



“Methodology and tools for Multilingual Linguistic Linked Data generation”

Jorge Gracia, Daniel Vila-Suero

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)
jgracia, dvila@fi.upm.es

EKAW Tutorial on Language Resources and Linked Data
Linköping (Sweden), 25th November 2014

Different methods and guidelines available:

- LOD2
- Datalift
- W3C Linked Data cookbook
- W3C Best Practices for Linked Data
- **Guidelines for Multilingual Linked Data**
- **W3C Best Practices for Multilingual Linked Open Data (BPMLOD) community group [Get Involved!](#)**

- We will use a particular use case: “*bilingual dictionaries*” as **running example** (although the proposed methodology is general)
- W3C BPMLOD community group draft report: “*Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries*”
- Available at:
<http://bpmlod.github.io/report/bilingual-dictionaries/index.html>

Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication

Each activity composed of several tasks

Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication

The goal is to:

- Specify and analyse the data sources in order to plan and manage the subsequent activities
 - Main aspects to specify are:
 - Format
 - Identifiers structure
 - Access methods: *file, webservice, etc.*
 - Data models: *Standards, terminologies, etc.*
 - Language representation: *how languages are tagged, represented, etc.*
 - License and provenance: *existing license of data sources*
-

Documentation of data sources:

- Type of data: *Bilingual dictionary (English and Spanish)*
- Data model: *LMF (Lexical Markup Framework)*
- Format: XML files
- License: GPL 3.0
- Provenance: Apertium EN-ES
-



```
<Lexicon>
  <feat att="language" val="en"/>
  ...
  <LexicalEntry id="bench-n-en">
    <feat att="partOfSpeech" val="n"/>
    <Lemma>
      <feat att="writtenForm" val="bench"/>
    </Lemma>
    <Sense id="bench_banco-n-1"/>
  </LexicalEntry>
  ...
```


- Main activities:
 1. Analysis of data sources
 2. **Modelling**
 3. URI/IRI design
 4. Generation
 5. Publication

Modelling tasks

1. Analysis and selection of domain vocabularies
2. Mapping of data sources and vocabularies
3. Vocabulary for representing licensing and provenance information

