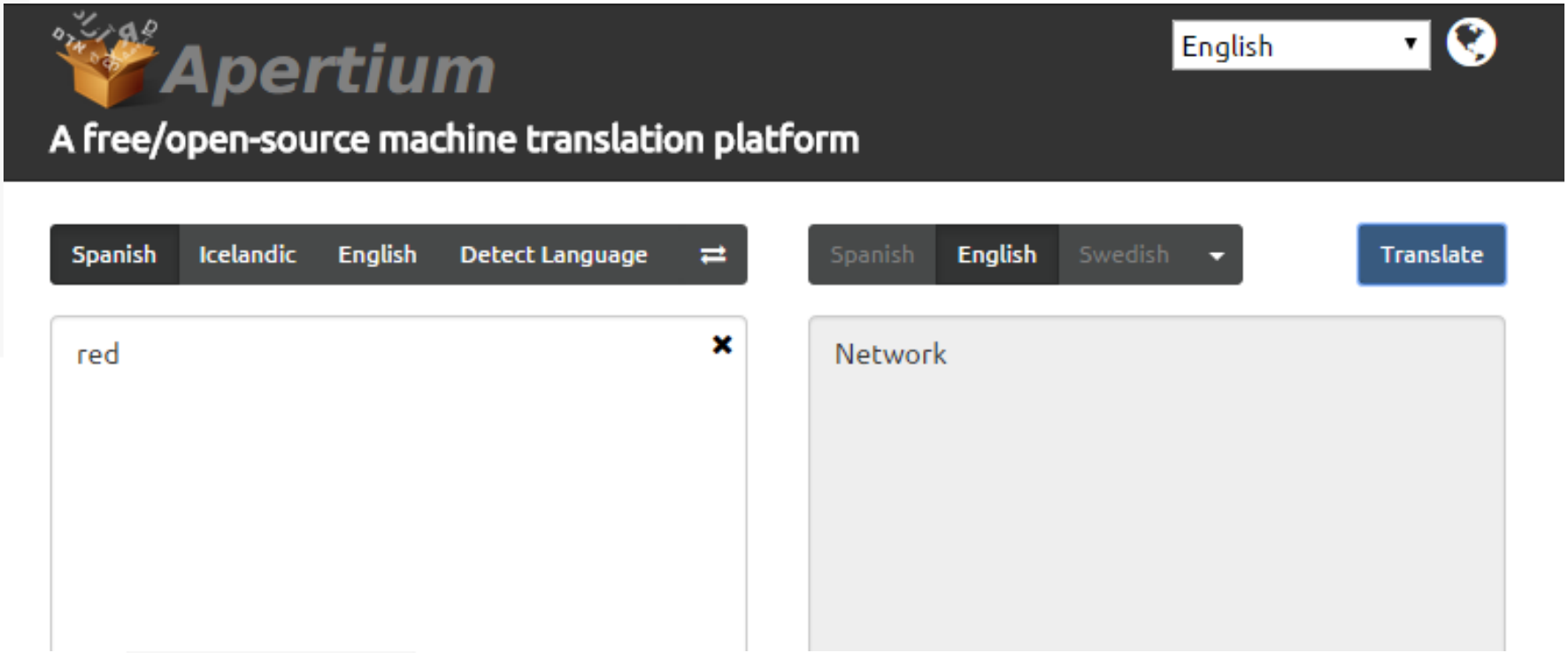
Motivation

GOAL: to expose translations contained in bilingual dictionaries as Linked Data on the Web for their consumption by semantic enabled applications in a direct manner, not relying on application-specific formats



The Apertium platform

Apertium [http://www.apertium.org] open source platform for Machine Translation. Bilingual dictionaries available in XML. We use the LMF version o such dictionaries.



