# Guidelines for Multilingual Linked Data generation and publication

Jorge Gracia, Daniel Vila-Suero

jgracia, dvila@fi.upm.es

ISWC Tutorial “Building the Multilingual Semantic Web", Trentino (Italy)

20th October 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!

- Report: Guidelines for Linguistic Linked Data Generation: Bilingual Dictionaries
- Available at: http://bpmlod.github.io/report/bilingual-dictionaries/index.html
- We will use this use case (bilingual dictionary) as **running example** but guidelines are general.

# Guidelines for ML Linked Data

- Set of main activities:
  1. Analysis of data sources
  2. Modelling
  3. URI/IRI design
  4. Generation
  5. Publication
- Each activity composed of several tasks

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

# Analysis of data sources: Intro

The goal is to:

- Specify and analyse the data sources in order to plan and manage the following activities.
- Important 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
  - ….

# EXAMPLE

## Identification and analysis of data sources

```
<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-l"/>
  </LexicalEntry>
   …
```

# Guidelines for ML Linked Data

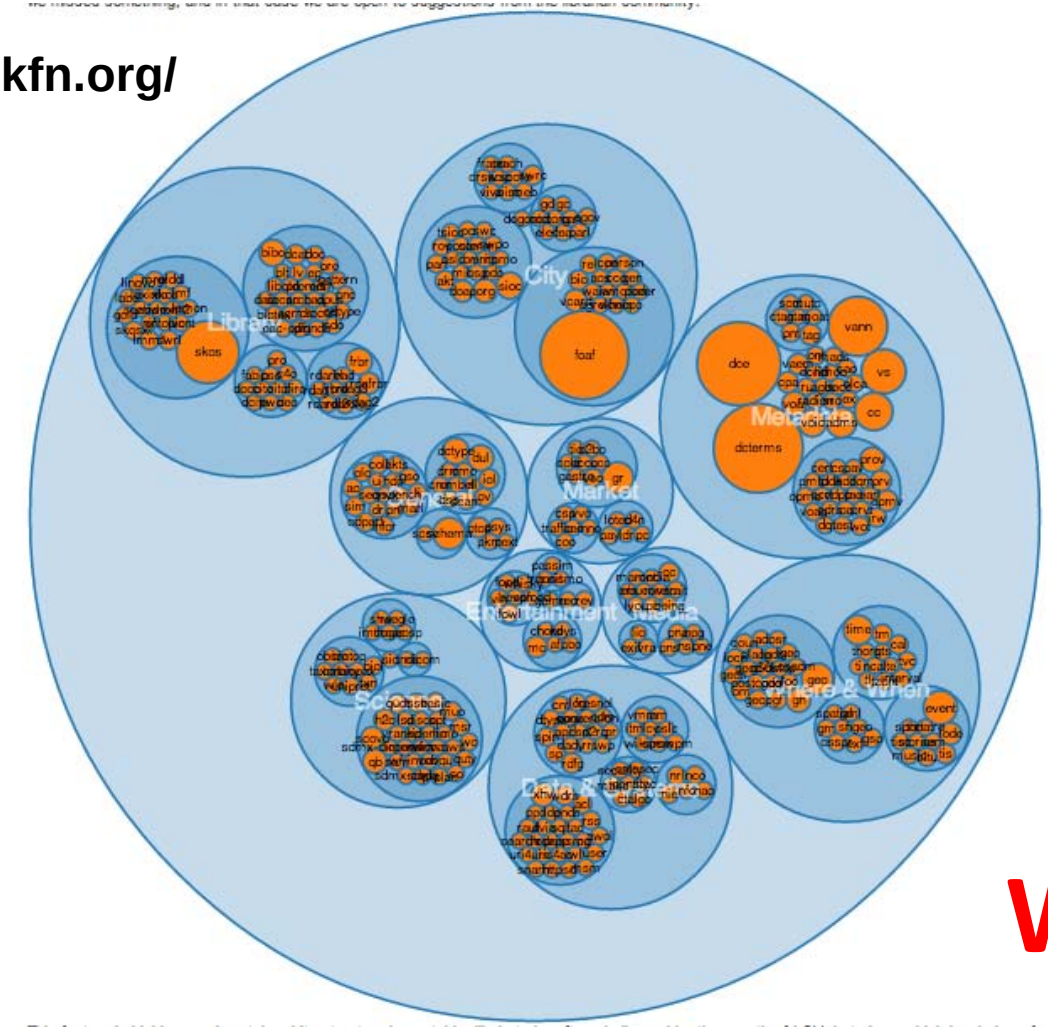
- Set of main activities:
  1. Analysis of data sources
  2. Modelling
  3. URI/IRI design
  4. Generation
  5. Publication

