



# Linked Data and Lexicography

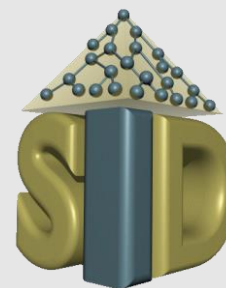
Jorge Gracia (University of Zaragoza, Spain)

# Acknowledgements

This presentation relies on work developed by a larger group of collaborators



Julia Bosque-Gil



Ontology-Lexica community group

# OUTLINE

1. *Benefits of LLD in Lexicography*
2. *K Dictionaries Global series*
3. *Other examples*
4. *Lexicog module*

# *Benefits of LLD in Lexicography*

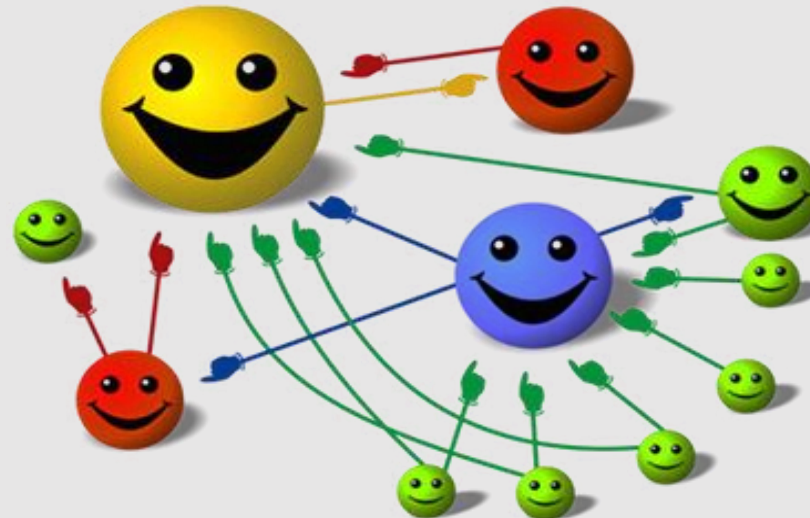


J. Bosque-Gil, J. Gracia, and A. Gómez-Pérez, “Linked data in lexicography,” *Kernerman Dictionary News*, K  
DICTIONARIES LTD, pp. 19–24, Jul-2016.

# Linked Data for lexicography: the value of linking

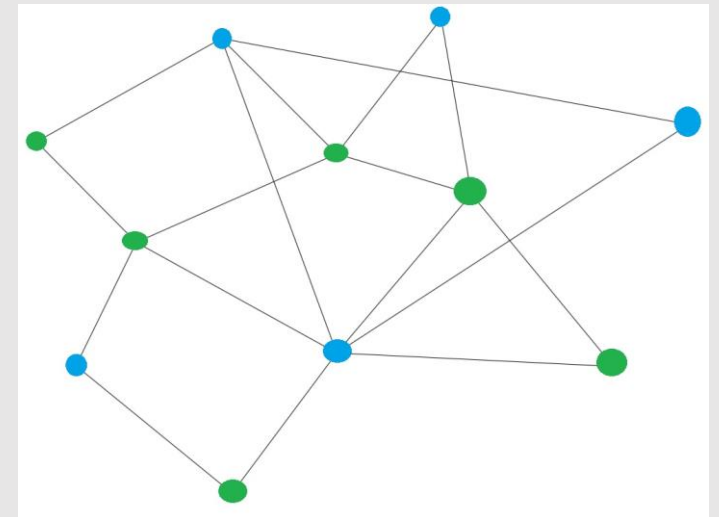
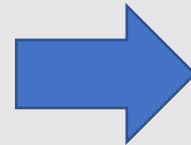
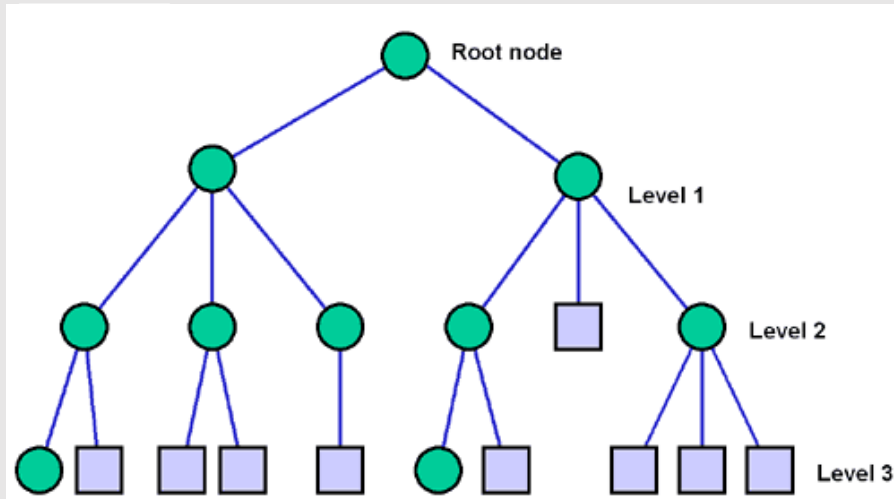
LD as an ideal [common representation framework](#) for lexicographic resources independently of theory, target audience, language, etc.

Recent work focused on the [conversion of already available dictionaries](#) on the basis of de-facto standards such as *lemon* (El Maarouf et al. 2014, Villegas and Bel 2015, Declerck and Mörth 2016, Khan et al. 2016, etc., Bosque et al. 2016) or new ontologies (Parvizi et al. 2016).



# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies **shifting from a tree-structure based view to a graph-based view** on the data.



# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with tree-structures:

- Hindered direct access to embedded data vs. multiple access points
- Internal reuse during compilation not ensured:



# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with tree-structures:

- Hindered direct access to embedded data vs. multiple access points
- Internal reuse during compilation not ensured:

E.g. 1) Multiword items

## *bow and scrape*



*» to treat someone who is powerful or wealthy in an extremely respectful way especially in order to get approval, friendship, etc.*

<https://www.merriam-webster.com/dictionary/bow%20and%20scrape>



# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with tree-structures:

- **Hindered direct access** to embedded data vs. **multiple access points**
- **Internal reuse** during compilation **not ensured**:

E.g. 1) Multiword items



**bow** and scrape

**bow.** *v. intransitive*

» **sense 2.** *to bend the head, body, or knee in reverence, submission, or shame*

# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with tree-structures:

- **Hindered direct access** to embedded data vs. **multiple access points**
- **Internal reuse** during compilation **not ensured**:  
E.g. 1) Multiword items



*bow and scrape* - - - - -

***scrape.*** *v. intransitive*

» ***sense 3.*** *to draw back the foot along the ground in making a bow*

# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies **shifting from a tree-structure based view to a graph-based view** on the data.

Some issues with **tree-structures**:

- **Hindered direct access** to embedded data vs. **multiple access points**
- **Internal reuse** during compilation **not ensured**:

E.g. 2) Synonyms/antonyms/cross-references embedded in an entry

Information is not always shared across entries, redundancies are not prevented, embedded entries might have not being defined



```
<Headword>antiguado</Headword>
[... ]
<SenseBlock>
  <SenseGrp identifier="SE00005713"
version="1">
    <Synonym>antiguo</Synonym>
```

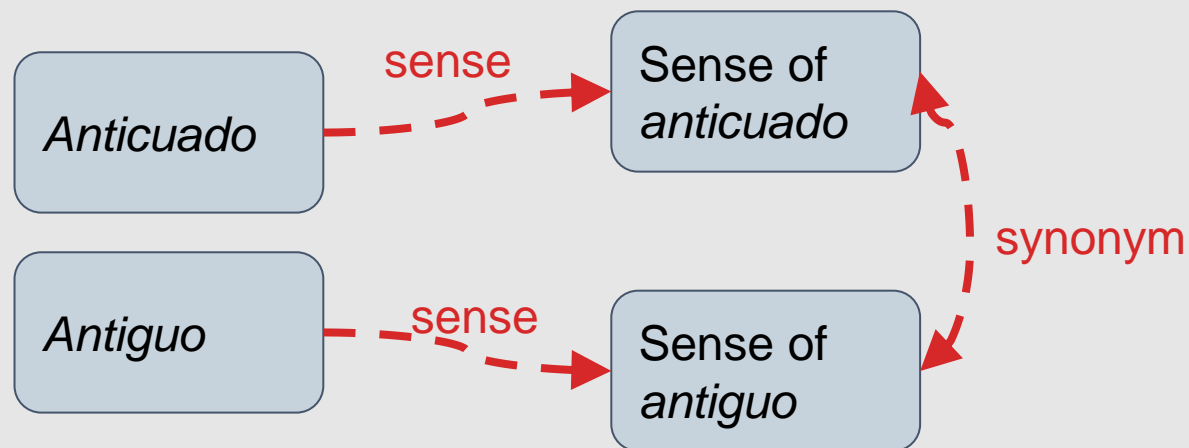
# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with tree-structures:

- **Hindered direct access** to embedded data vs. **multiple access points**
- **Internal reuse** during compilation **not ensured**:

E.g. 2) Synonyms/antonyms/cross-references embedded in an entry



# Linked Data for lexicography: the value of graphs

Adopting the LD paradigm implies shifting from a tree-structure based view to a graph-based view on the data.