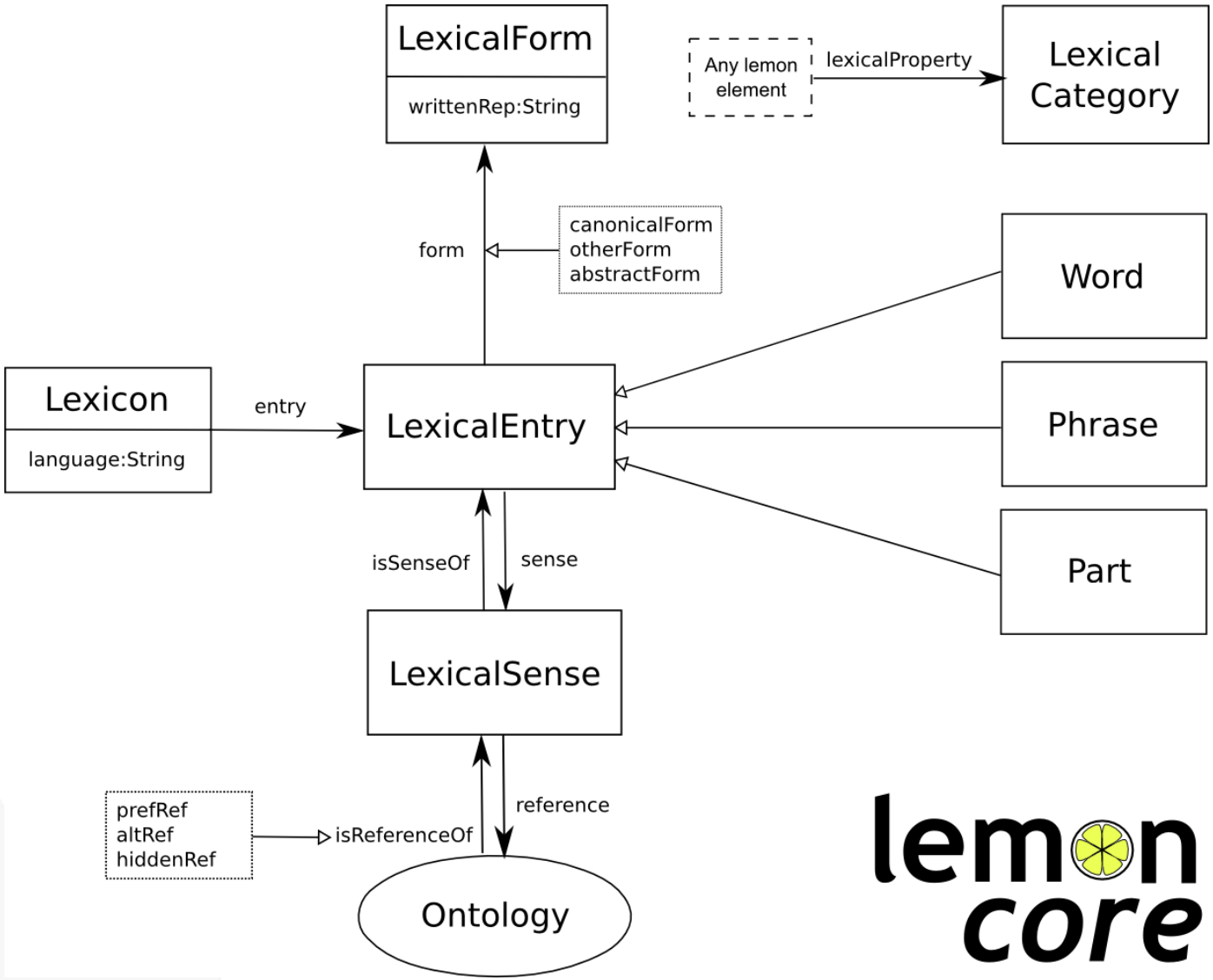
More that 40 language pairs

Afrikaans <-> Dutch
Breton --> French
Catalan <-> Italian
Welsh <-> English
Danish <-- Norwegian
English <-> Catalan
English <-> Spanish
English <-> Galician
Esperanto <-- Catalan
Esperanto <-> English
Esperanto <-- Spanish
Esperanto <-- French
Spanish <-> Aragonese
Spanish <-> Asturian
Spanish <-> Catalan
Spanish <-> Galician
Spanish <-> Italian
Spanish <-> Portuguese
Spanish <-> Romanian
Basque --> English
Basque --> Spanish
French <-> Catalan
French <-> Spanish
Serbo-Croatian <-> English
Serbo-Croatian <-> Macedonian
Serbo-Croatian <-> Slovenian
Indonesian <-> Malaysian
Icelandic <-> Swedish
Icelandic --> English
Kazakh <-> Tatar
Macedonian <-> Bulgarian
Macedonian --> English
Norwegian Nynorsk <-> Norwegian Bokmål
Occitan <-> Catalan
Occitan <-> Spanish
Portuguese <-> Catalan
Portuguese <-> Galician
Northern Sami --> Norwegian Bokmål
Swedish <-> Danish
……