Analysis of vocabularies

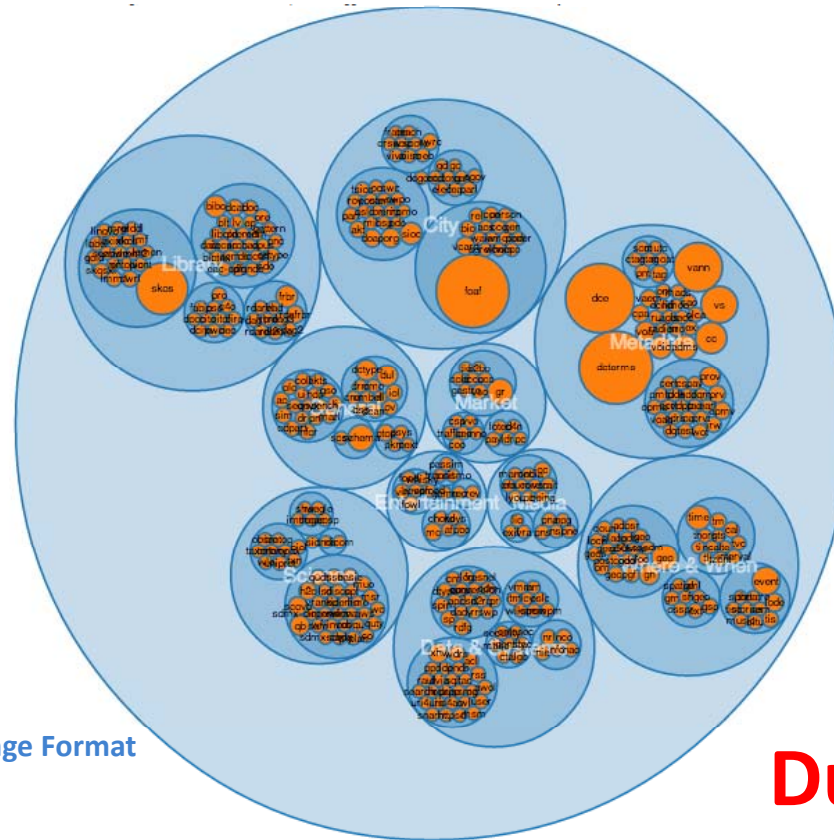
Use <http://lov.okfn.org/>

LexInfo

lemon

NIF

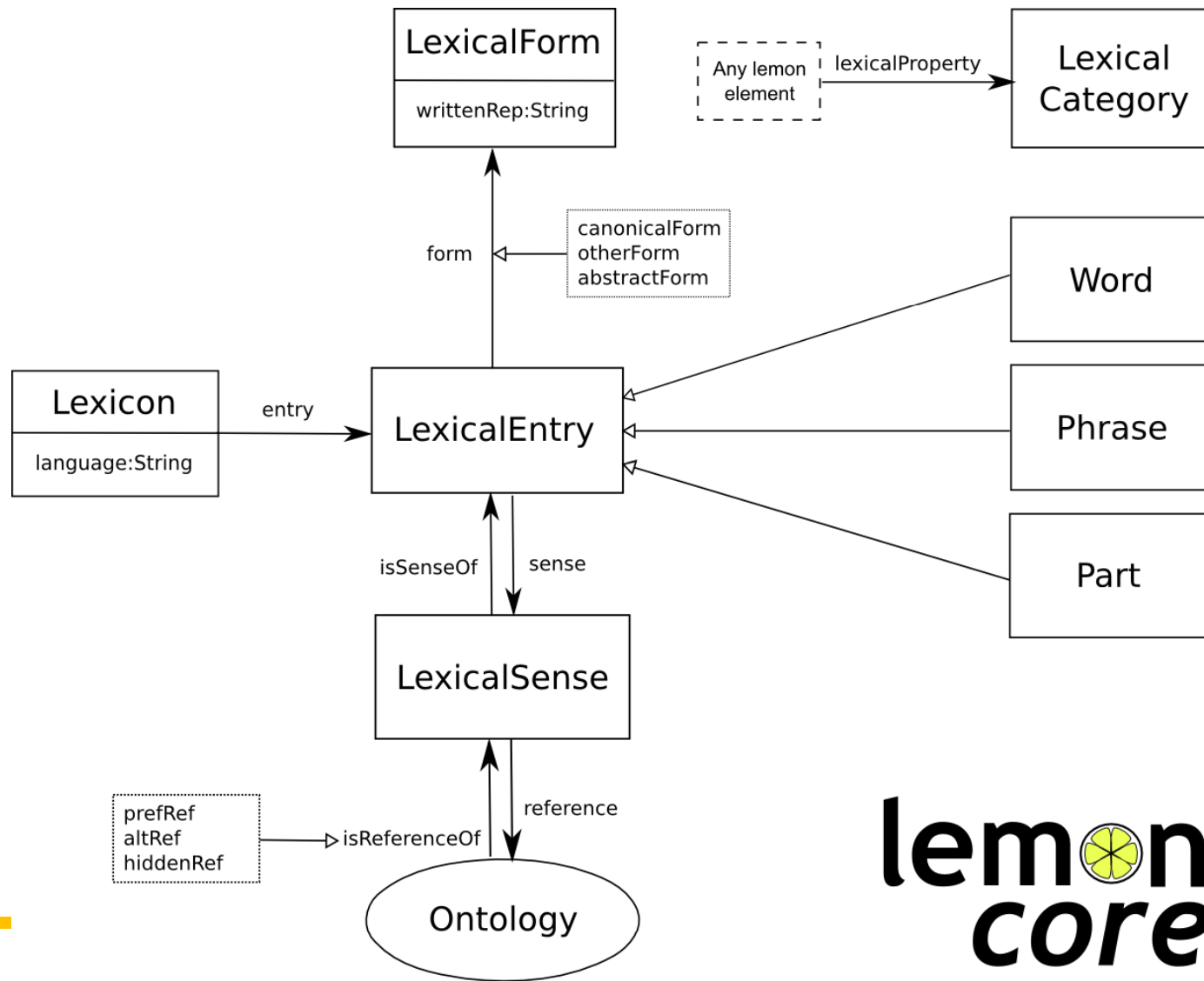
NLP Interchange Format



DCAT

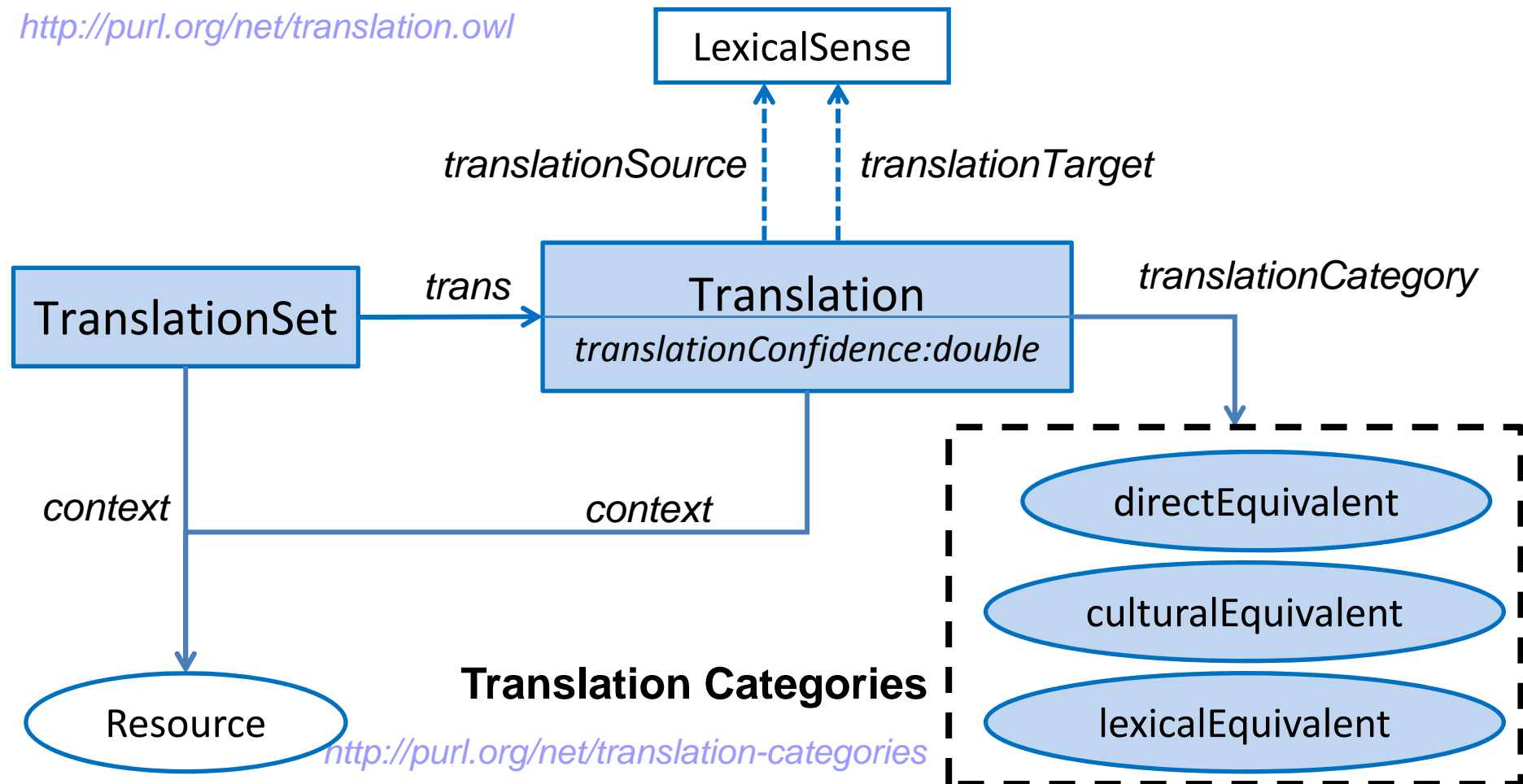
PROV-O

Dublin Core

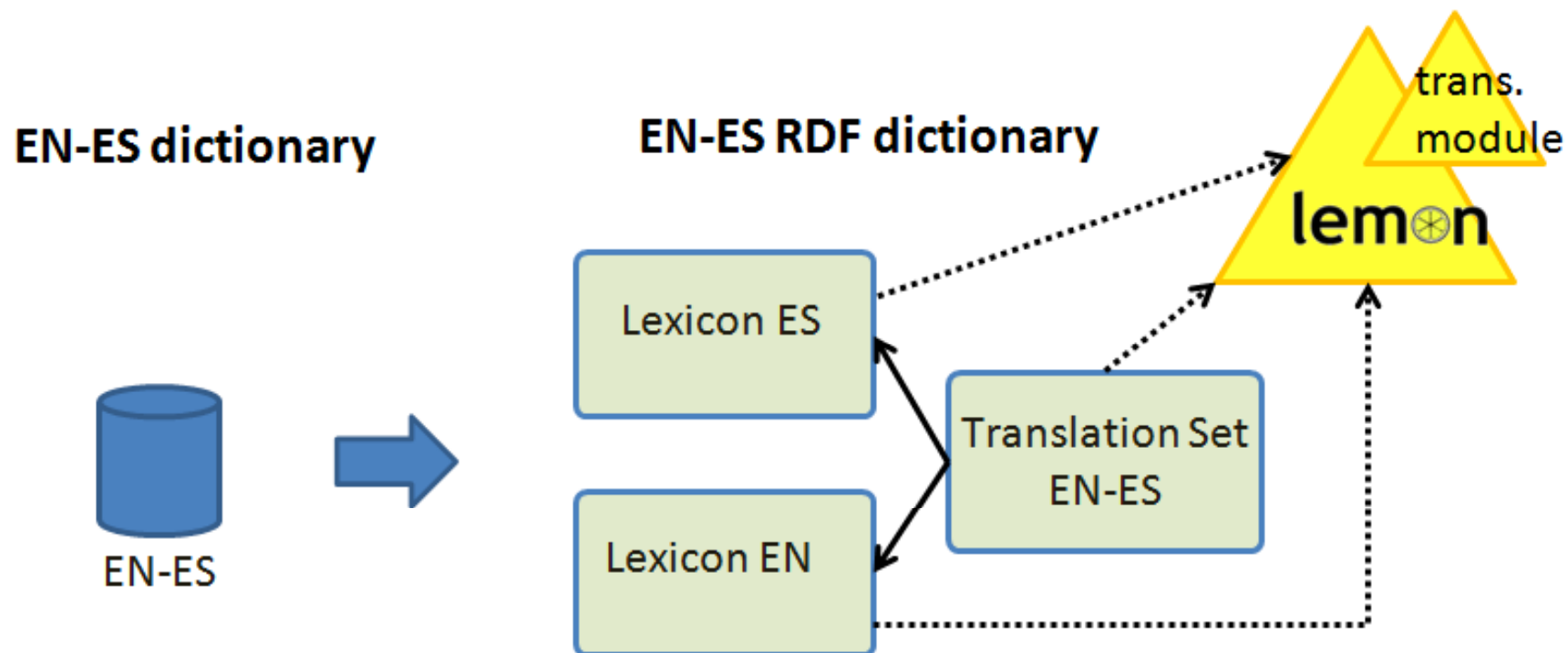


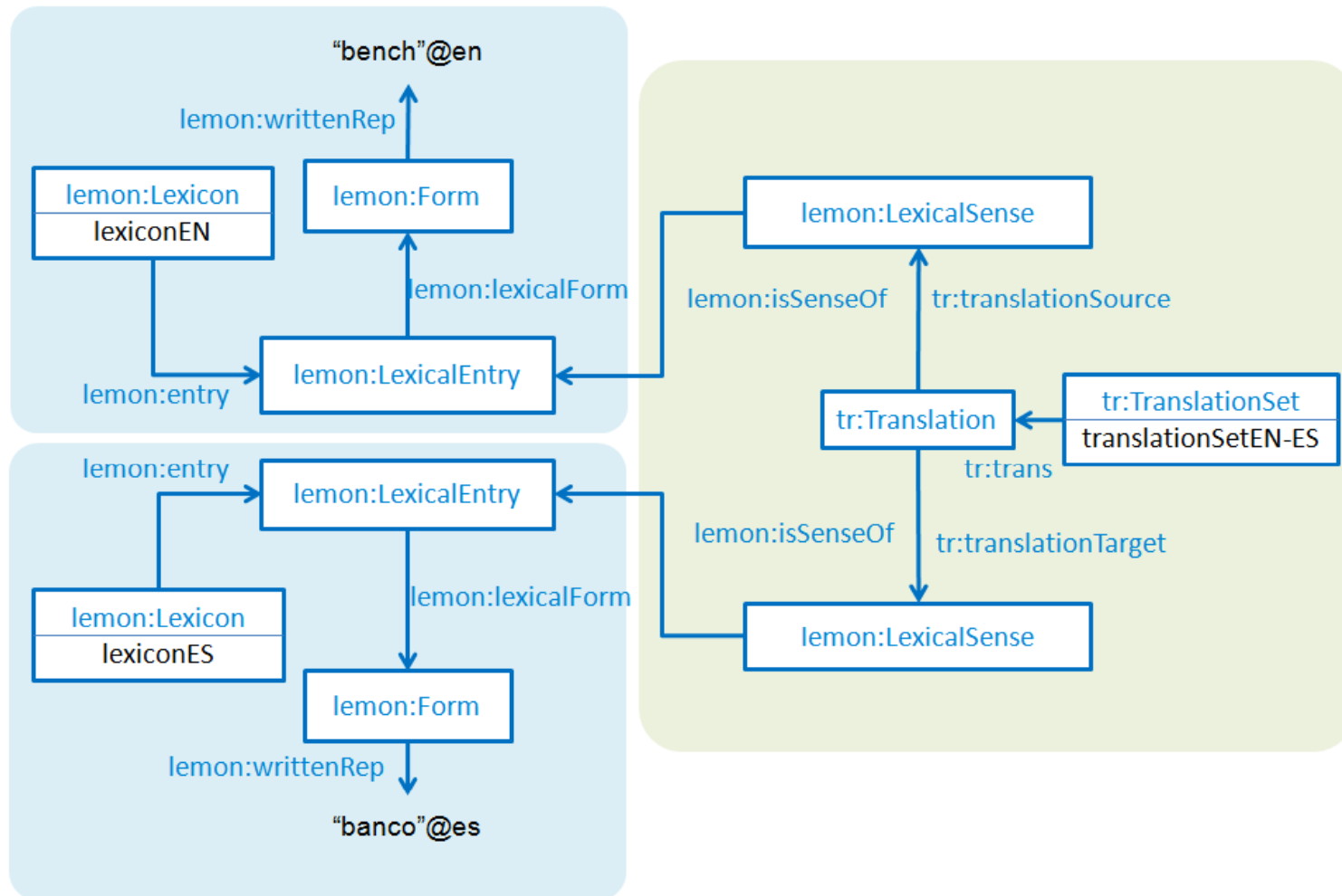
Translation Module

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



Mapping of data sources





Main activities:

1. Analysis of data sources
2. Modelling
- 3. URI/IRI design**
4. Generation
5. Publication

The goal is to:

- Define **URI/IRI patterns and namespaces to be used**
- Ensure that LD best practices are followed

Some good practises...

1. Define **namespace(s)** (that you own or have control over).
2. Define how to create the **ID of resources** (reuse original data source keys if possible)
3. Define the structure of the **URI space** to organize the resources in different addresses and **avoid collision**.

Useful guidance at:

ISA - Study on persistent URIs Archer et al.,

Linked Data patterns book online → URI patterns

Following ISA* recommendations:

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

where:

- `{type}` : a value from the set of type of resources, examples are 'id' or 'item' for real world objects; 'doc' for documents that describe those objects; 'def' for concepts; 'set' for datasets

*ISA - Study on persistent URIs, 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`



- Set of main activities:
 1. Analysis of data sources
 2. Modelling
 3. URI/IRI design
 - 4. Generation**
 5. Publication

1. Selection, extension or development of technologies for RDF generation
2. Mapping of data sources to RDF
3. Transformation of data sources to RDF

Goal:

```

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 .