Some issues with common dictionary design decisions:

- Cross-references not typed with ontologically defined properties
- Reliance on internal IDs vs. transparent URIs naming strategies
- Need to keep track of order (e.g. in homographs)

# *KDictionaries Global series (an example of a dictionary converted into LD)*

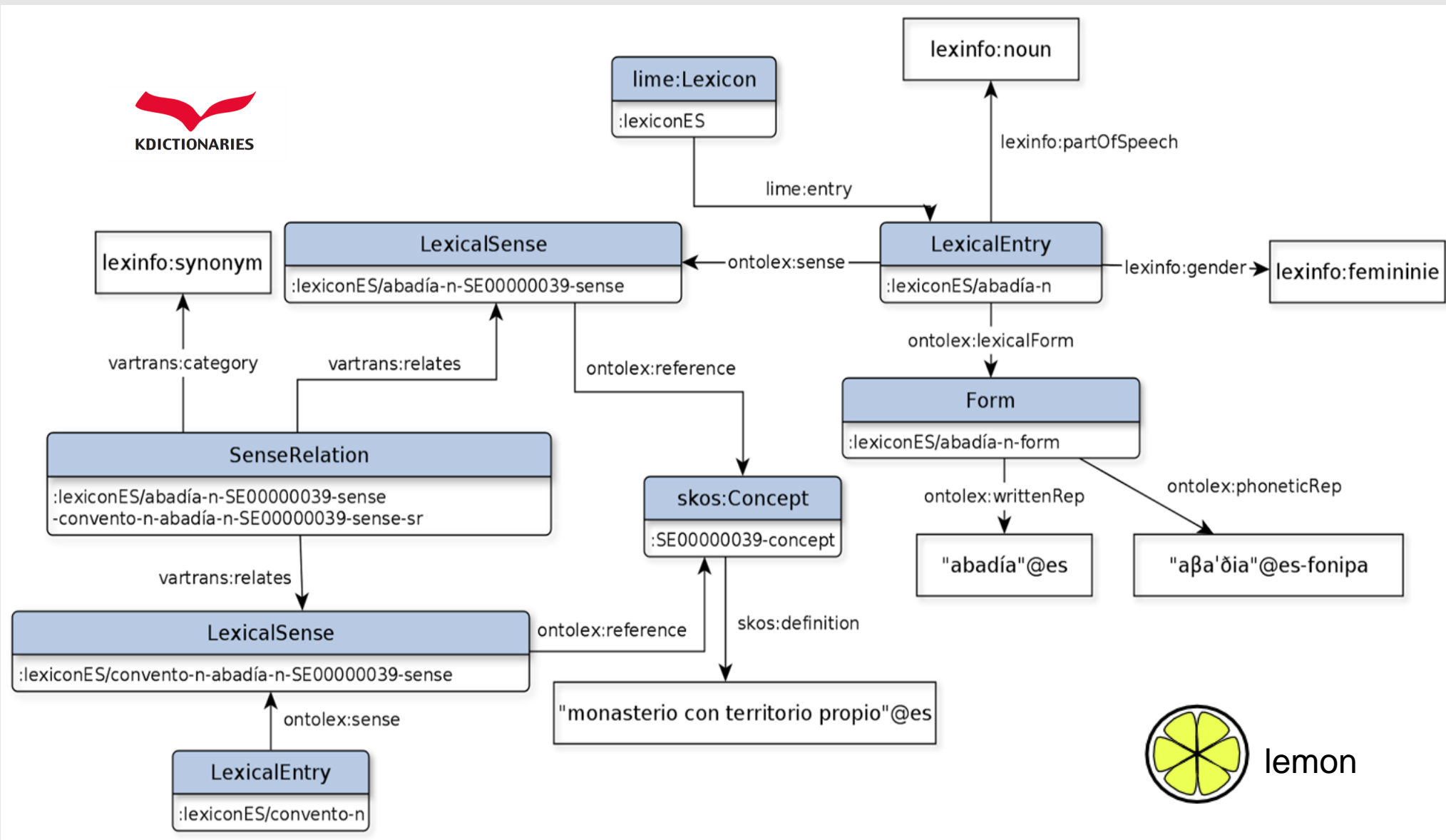


J. Bosque-Gil, J. Gracia, E. Montiel-Ponsoda, and G. Aguado-de-Cea, “Modelling Multilingual Lexicographic Resources for the Web of Data: the K Dictionaries case,” in *Proc. of GLOBALEX’16 workshop at LREC’15, Portoroz, Slovenia*, 2016.



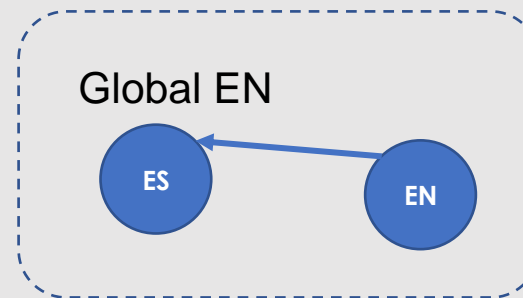
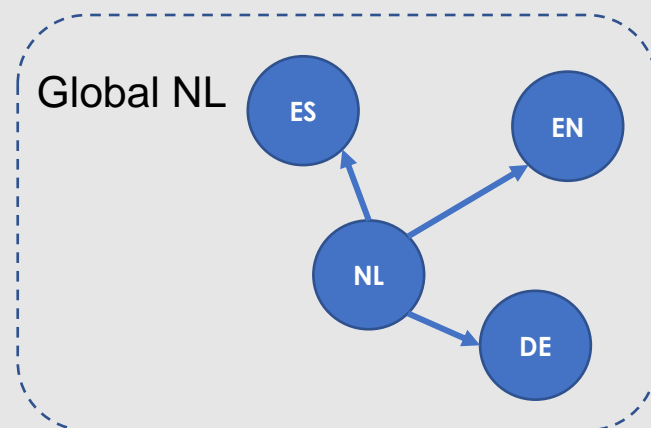
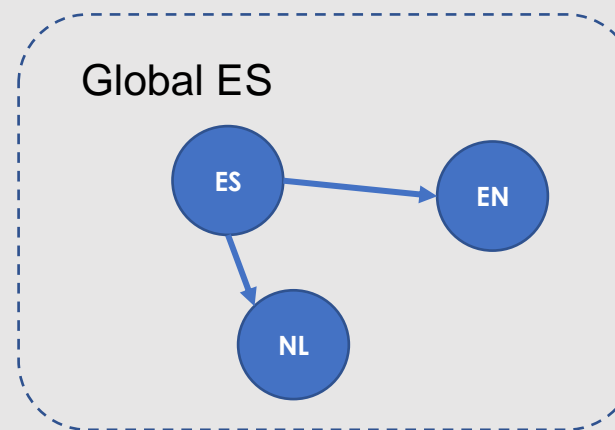
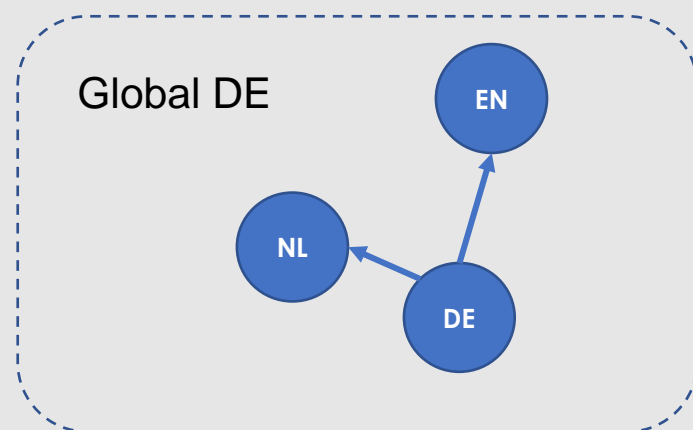
J. Bosque-Gil, D. Lonke, J. Gracia, and I. Kernerman, “Validating the OntoLex-lemon Lexicography Module with K Dictionaries’ Multilingual Data,” in *Proc. of 6th biennial conference on electronic lexicography, eLex 2019*, 2019, pp. 726–746.