22 of them (more stable) available in LMF

Representing translations in RDF

lemon

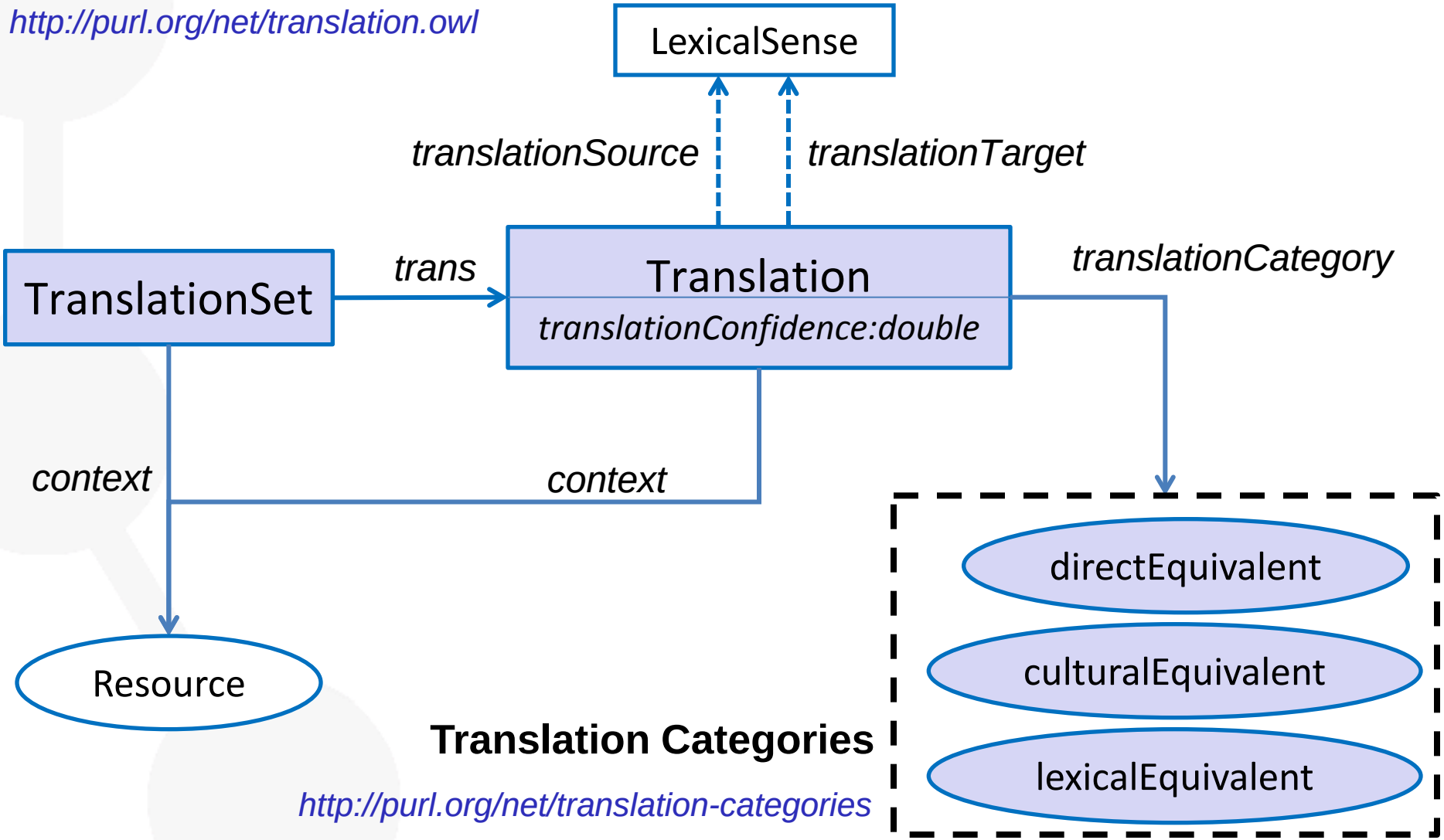


lemon core

The translation module

Translation Module

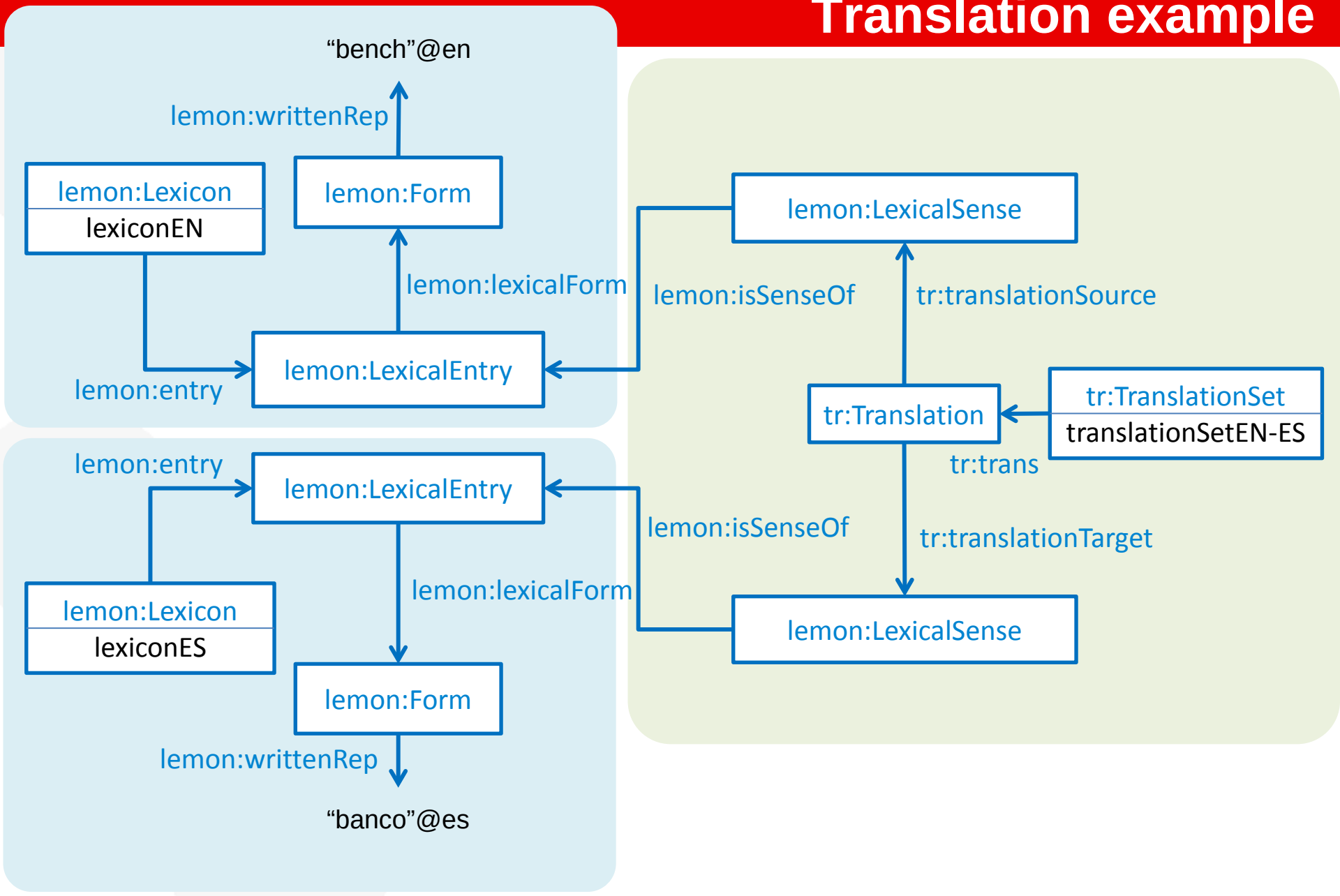
http://purl.org/net/translation.owl



Translation Categories

http://purl.org/net/translation-categories

Translation example

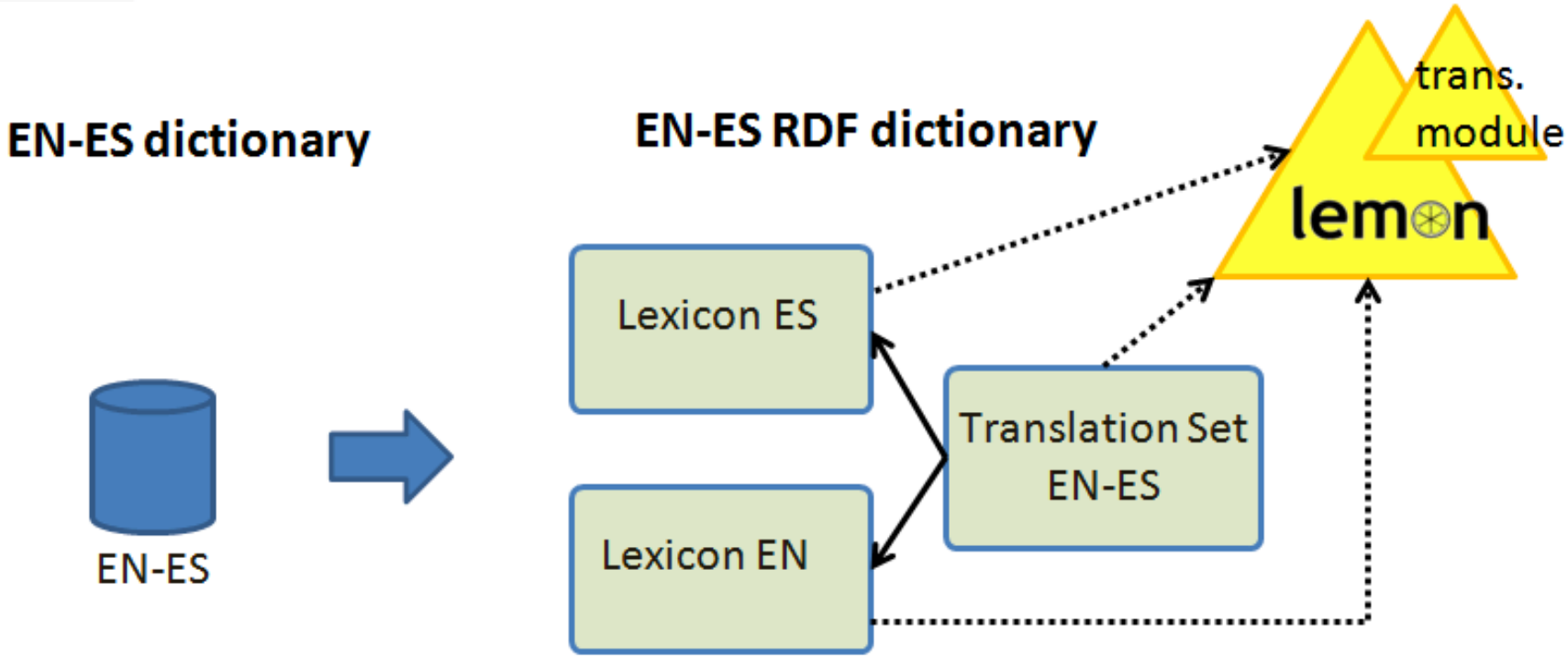


Building the Apertium RDF graph

Methodology

1. Data analysis and vocabulary selection
2. Modelling
3. URIs design
4. RDF generation
5. Publication as linked data

Mapping of data sources



URIs design

Following ISA recommendations [Archer et al.]:

http://{domain}/{type}/{concept}/{reference}

{domain}: **http://linguistic.linkeddata.es/**
{type}: **id** (real-world object)
{concept}: **apertium**
{reference}: **resource ID**

Apertium English lexicon:
http://linguistic.linkeddata.es/**id**/apertium/lexiconEN

Apertium Spanish lexicon:
http://linguistic.linkeddata.es/**id**/apertium/lexiconES

Apertium English-Spanish translation set:
http://linguistic.linkeddata.es/**id**/apertium/tranSetEN-ES

RDF generation based on Open Refine

- E.g., RDF generated:

```
apertium:lexiconEN a lemon:Lexicon ;
        dc:source <http://hdl.handle.net/10230/17110> .
...
apertium:lexiconEN lemon:entry apertium:lexiconEN/bench-n-en .

apertium:lexiconEN/bench-n-en a lemon:LexicalEntry ;
        lemon:lexicalForm apertium:lexiconEN/bench-n-en-form ;
        lexinfo:partOfSpeech lexinfo:noun .

apertium:lexiconEN/bench-n-en-form a lemon:Form ;
        lemon:writtenRep "bench"@en .
```