# 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 and selection of domain vocabularies

Use **http://lov.okfn.org/**




lemon

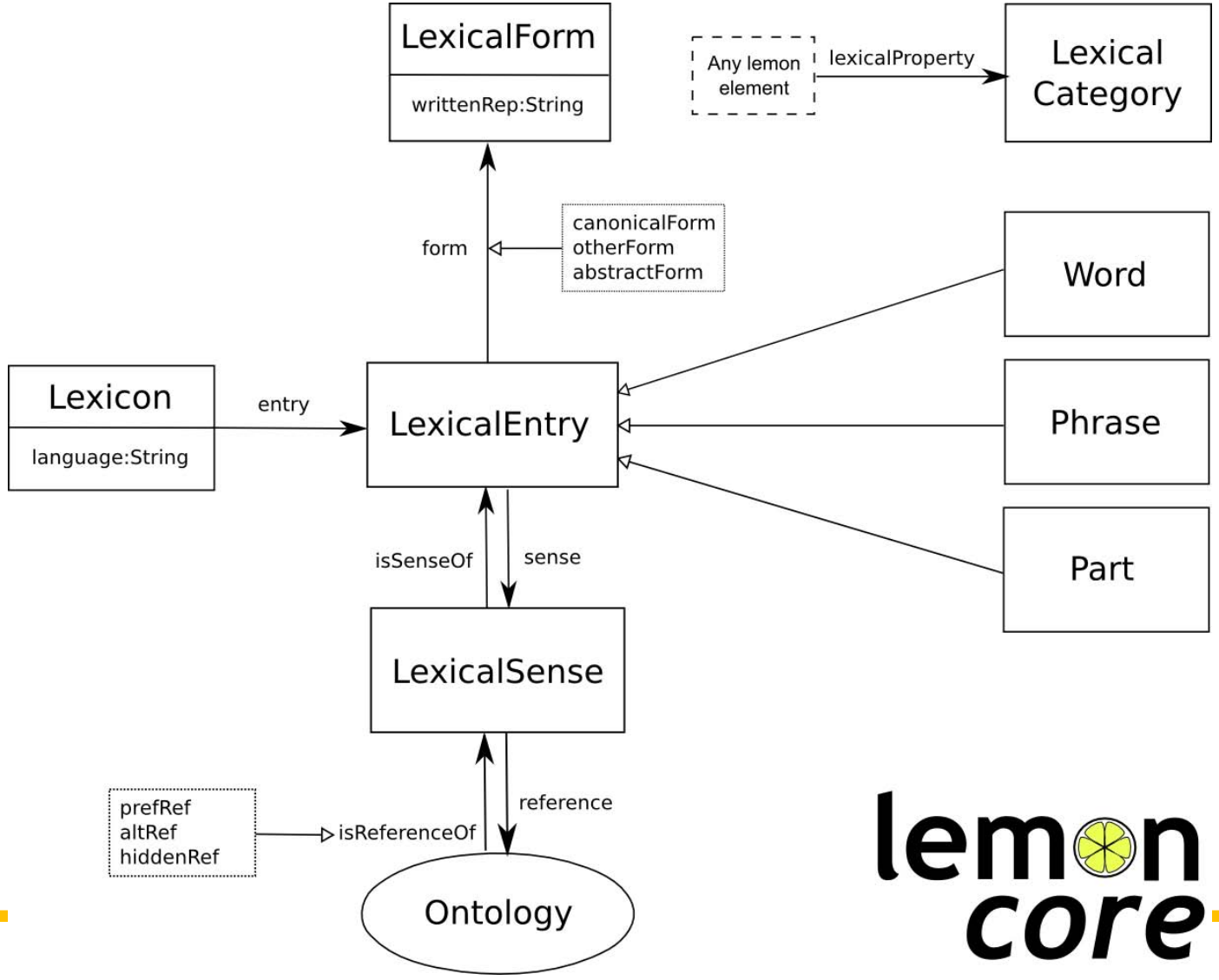
**LexInfo**

**NIF**

NLP Interchange Format