# A LD example from KDictionaries

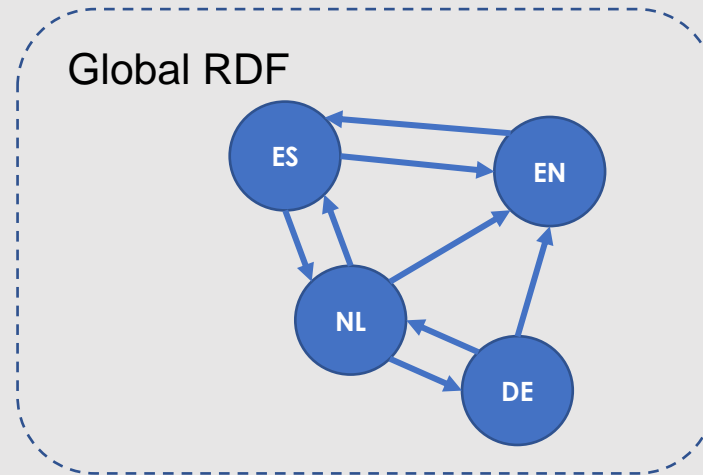




# From multilingual to cross-lingual linked dictionaries



# From multilingual to cross-lingual linked dictionaries



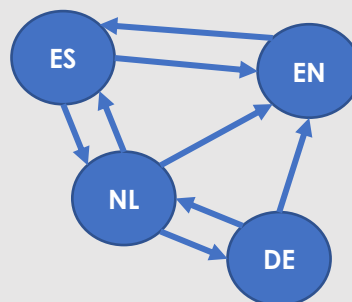
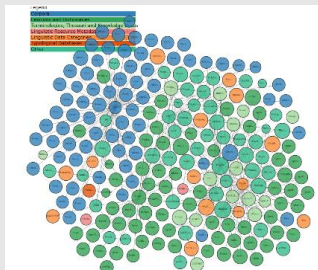
# In the Lynx project

Lynx EU Project (“Legal Knowledge Graph for Multilingual Compliance Services”)

Lynx Services (WSD, WSI, ...)



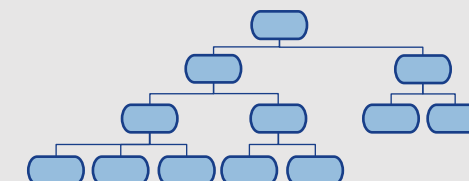
**External  
vocabularies and  
linked data  
resources**



**Lynx domain  
independent  
vocabularies**



**Lynx domain  
dependent  
vocabularies**



# Example query

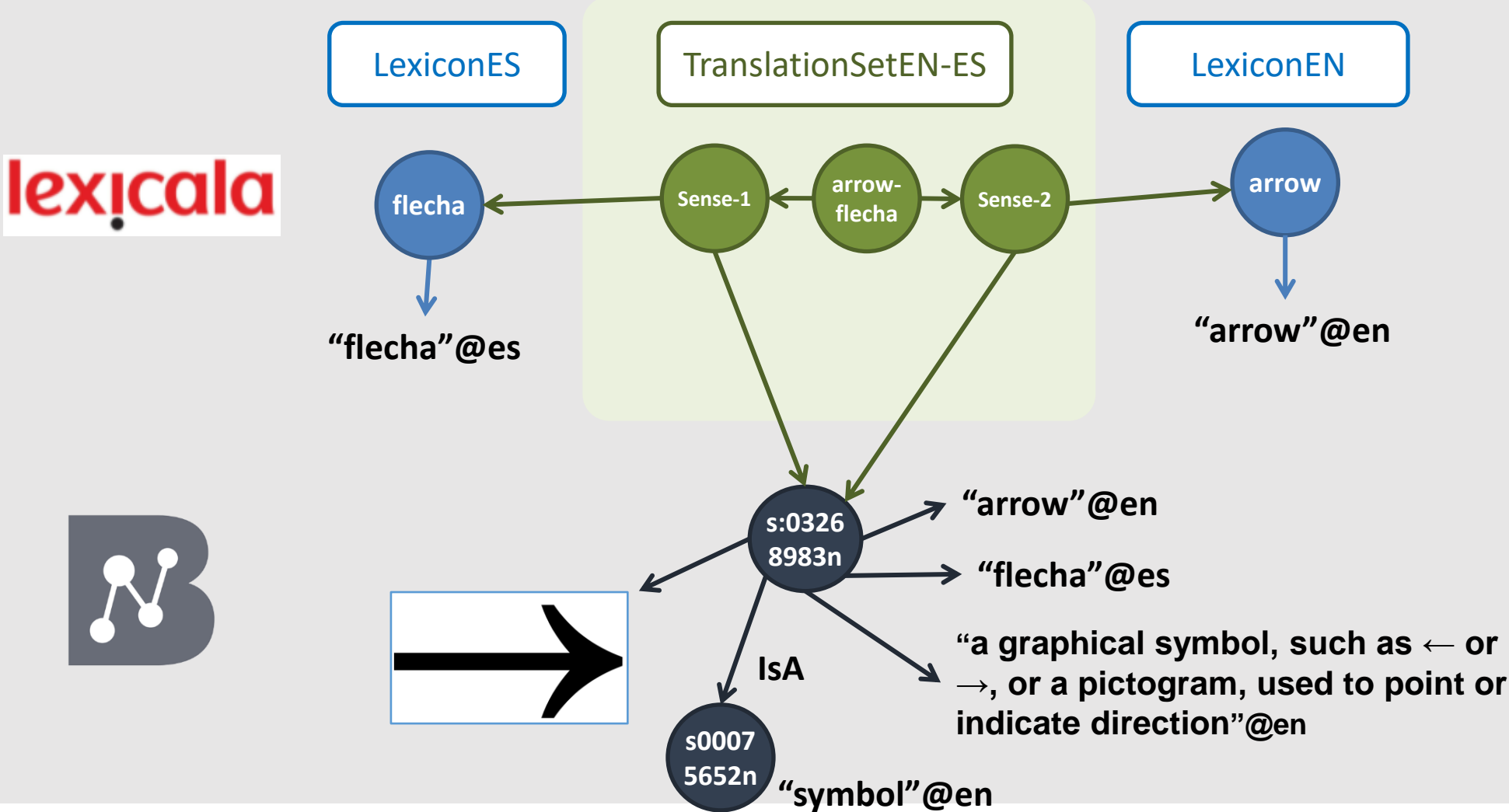
*Give me all lexical entries that evoke the concept of  
**law** or **labour law***

```
SELECT DISTINCT ?entryA {  
  ?concept a ontalex:LexicalConcept;  
           skos:definition ?def .  
  ?senseA ontalex:isLexicalizedSenseOf ?concept .  
  ?senseB ontalex:isLexicalizedSenseOf ?concept .  
  ?entryA ontalex:sense ?senseA .  
  ?entryB ontalex:sense ?senseB .  
  FILTER (?senseA!=?senseB)  
  FILTER regex (?def, "ley | derecho laboral")  
}
```