Publication

- SPARQL endpoint

  http://linguistic.linkeddata.es/apertium/sparql-editor/

- Web interface

  http://linguistic.linkeddata.es/apertium/

- Datahub

  http://datahub.io/dataset?q=apertium+rdf&organization=oeg-upm

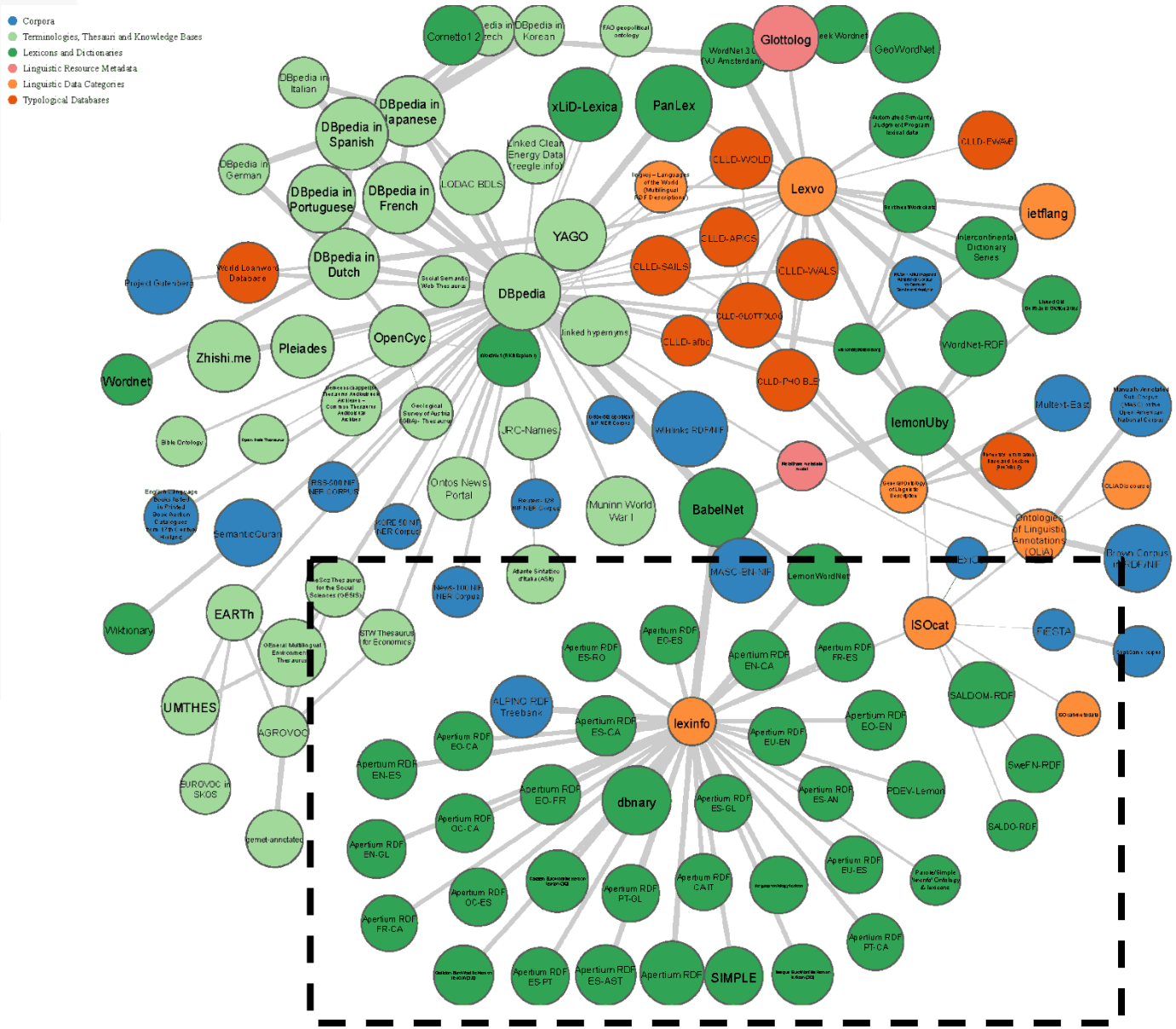
Traversing the graph

22 generated datasets

Lang. pair	# triples	# trans.
CA-IT	180,851	7,869
EN-CA	759,601	33,029
EN-ES	576,316	25,83
EN-GL	425,117	20,034
EO-CA	426,301	19,964
EO-EN	617,772	31,474
EO-ES	380,198	17,212
EO-FR	726,281	35,791
ES-AN	71,997	3,11
ES-AST	825,54	36,096
ES-CA	730,501	31,291

Lang. pair	# triples	# trans.
ES-GL	206,284	8,985
ES-PT	279,245	12,054
ES-RO	400,366	17,318
EU-ES	262,336	11,838
EU-EN	265,466	13,089
FR-CA	152,002	6,55
FR-ES	495,614	21,475
OC-CA	346,346	15,983
OC-ES	317,162	14,561
PT-CA	163,149	7,111
PT-GL	234,065	10,144

Apertium RDF in the LLOD cloud



Direct translations

- Direct translations for “bank”@en

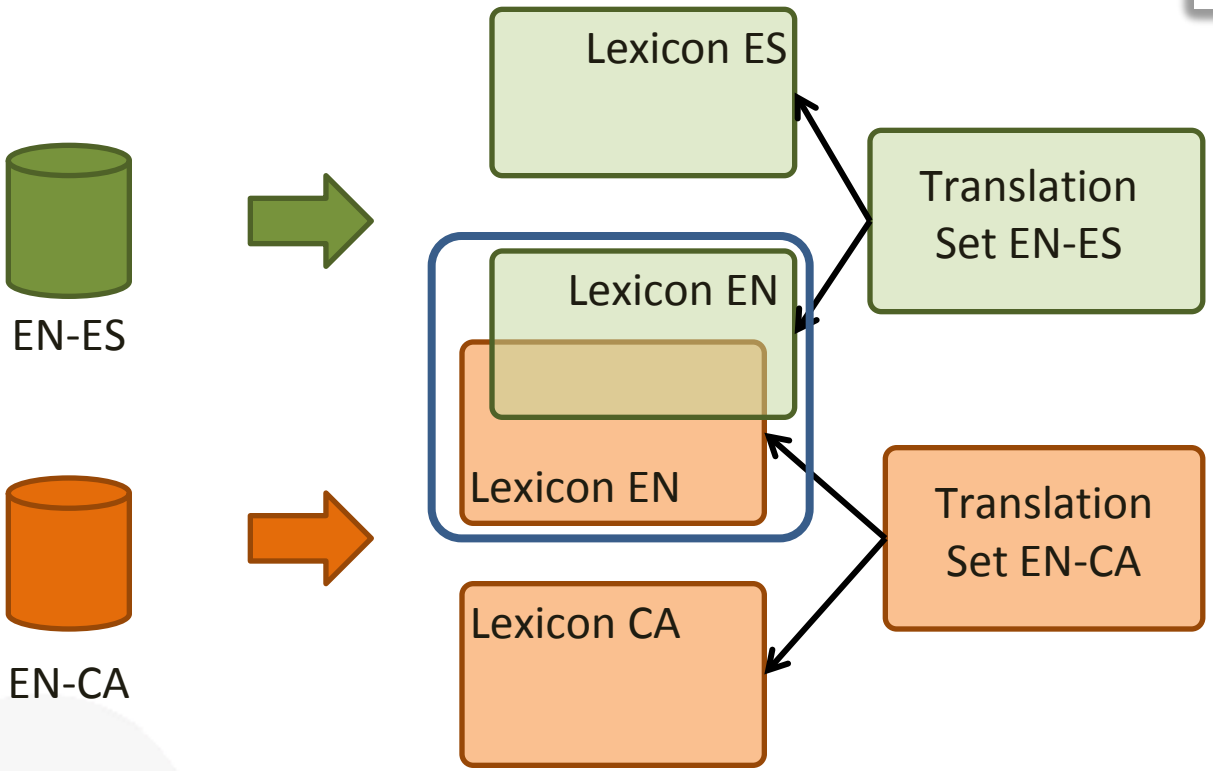
Translated written repr.	Part of Speech
"banc"@ca	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"riba"@ca	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"banco"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"orilla"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"ribera"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"beira"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"banco"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"ourela"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"orela"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"banku"@eu	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"erribera"@eu	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"ertz"@eu	http://www.lexinfo.net/ontology/2.0/lexinfo#noun
"amuntegar"@ca	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"agolpar"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"amontonar"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"apelotonar"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"hacinar"@es	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"apiñar"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"amontoar"@gl	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"kontua_izan"@eu	http://www.lexinfo.net/ontology/2.0/lexinfo#verb
"pilatu"@eu	http://www.lexinfo.net/ontology/2.0/lexinfo#verb