```

Google refine Apertium-en-es-pol-v2 LexiconES [Permalink](#) Open... Export Help

Facet / Filter **RDF Schema Alignment** Freebase RDF

Using facets and filter methods from each data column

Use facets and filter methods from each data column

Not sure how to get Watch these screens

The RDF schema alignment skeleton below specifies how the RDF data that will get generated from your grid-shaped data. The cells in each record of your data will get placed into nodes within the skeleton. Configure the skeleton by specifying which column to substitute into which node.

Base URI: <http://linguistic.linkeddata.es/id/apertium/> [edit](#)

RDF Skeleton [RDF Preview](#)

Available Prefixes: dc rdfs lexinfo foaf owl xsd rdf lemon [+ add prefix](#) [manage prefixes](#)

LexicalEntry-id URI X lemon:LexicalEntry add rdf:type	X >-lexinfo:partOfSpeech->	LexicalEntry-POS URI add rdf:type	...
	X >-lemon:lexicalForm->	LexicalEntry-id URI X lemon:Form add rdf:type	X >-lemon:writtenRep-> LexicalEntry-writtenFo add property
	X >-dc:source->	http://hdl.handle.net/10230/17110 add rdf:type	...
	add property		
lexiconES X lemon:Lexicon add rdf:type	X >-lemon:entry->	LexicalEntry-id URI add rdf:type	...
	X >-lemon:language->	es	

Add another root node Save

OK Cancel

Main activities:

1. Analysis of data sources
2. Modelling
3. URI/IRI design
4. Generation
5. Publication

- The goal is to:
 - Make available the RDF dataset following [Linked Data best practices](#)
 - Facilitate dataset discovery and consumption

Vocabulary for licensing and provenance

→ INPUT:

- Documentation of data sources (licensing and provenance)

OUTPUT →

- Selection of standard vocabs



1 Add "rights" metadata in the dataset description (e.g., VoID, DCAT)



DCAT W3C
Data catalog vocabulary



2 Use standard predicates to declare "rights" statements (e.g., Dublin Core terms: `dc:rights`, `dct:license`)

Standard license available ?



Yes 

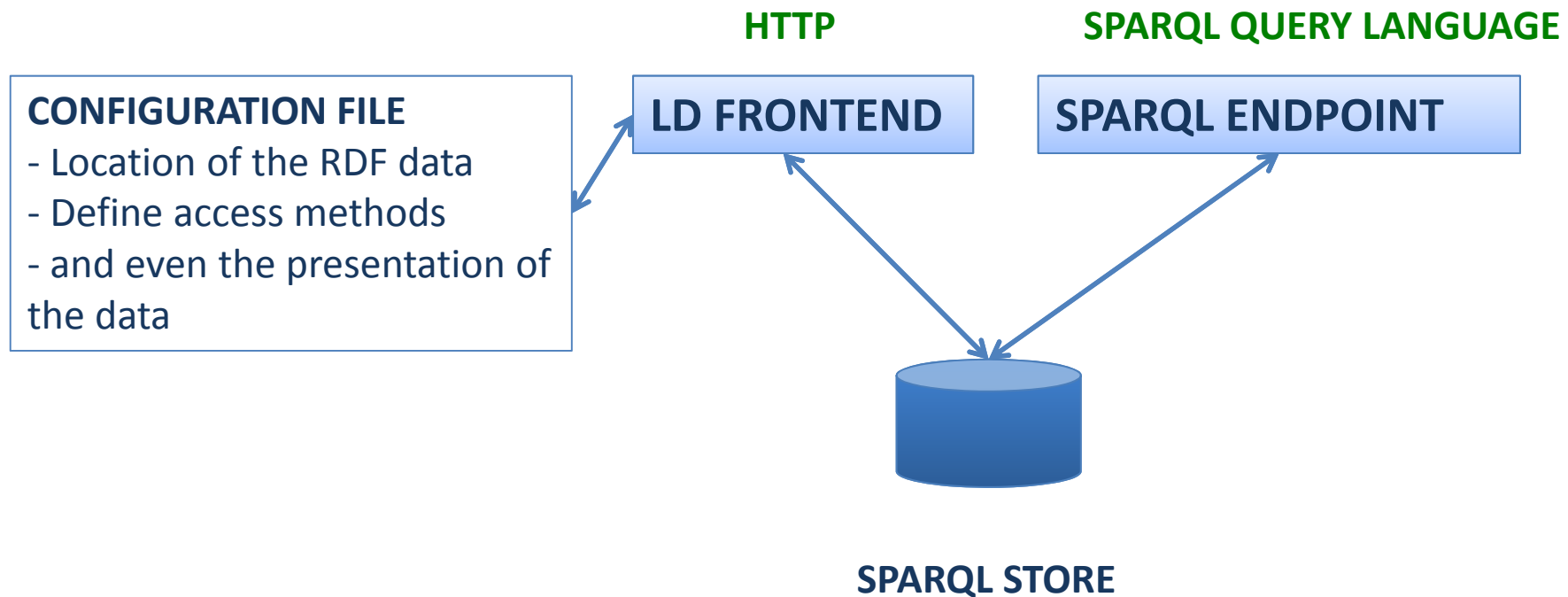
3a Use **URI of standard license** e.g., CC0