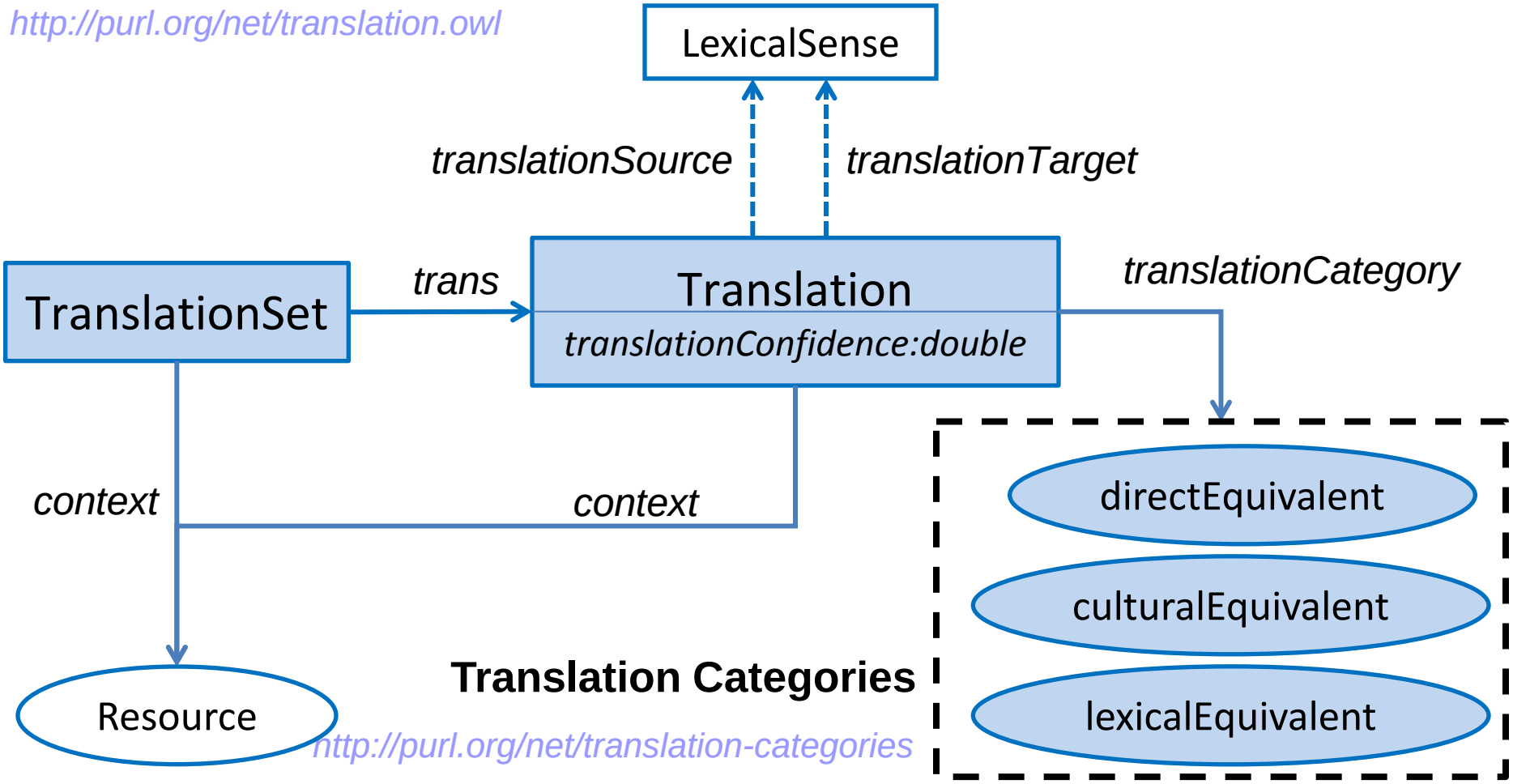
**Wordnet**

# lemon

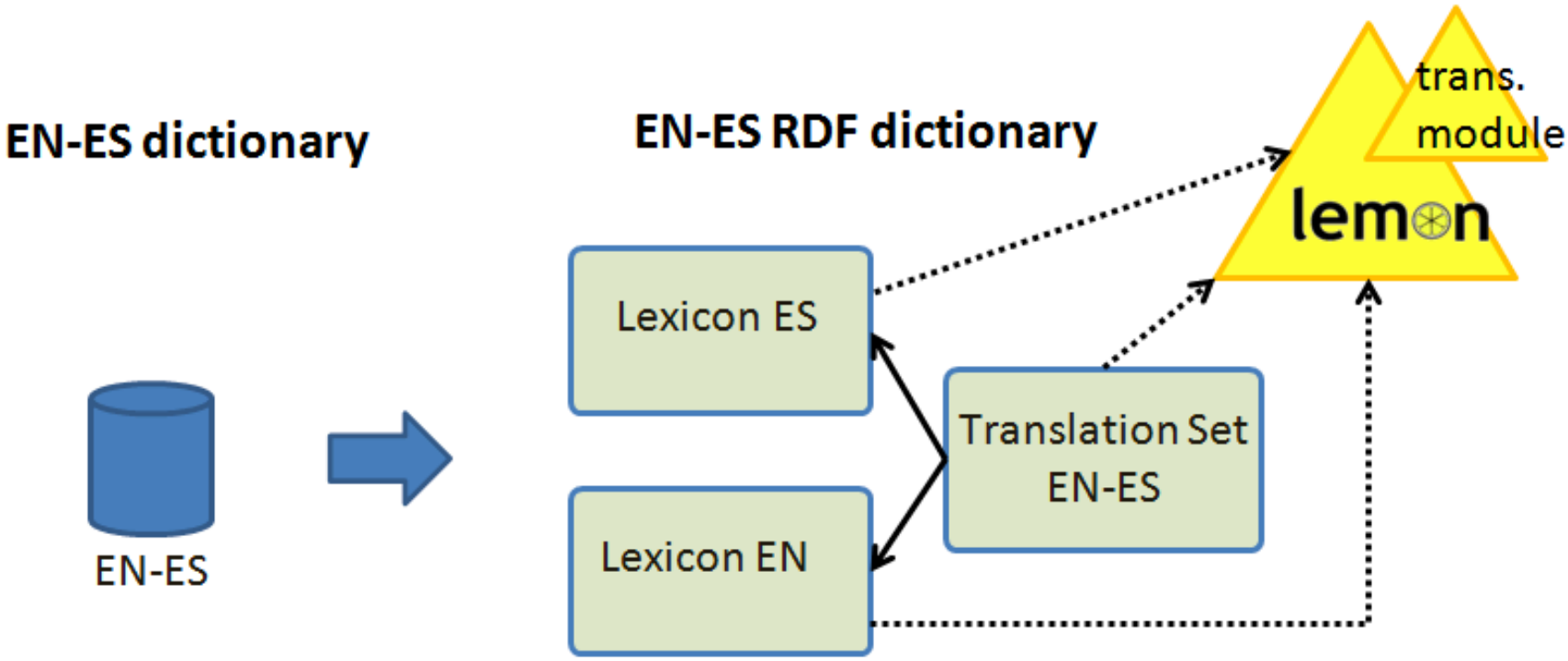

LexicalForm
writtenRep:String
Any lemon element
lexicalProperty
Lexical Category
form
canonicalForm
otherForm
abstractForm
Word
Lexicon
language:String
entry
LexicalEntry
Phrase
Part
isSenseOf
sense
LexicalSense
prefRef
altRef
hiddenRef
isReferenceOf
reference
Ontology
lemon core