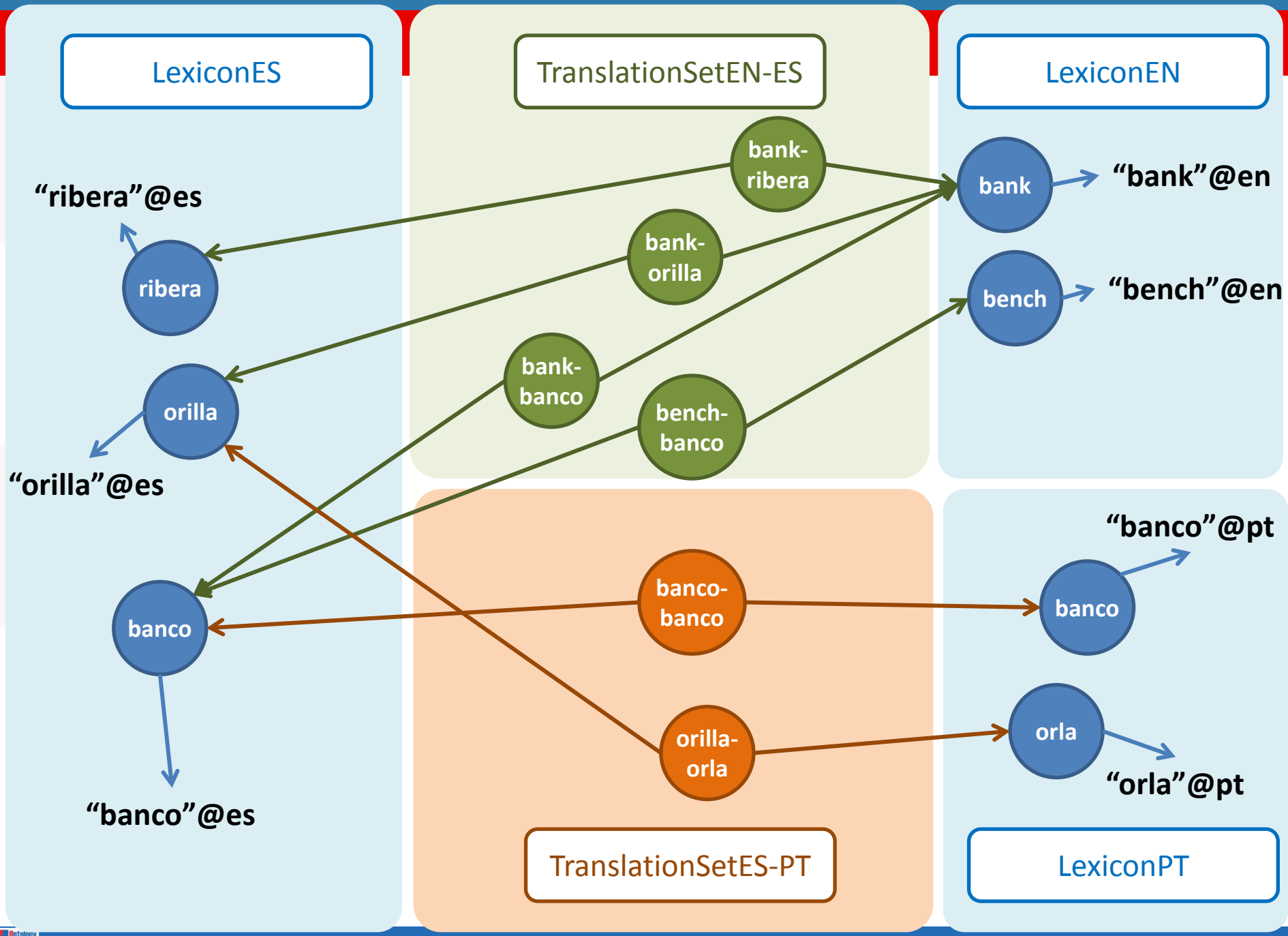
Apertium LMF

Apertium RDF

Monolingual lexicons

Translation sets

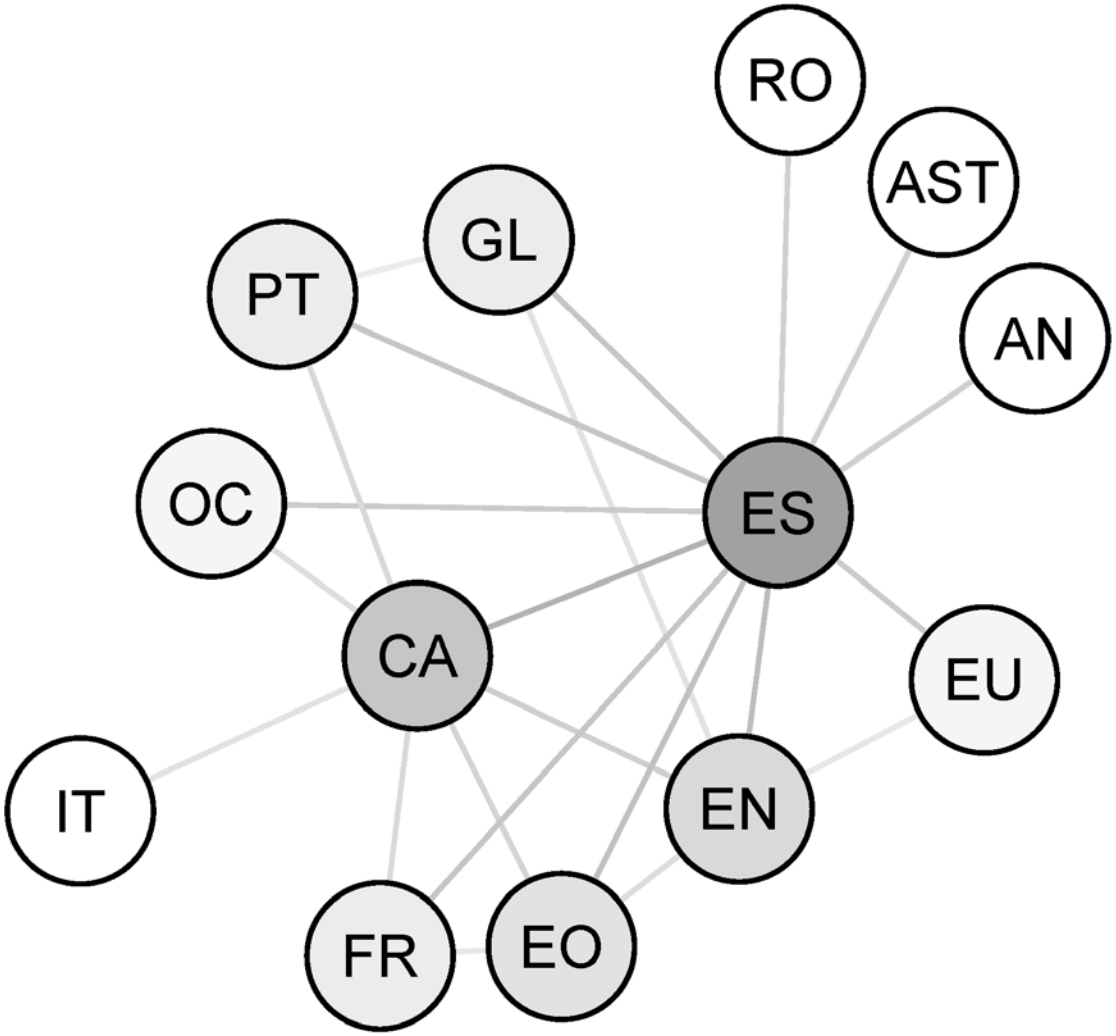



LexiconES
TranslationSetEN-ES
LexiconEN
"ribera"@es
ribera
orilla
"orilla"@es
banco
"banco"@es
bank-ribera
bank-orilla
bank-banco
bench-banco
bank
"bank"@en