No 

ODRL
Open Digital Rights Language

3b Use **rights declaration language**, e.g., ODRL

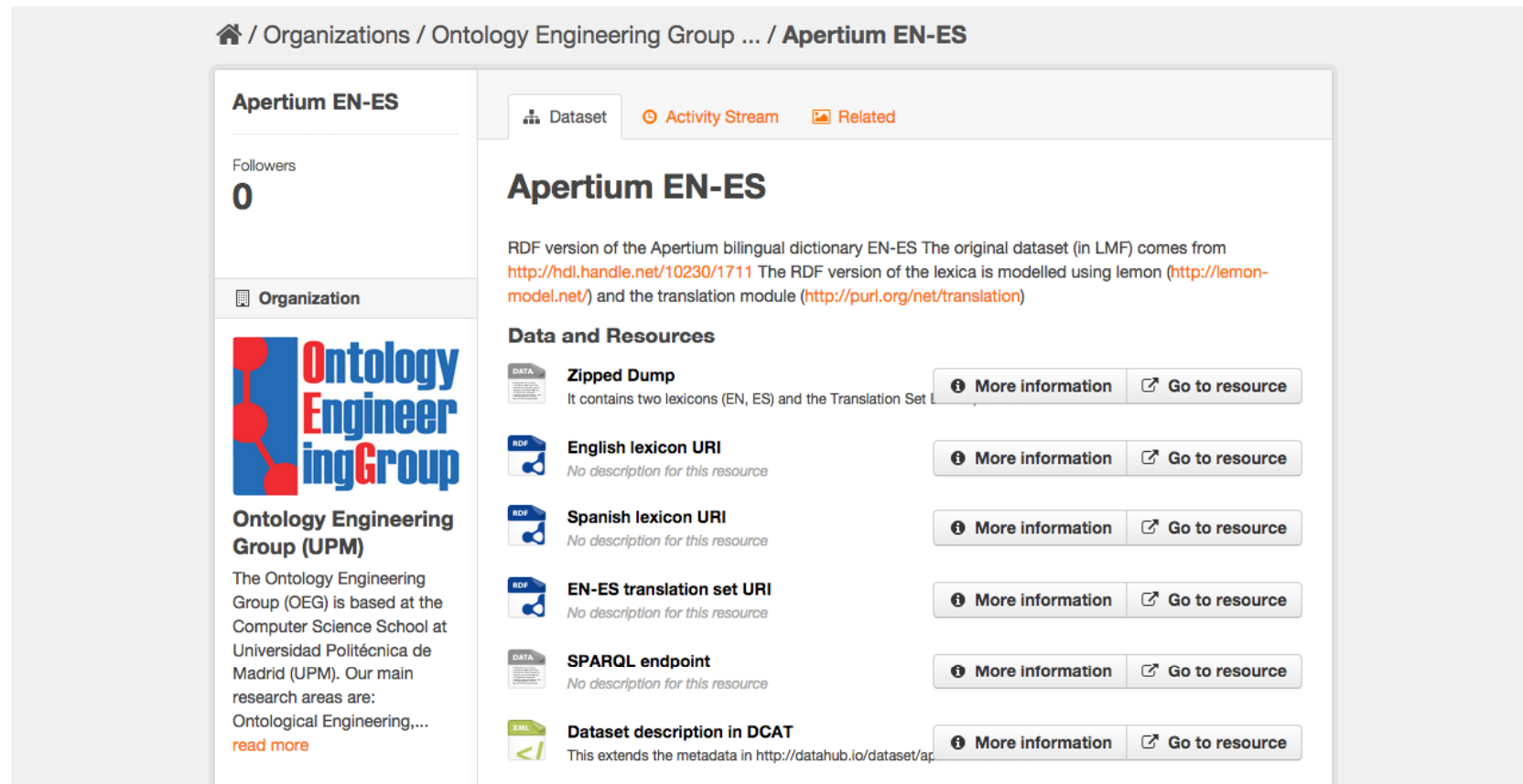
Dataset and vocabulary publication on the Web