## RESULTS

```
"http://lexicala.com/id/LexiconES/acatamiento-n"  
"http://lexicala.com/id/LexiconES/abrogar-vb" ,  
"http://lexicala.com/id/LexiconEN/to_abrogate-vb" ,  
"http://lexicala.com/id/LexiconEN/to_repeal-vb"  
"http://lexicala.com/id/LexiconEN/abrogation-n" ,  
....
```

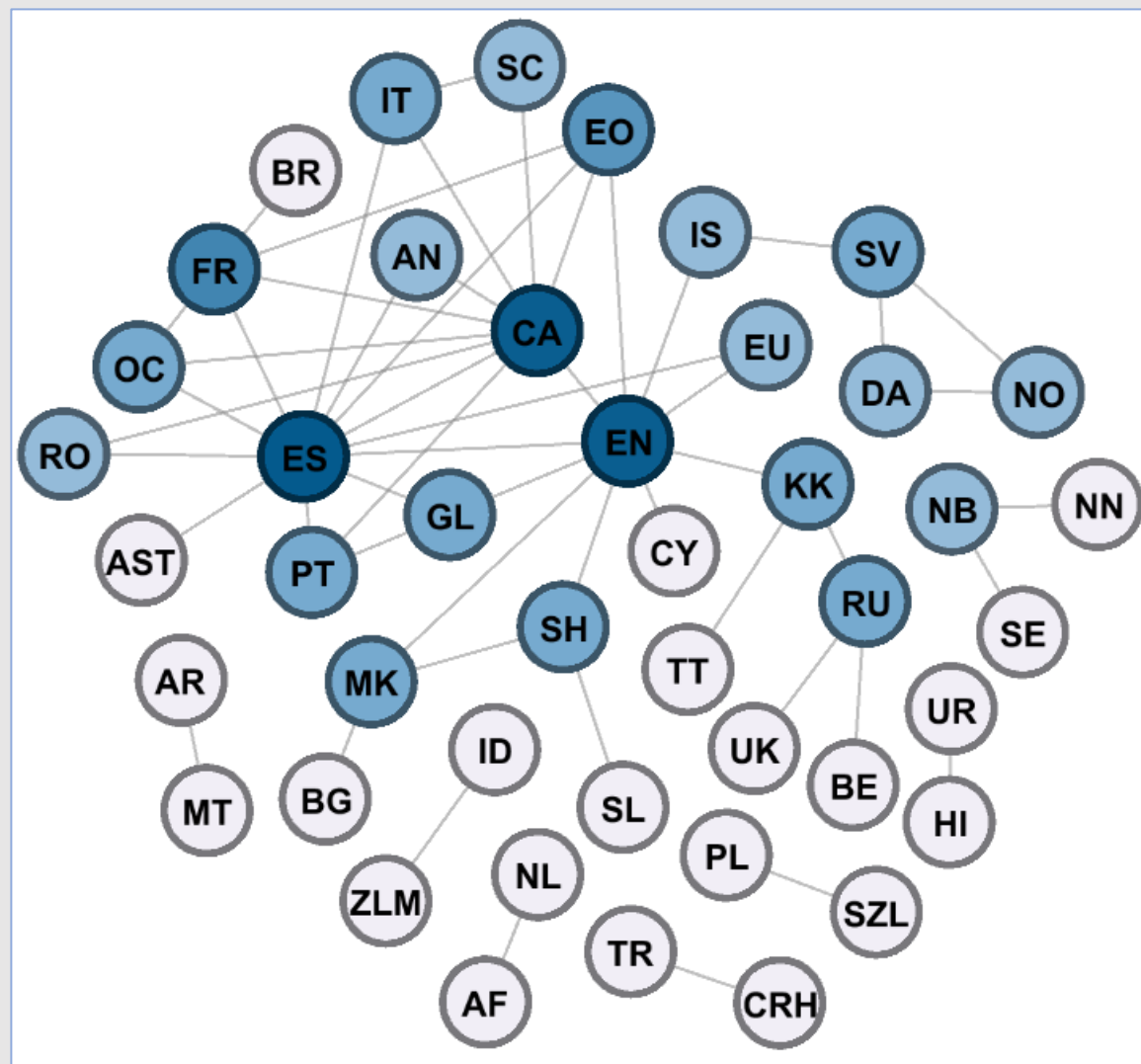
# Example of external linking



# Other examples

# Apertium RDF graph

Apertium RDF v2.0 graph (2020)



J. Gracia *et al.*, “Leveraging Linguistic Linked Data for Cross-Lingual Model Transfer in the Pharmaceutical Domain,” in *Proc. of 19th International Semantic Web Conference (ISWC 2020)*, 2020, pp. 499–514.



# Wikidata



<https://www.wikidata.org/>

“Wikidata is a free and open knowledge base that can be read and edited by both humans and machines [...] The content of Wikidata is [available under a free license](#), [exported using standard formats](#), and [can be interlinked to other open data sets](#) on the linked data web.”



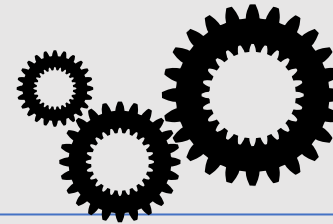
Wikidata [SPARQL service](#)

# DBnary

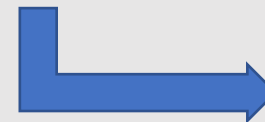


<https://www.wiktionary.org/>

“**Collaborative** project to produce a free-content **multilingual dictionary**. It aims to describe all words of all languages [...].”