## Translation Module

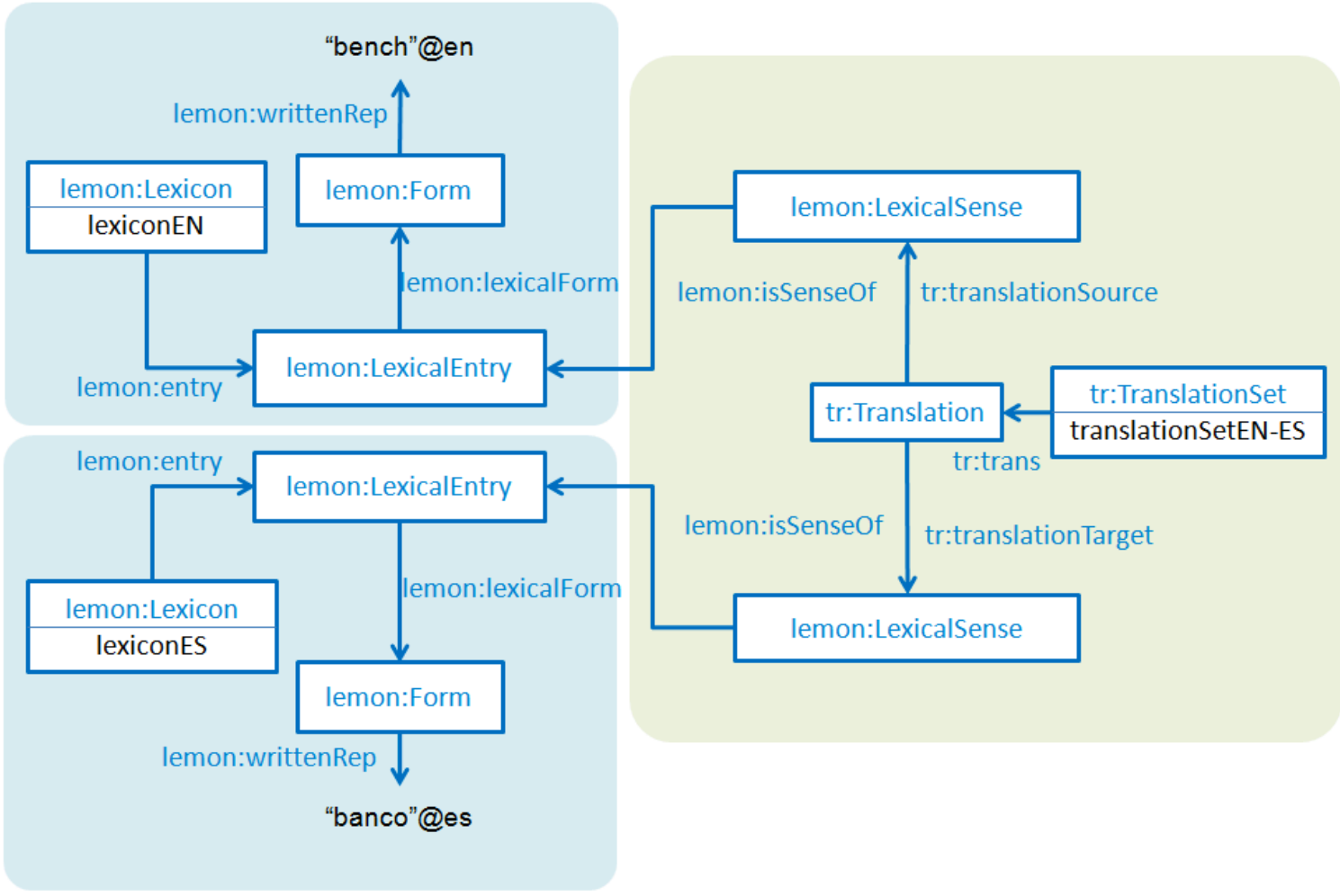
http://purl.org/net/translation.owl



# Mapping of data sources



# Mapping of data sources in detail



- Set of 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 colision**.