Metadata definition and publication using DCAT

- 1) Register dataset in datahub.io
- 2) Extend generated DCAT file and link to datahub.io one

<http://datahub.io/dataset/apertium-rdf-en-es>



The screenshot shows the DataHub.io interface for the dataset 'Apertium EN-ES'. The page is organized into several sections:

- Header:** Shows the breadcrumb path: Home / Organizations / Ontology Engineering Group ... / Apertium EN-ES. Below this are tabs for 'Dataset', 'Activity Stream', and 'Related'.
- Left Sidebar:**
 - Apertium EN-ES:** Shows 0 followers.
 - Organization:** Features the logo for the 'Ontology Engineering Group' and text identifying it as part of the 'Ontology Engineering Group (UPM)'. It mentions the group is based at the Computer Science School at Universidad Politécnica de Madrid (UPM) and lists research areas like Ontological Engineering.
- Main Content Area:**
 - Apertium EN-ES:** A descriptive paragraph stating it is the RDF version of the Apertium bilingual dictionary EN-ES, originally in LMF format, modeled using lemon and a translation module.
 - Data and Resources:** A list of six resources, each with an icon, a title, a brief description, and two buttons: 'More information' and 'Go to resource'.
 - Zipped Dump:** Contains two lexicons (EN, ES) and the Translation Set.
 - English lexicon URI:** No description for this resource.
 - Spanish lexicon URI:** No description for this resource.
 - EN-ES translation set URI:** No description for this resource.
 - SPARQL endpoint:** No description for this resource.
 - Dataset description in DCAT:** Extends the metadata in the dataset's URL.

Extending DCAT description

```
<dcat:Dataset rdf:about="http://linguistic.linkeddata.es/set/apertium/EN-ES">  
  
  <owl:sameAs rdf:resource="http://datahub.io/dataset/apertium-rdf-en-es"></owl:sameAs>  
  
  < dct:source rdf:resource="http://hdl.handle.net/10230/17110"></dct:source>  
  
  < dct:license rdf:resource="http://purl.oclc.org/NET/rdflicense/gpl-3.0"></dct:license>  
  
  < rdfs:seeAlso rdf:resource="http://dbpedia.org/resource/Apertium"></rdfs:seeAlso>  
  
  < rdfs:seeAlso rdf:resource="http://purl.org/ms-lod/UPF-MetadataRecords.ttl#Apertium-en-es_resource-5v2"></rdfs:seeAlso>  
  
</dcat:Dataset>
```


- Loading the RDF data into a **SPARQL** endpoint **not enough for publishing LD:**
 - Why? We provide a queryable repository, but **URIs are not de-referenceable**
- **We need a mechanism to make our URIs de-referenceable:**
 - Through a common web server (as files)
 - **Linked Data front-ends:**
 - Pubby
 - More sophisticated: LD APIs (Puelia, Elda)

- Documentation of data sources and issues
- Language issues have to be taken into account during the whole process
- Metadata description is key for enabling reusing and discovery
- Vocabulary have to be documented and published following LD BPs