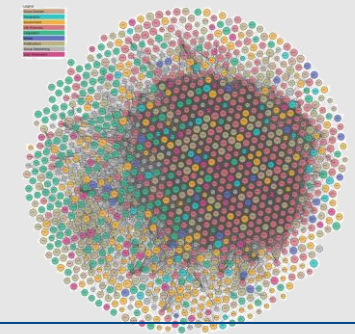


Conversion into RDF +  
publicaton on the Web of Data



**Dbnary**

<http://kaiko.getalp.org/>

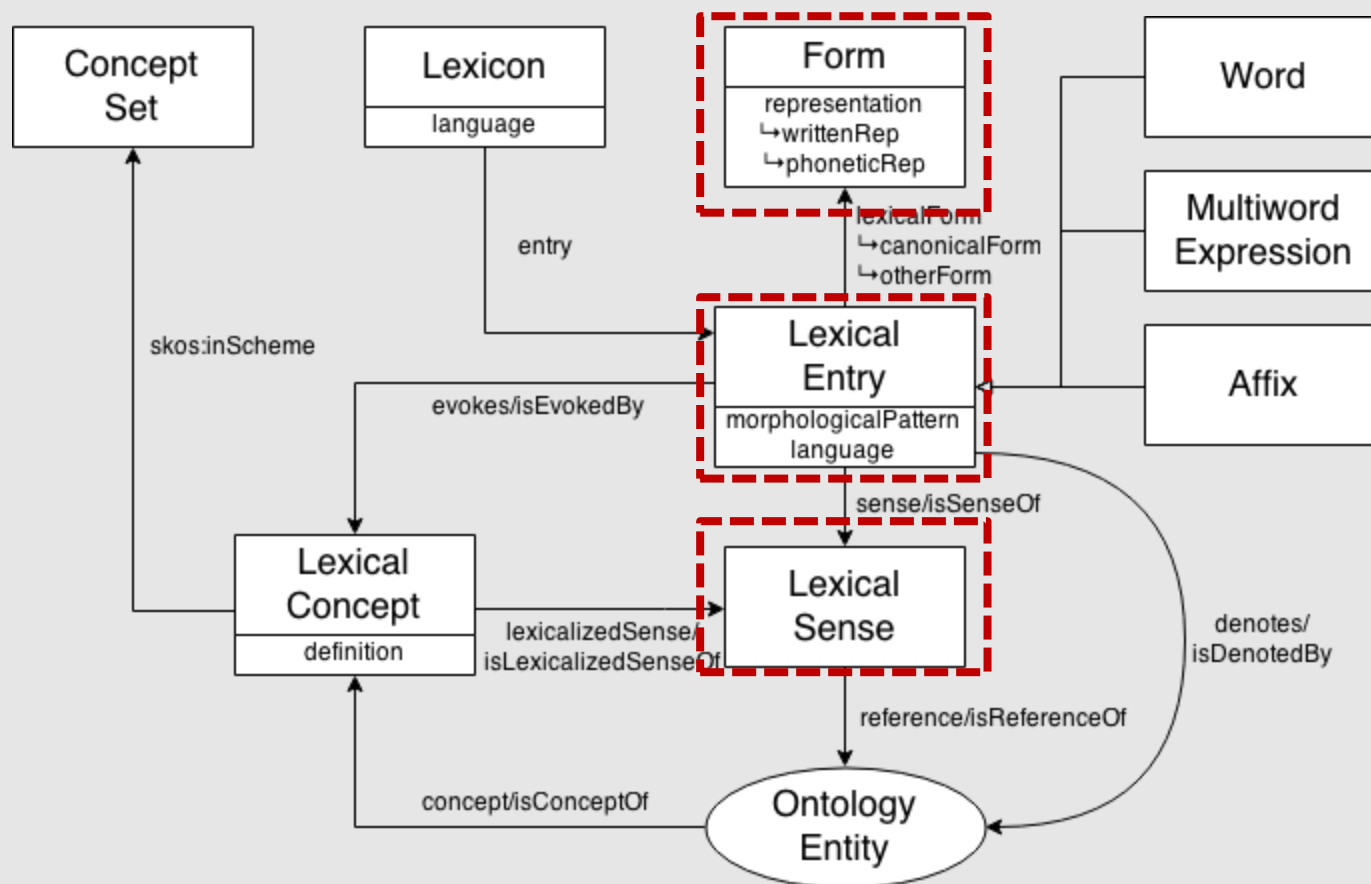


“Dbnary is an effort to provide **multilingual lexical data** extracted from wiktionary. The extracted data is made available as **Linguistic Linked Open Data**”

# The Lexicog model

# Ontolex lemon

## Core of the model



<https://www.w3.org/2016/05/ontolex/>

# Ontolex lexicog module



<https://www.w3.org/2019/09/lexicog/>



## TABLE OF CONTENTS

1. Introduction
  - 1.1 Background and motivation
  - 1.2 Aim and scope
  - 1.3 Namespaces
2. Lexicography Module (lexicog)
  - 2.1 Lexicographic Resource
  - 2.2 Entry
  - 2.3 entry
  - 2.4 Lexicographic Component
  - 2.5 describes
  - 2.6 subComponent
  - 2.7 FormRestriction
  - 2.8 restrictedTo
  - 2.9 UsageExample
  - 2.10 usageExample

## The OntoLex Lemon Lexicography Module

Final Community Group Report 17 September 2019



### Editors:

[Julia Bosque-Gil](#) (Ontology Engineering Group, Universidad Politécnica de Madrid)  
[Jorge Gracia](#) (Aragon Institute of Engineering Research, University of Zaragoza)

### Authors:

[Julia Bosque-Gil](#) (Ontology Engineering Group, Universidad Politécnica de Madrid)  
[Jorge Gracia](#) (Aragon Institute of Engineering Research, University of Zaragoza)  
[John McCrae](#) (Insight Centre for Data Analytics, National University of Ireland, Galway)  
[Philipp Cimiano](#) (Cognitive Interaction Technology Excellence Center, Bielefeld University)  
[Sander Stolk](#) (Centre for the Arts in Society, Leiden University)  
[Fahad Khan](#) (Istituto di Linguistica Computazionale "Zampolli", CNR, Pisa)  
[Katrien Depuydt](#) (Institute for Dutch Lexicology, Leiden, Netherlands)  
[Jesse de Does](#) (Institute for Dutch Lexicology, Leiden, Netherlands)  
[Francesca Frontini](#) (Paul-Valéry University, Montpellier III)  
[Ilan Kernerman](#) (K Dictionaries)

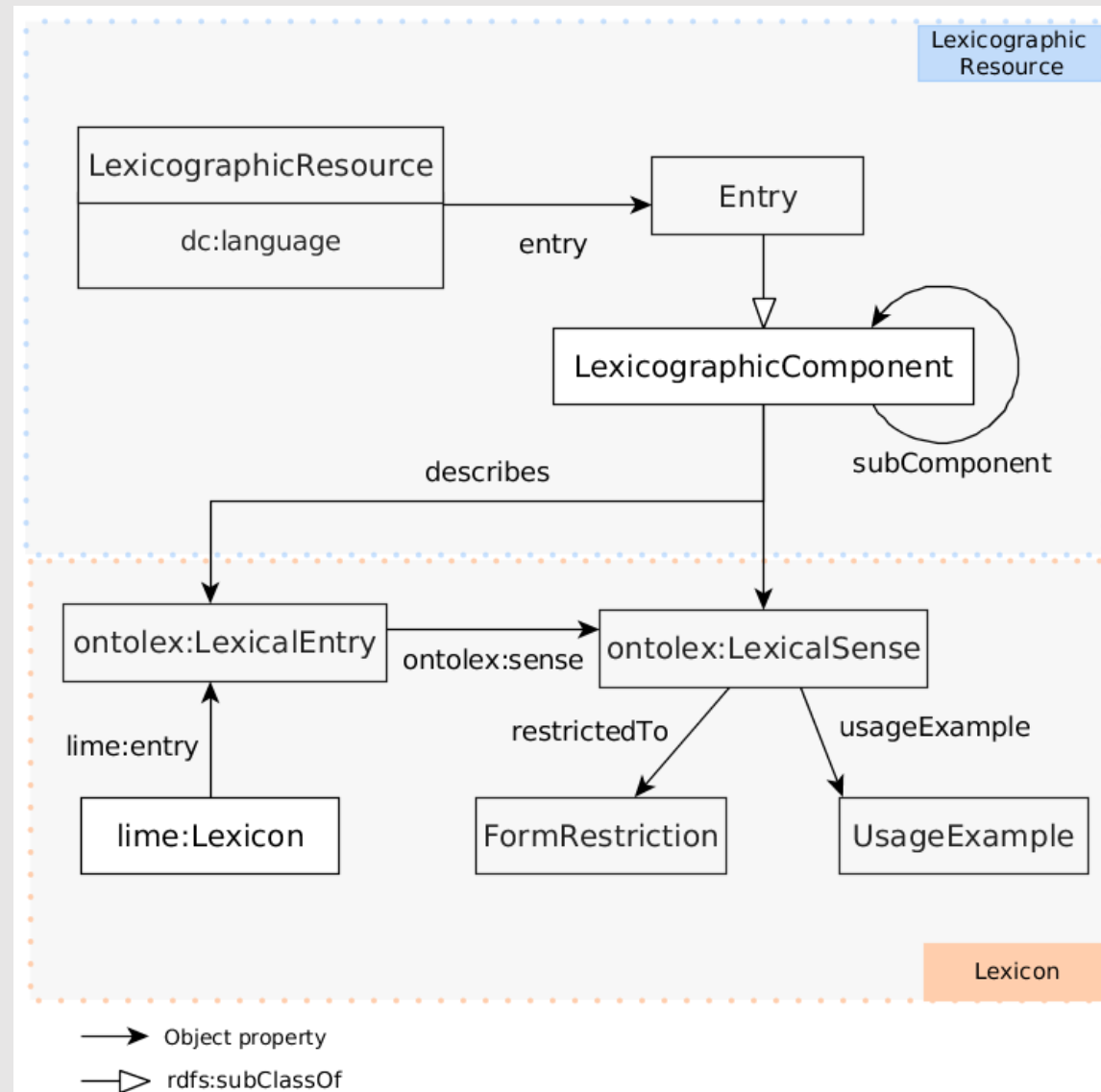
Copyright © 2019 the Contributors to the The OntoLex Lemon Lexicography Module Specification, published by the [Ontology Lexica](#) under the [W3C Community Final Specification Agreement \(FSA\)](#). A human-readable [summary](#) is available.