*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.,

# URI/IRI design

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 technologies for RDF generation
2. Mapping of data sources to RDF
3. Transformation of data sources to RDF

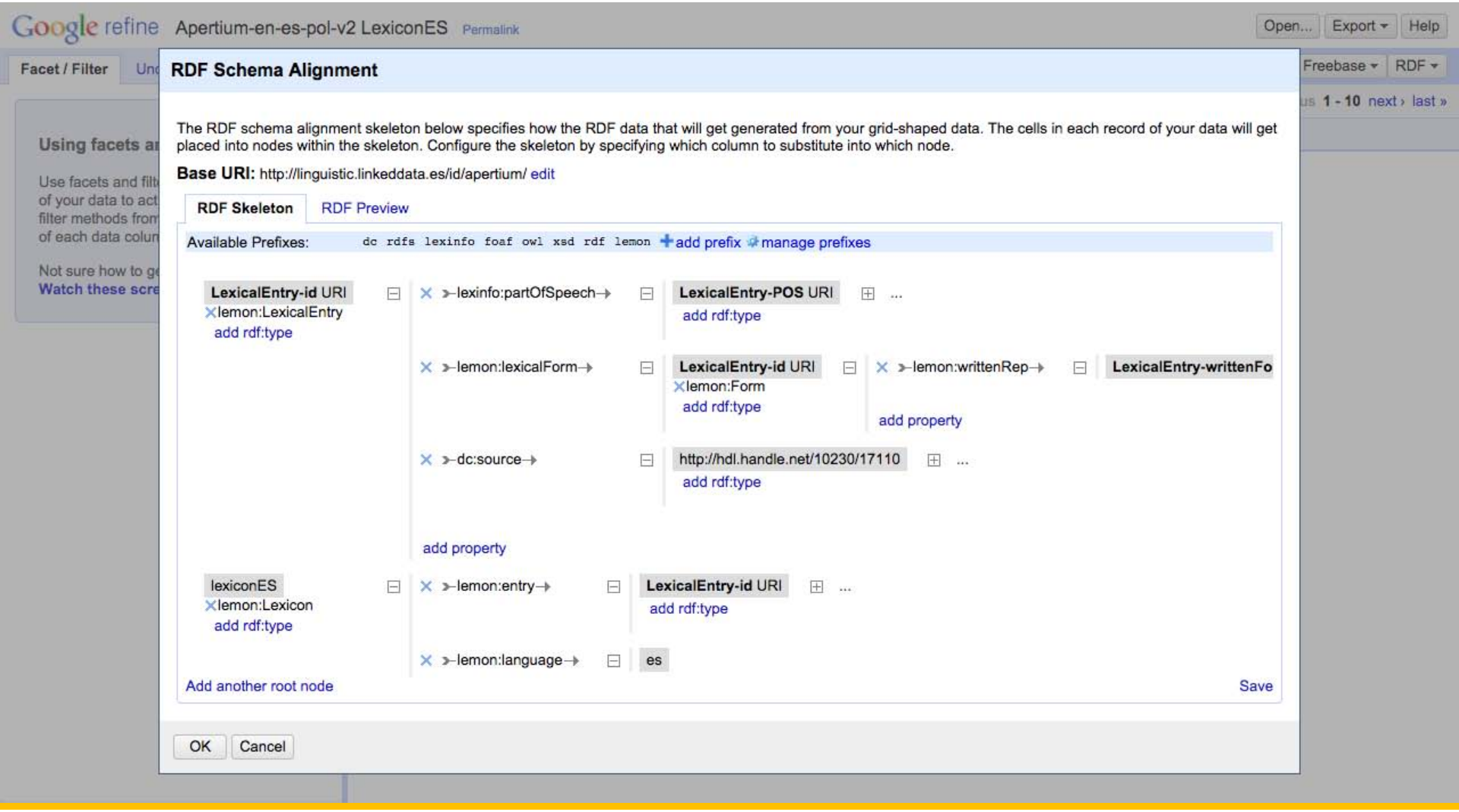
# Goal:

```
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 .
```