bench
"bench"@en
banco-banco
orilla-orla
TranslationSetES-PT
"banco"@pt
banco
orla
"orla"@pt
LexiconPT

- Indirect translations for “bank” EN-> ES -> PT

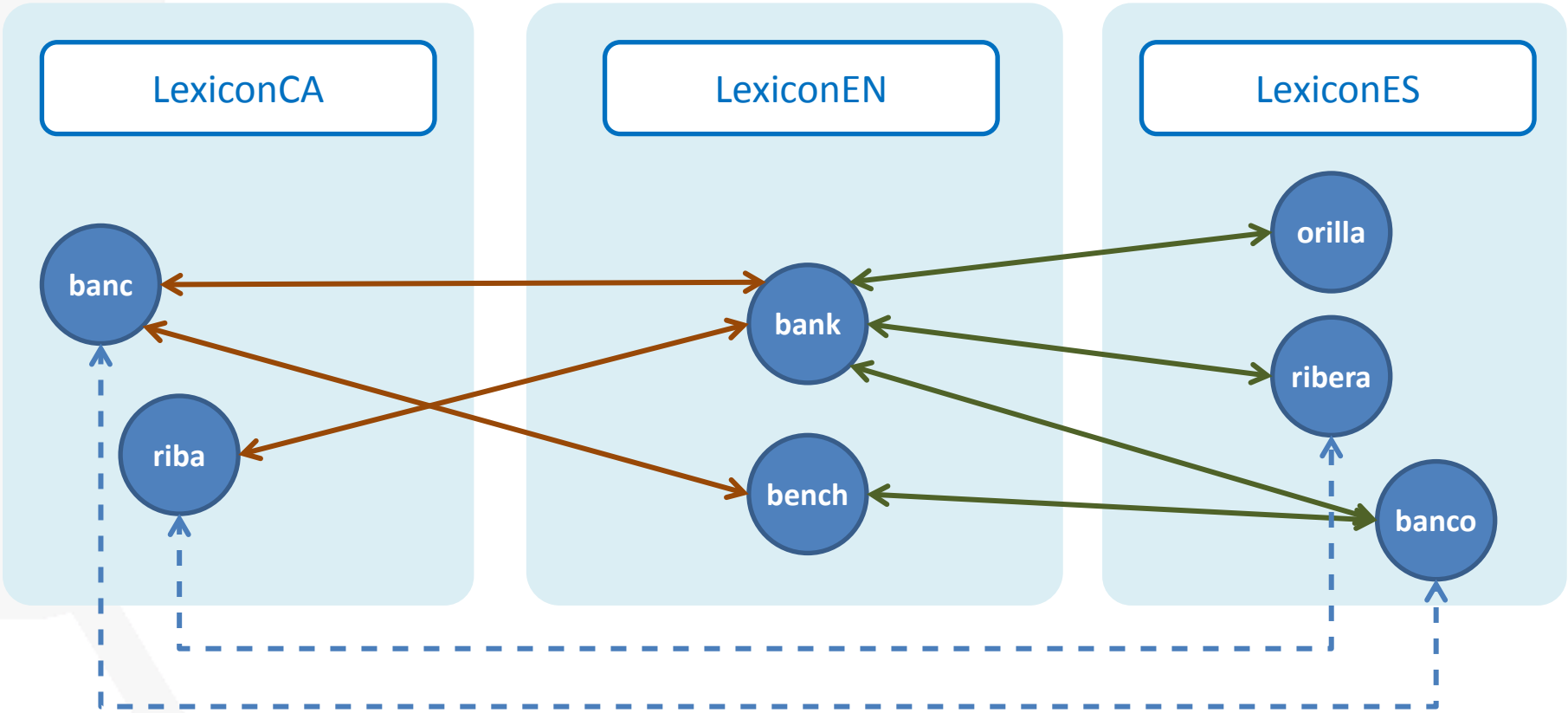
Pivot translation written repres.	Indirect translation written repres.
"banco"@es	"banco"@pt
"orilla"@es	"orla"@pt