### Abstract

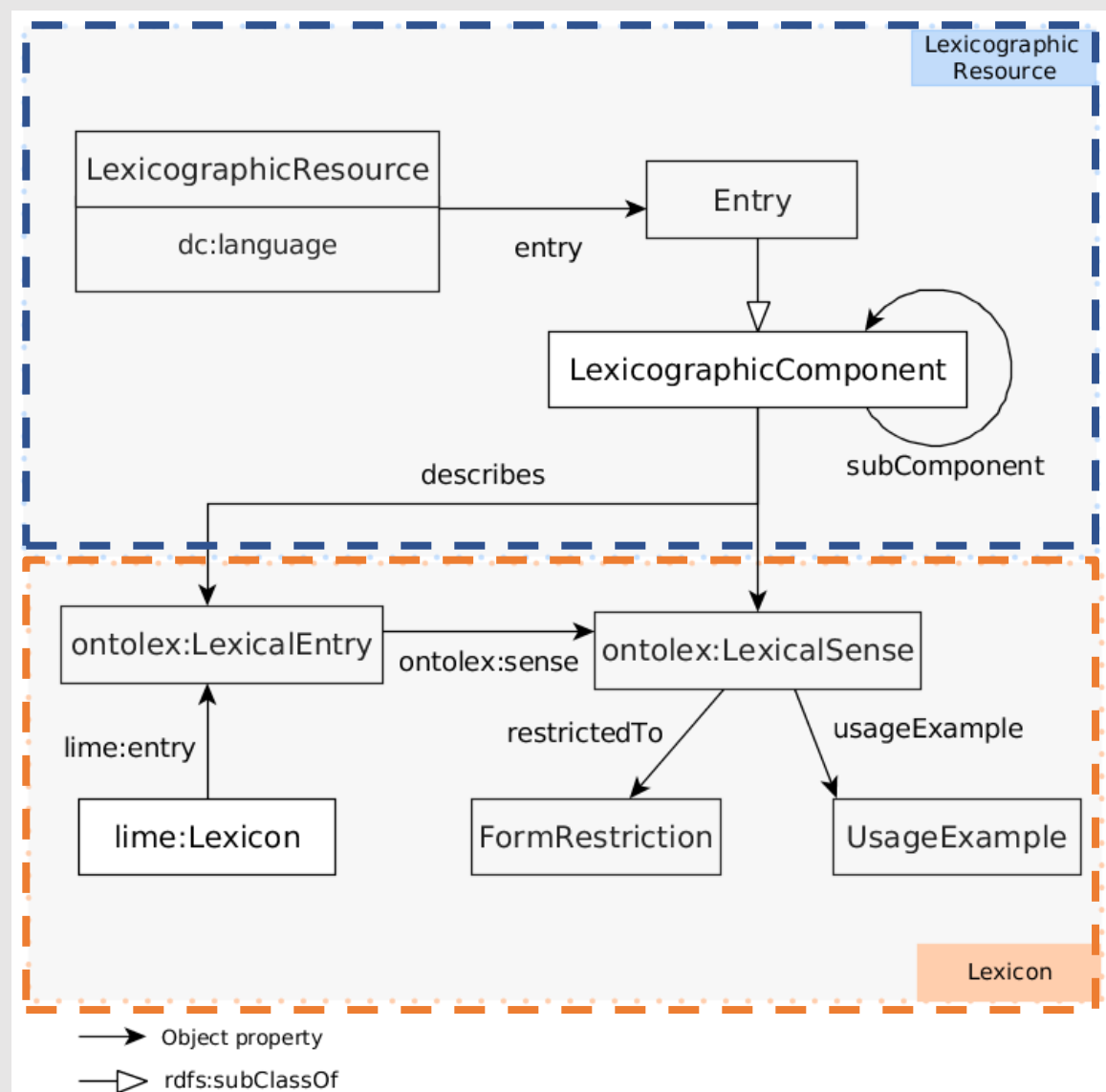
This document describes the *lexicography module* of the Lexicon Model for Ontologies (*lemon*) as a result of the work of the Ontology Lexicon community group (OntoLex). The module is targeted at the representation of

work of the Ontology Lexicon community group (OntoLex). The module is targeted at the representation of  
This document describes the lexicography module of the Lexicon Model for Ontologies (*lemon*) as a result of the

# Ontolex lexicog module



# Ontolex lexicog module





## Example extracted from the American Heritage Dictionary

### **an·i·mal**

n.

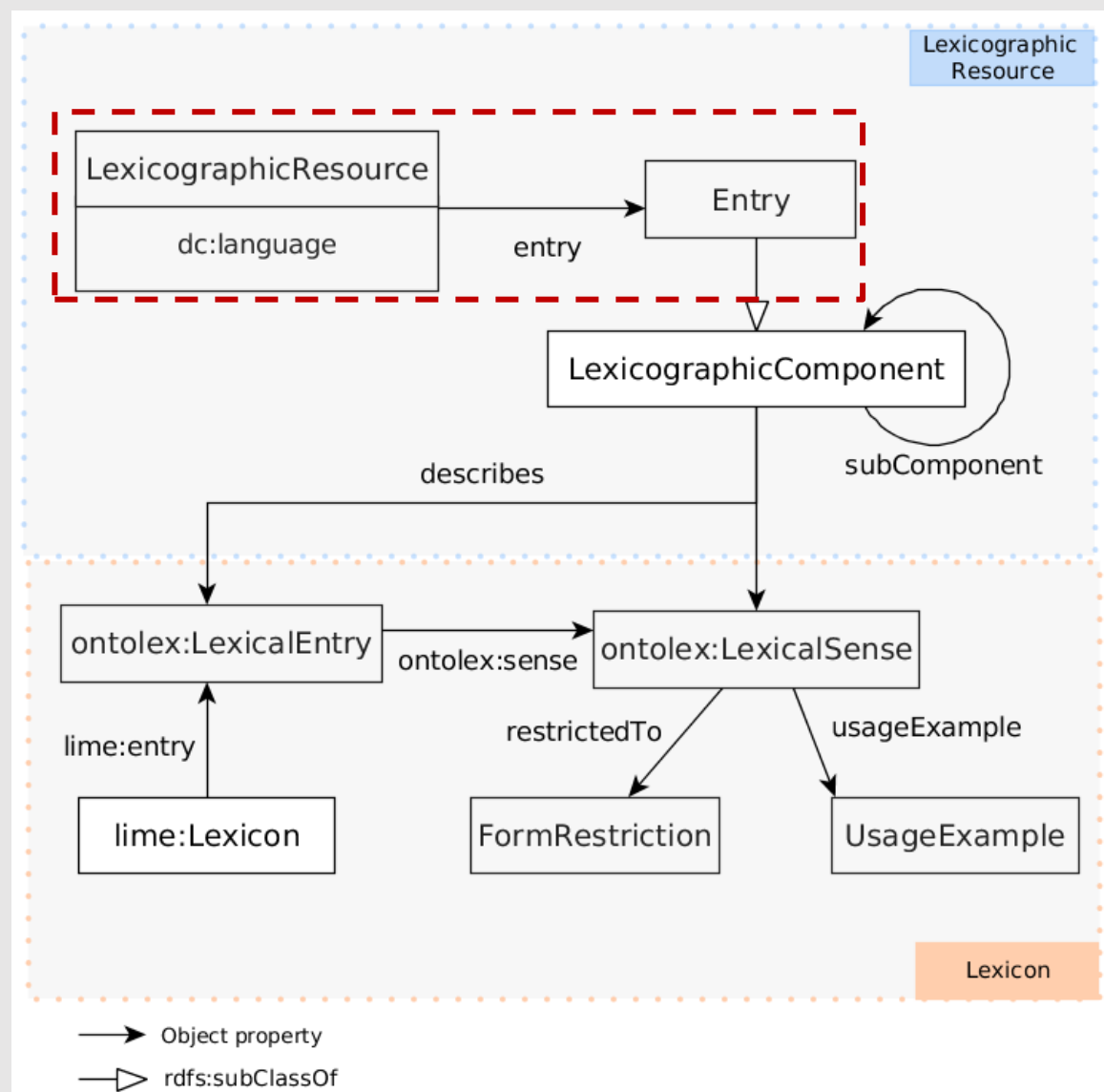
1. Any of numerous multicellular eukaryotic organisms of the kingdom Metazoa (or Animalia) [...]
2. An animal organism other than a human, especially a mammal.

[...]