# RDF schema alignment



# Guidelines for ML Linked Data

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

# Publication: Intro

- The goal is to:
  - Make available the RDF dataset following LD 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

ODRL
Open Digital Rights Language

PROV
W3C Provenance Ontology

# License vocabularies and mechanisms

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 ?**

creative commons

**Yes**

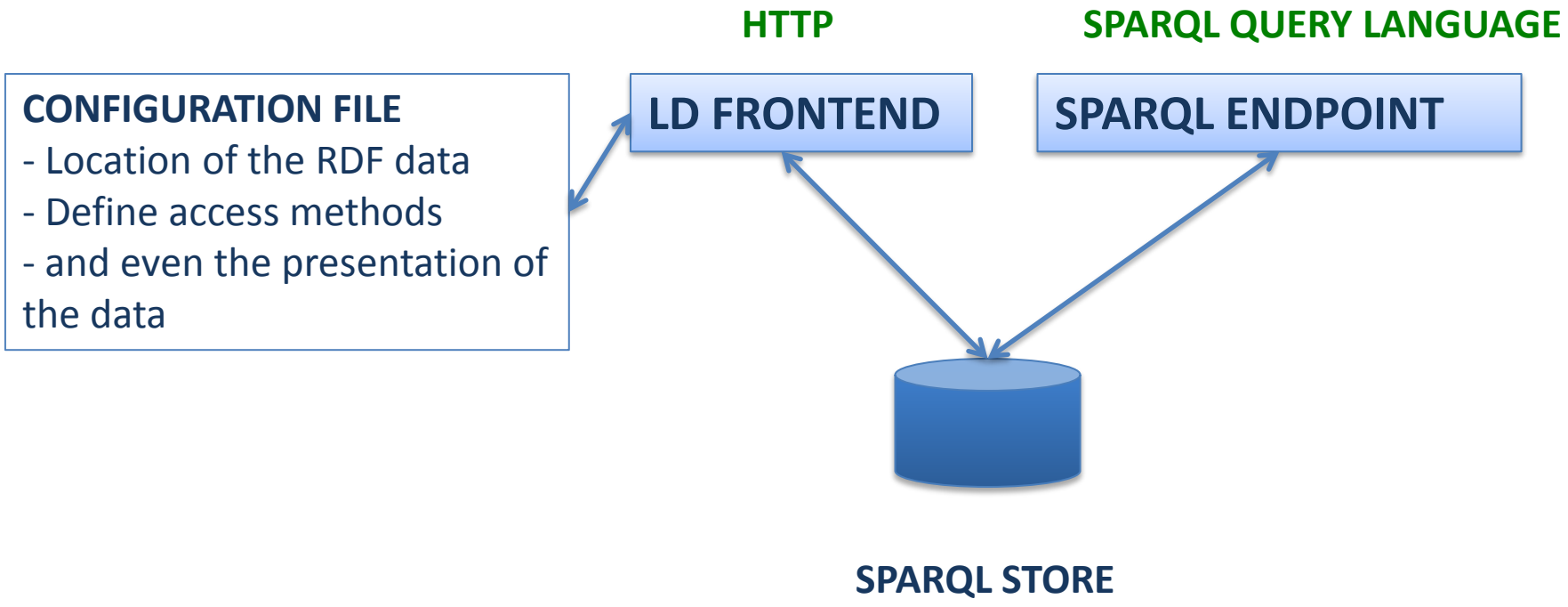
3a. Use **URI** of **standard license** e.g., CC0