Apertium RDF graph



Dijkstra algorithm to choose shortest path

How to measure confidence



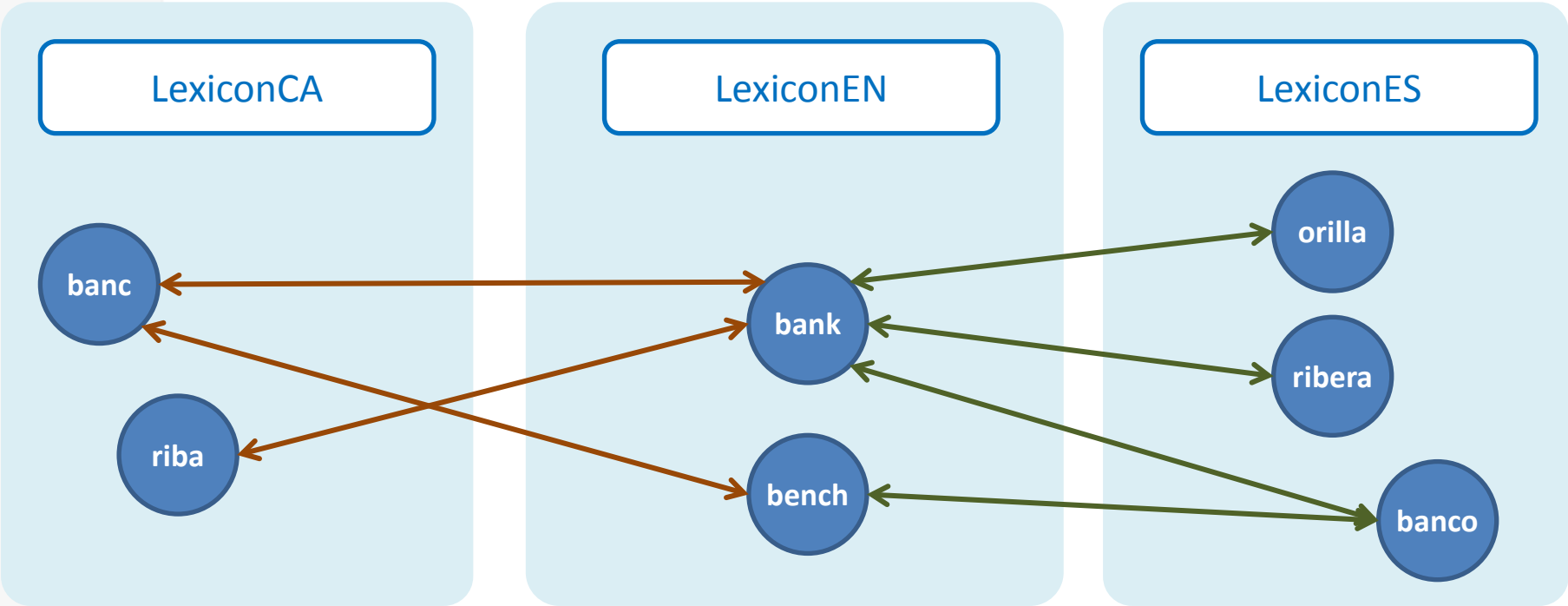
One time inverse consultation (OTIC) algorithm

Given a lexical entry *s:*

1. Get direct translations of s in the pivot language P_s
2. $\forall$ p $\in$ P_s, get its translations in the target language T_p
3. For every t $\in$ T_p,
   (a) gets its set of translations in the pivot language (P_t)
   (b) calculates the score for t:

$$score(t) = 2 * \frac{P_s \cap P_t}{|P_s| + |P_t|}$$

One time inverse consultation



s = “banco”@es

P_{banco}={“bank”@en, “bench”@en}

T_{bank}={“banc”@ca, “riba”@ca}

T_{bench}={“banc”@ca}

P_{banc}={“bank”@en, “bench”@en}

P_{riba}={“bank”@en}

score(“banc”@ca) = 1.0
score(“riba”@ca) = 0.5

Some results of aplying OTIC

Language path	Threshold	Precision	Recall	Effect on recall
EN-CA-ES	0.0	76%	48%	1.0
	0.5	77%	48%	0.99
	1.0	82%	43%	0.89
ES-EN-CA	0.0	53%	39%	1.0
	0.5	55%	39%	1.0
	1.0	61%	36%	0.92
EN-ES-CA	0.0	73%	38%	1.0
	0.5	76%	38%	0.99
	1.0	83%	33%	0.87

Linking with external sources

BabelNet 2.0

A very large multilingual encyclopedic dictionary and semantic network

Around 130.000 links between Apertium RDF – BabelNet already created

But this is another story….

Conclusions

Conclusions

- Apertium data on the Web following SW standards
- Common entry point for all the Apertium dictionaries (and other multilingual resources such as Terminesp)
- Direct and indirect translations can be easily obtained via SPARQL
- Confidence degree for indirect translations
- Linkable with other data sources in the LD cloud
- All the experimental data available at figshare

Thanks for your attention !

http://linguistic.linkeddata.es/apertium/