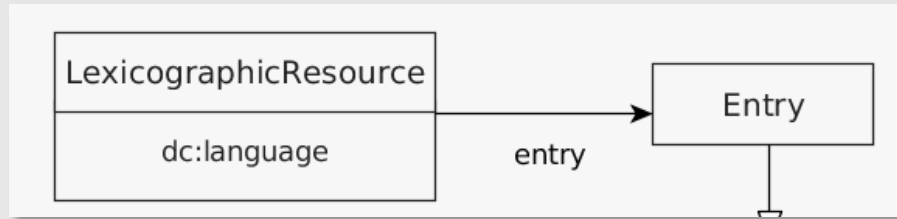
adj.

1. Relating to, characteristic of, or derived from an animal or animals, especially when not human: animal cells; animal welfare.
2. Relating to the physical as distinct from the rational or spiritual nature of people: animal instincts and desires.

# Ontolex lexicog module



# Ontolex lexicog module

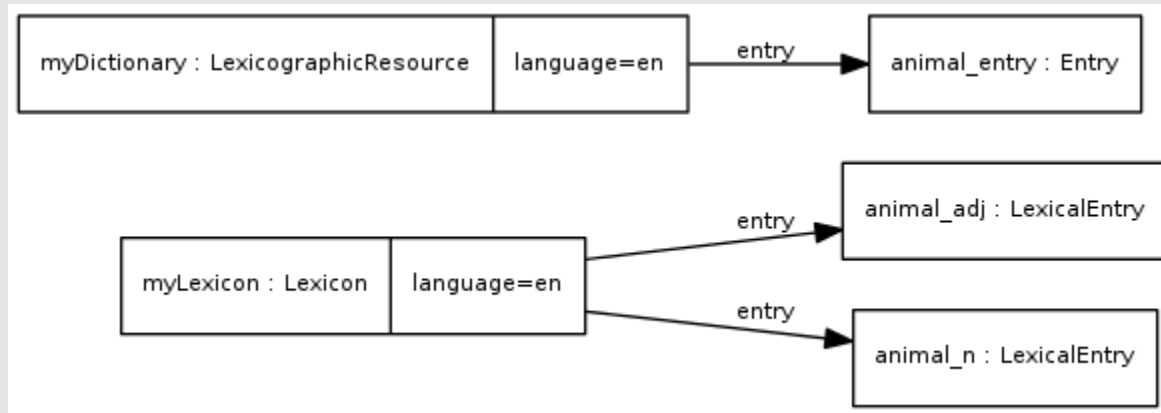
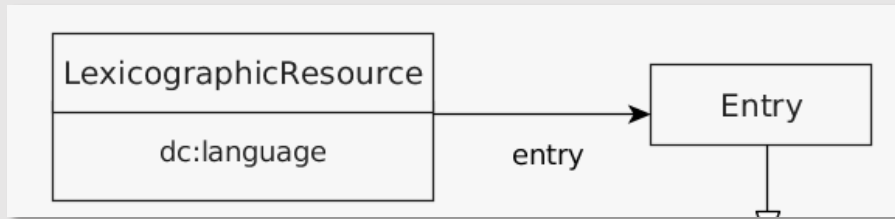


A **Lexicographic Resource** represents a collection of lexicographic entries (lexicog:Entry) in accord with the lexicographic criteria followed in the development of that resource.

An **Entry** is a structural element that represents a lexicographic article or record as it is arranged in a source lexicographic resource. As such, it supports the description of lexical entries or senses according to the lexicographic micro-structure, decided upon during a lexicographic resource compilation process.

The property **entry** relates a lexicog:LexicographicResource to a lexicog:Entry.

# Ontolex lexicog module



## # LEXICOGRAPHIC RESOURCE

```
:myDictionary a lexicog:LexicographicResource ;  
    dc:language "en" ;  
    lexicog:entry :animal_entry .
```

```
:animal_entry a lexicog:Entry .
```

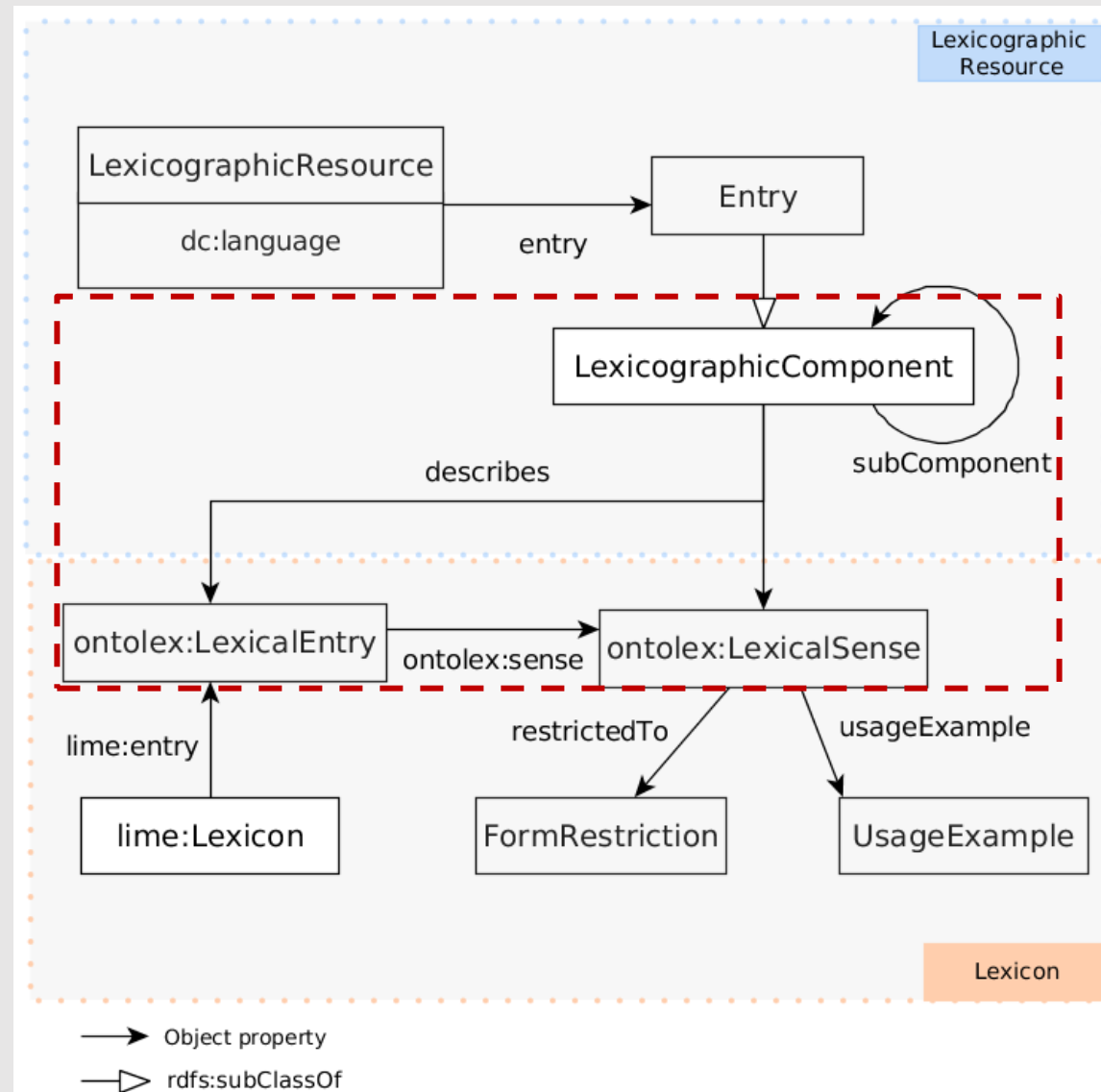
## # LEXICON

```
:myLexicon a lime:Lexicon ;  
    lime:language "en" ;  
    lime:entry :animal_n, :animal_adj .
```