**No**

ODRL Open Digital Rights Language

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

# Dataset and vocabulary publication on the Web



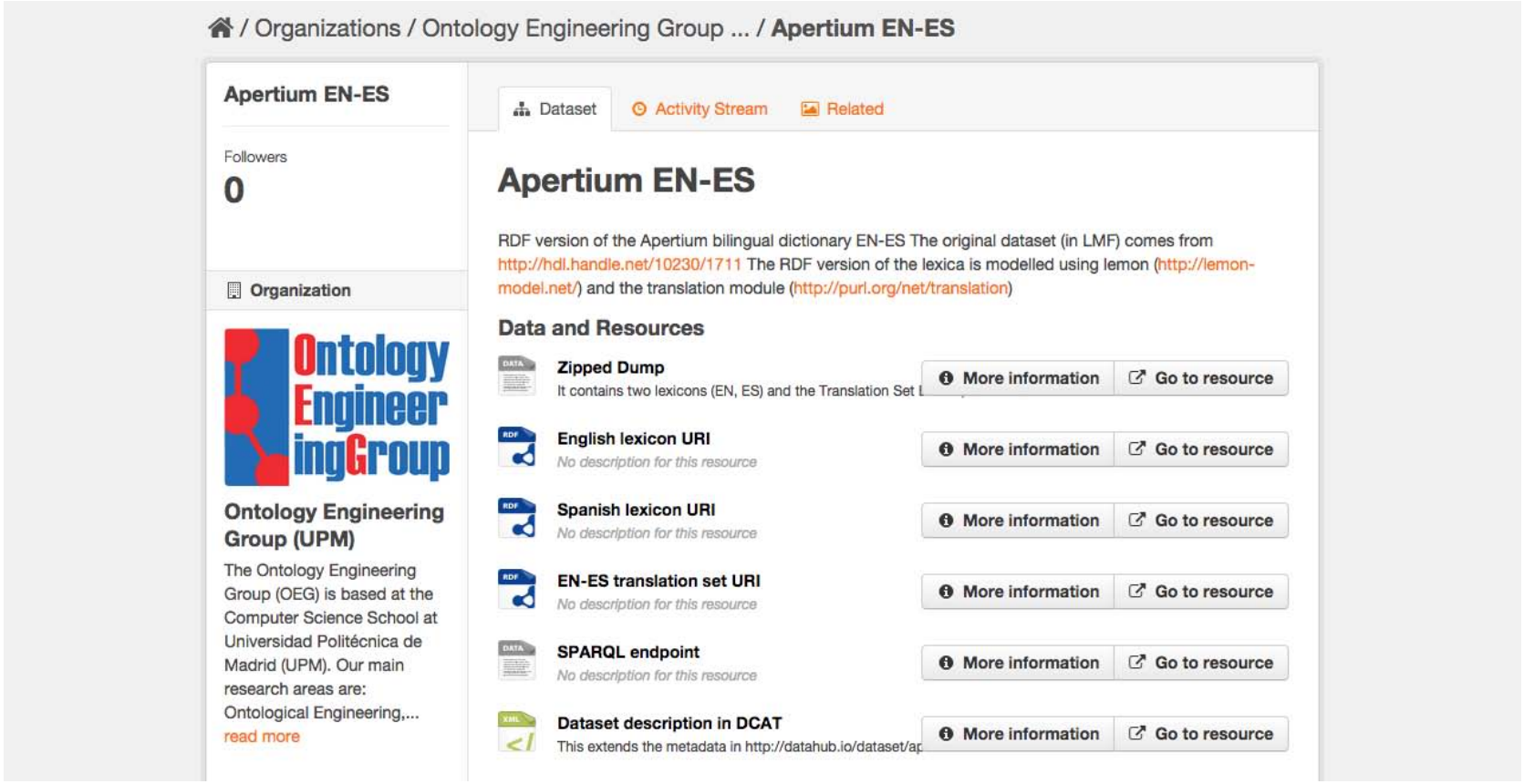
How:

1) Register dataset in datahub.io

2) (Extend generated DCAT file and link to datahub.io one)

# EXAMPLE
## Datahub entry

http://datahub.io/dataset/apertium-en-es



# Extending DCAT description

```
<dcat:Dataset rdf:about="http://linguistic.linkeddata.es/set/apertium/EN-ES">

   <owl:sameAs rdf:resource="http://datahub.io/dataset/apertium-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 an **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)

# Conclusions

- 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