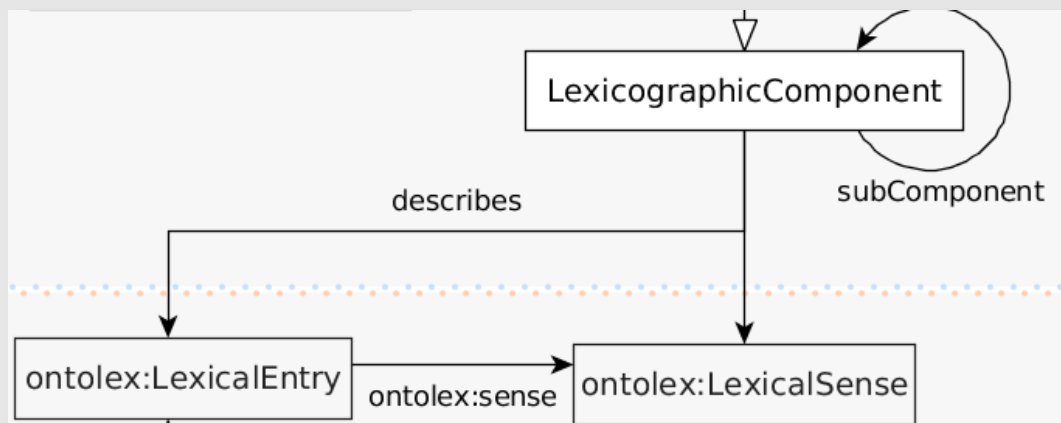
```
:animal_n a ontolex:LexicalEntry .
```

```
:animal_adj a ontolex:LexicalEntry .
```

# Ontolex lexicog module



# Ontolex lexicog module

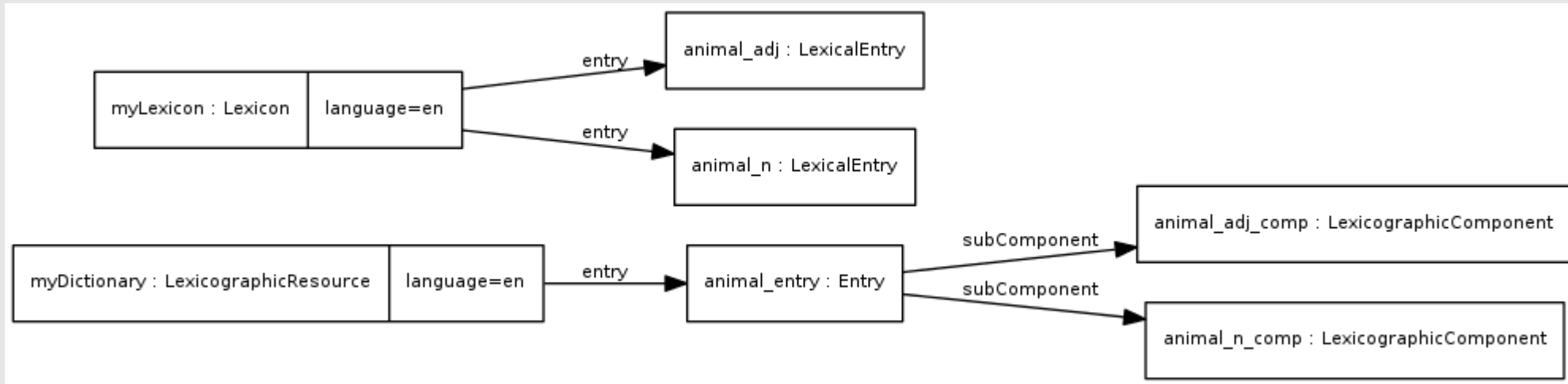
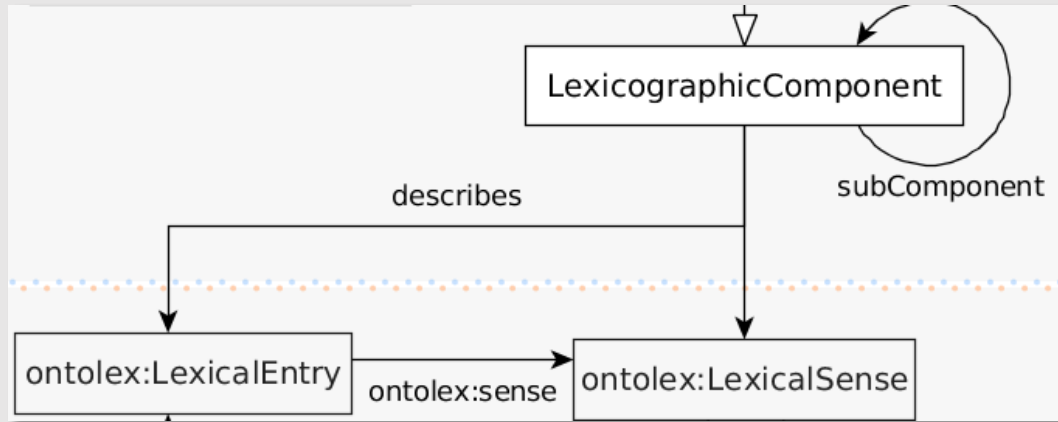


A **lexicographic component** is a structural element that represents the (sub-)structures of lexicographic articles providing information about entries, senses or sub-entries. If desired, lexicographic components can be arranged in a specific order and/or hierarchy.

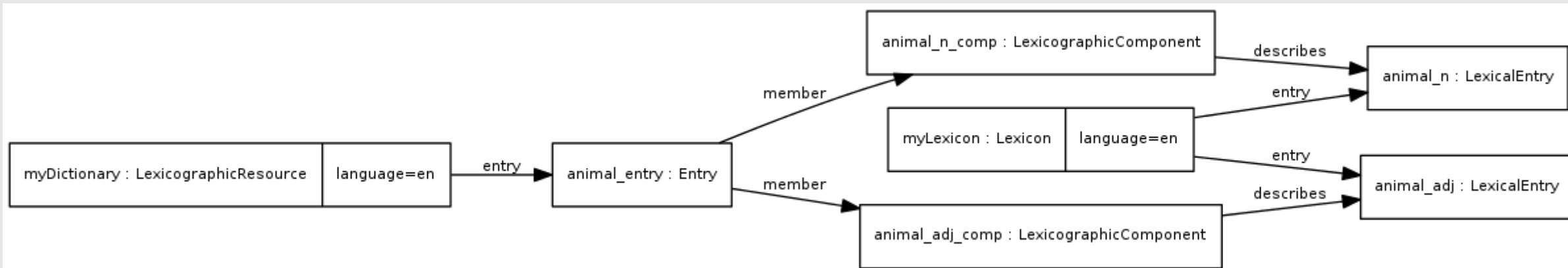
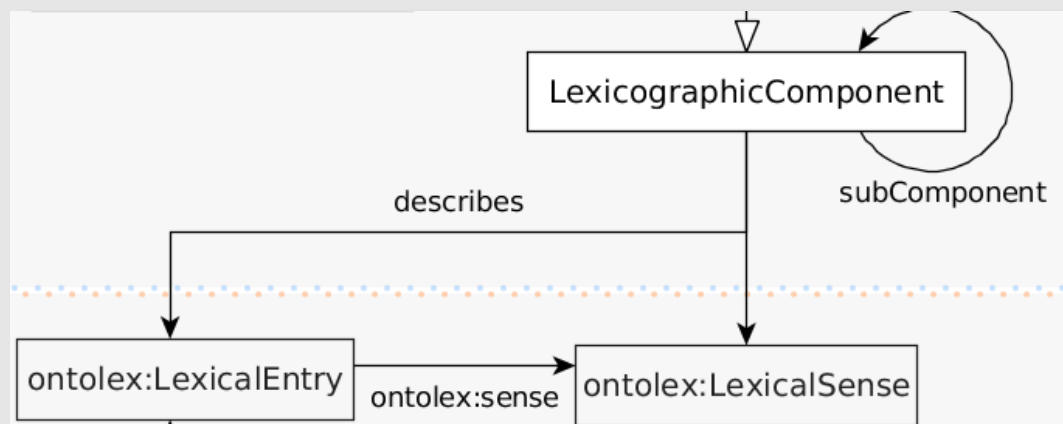
The property **describes** relates a lexicographic component to an element that represents the actual information provided by that component in the lexicographic resource. In most cases, this information will be lexical, and hence the object of the property will be an instance of **ontolex:LexicalEntry** or **ontolex:LexicalSense**.

The property **subComponent** encodes a hierarchical relation between two lexicographic components

# Ontolex lexicog module



# Ontolex lexicog module





# Conclusions

# Conclusions

- **Linked data** have proved to be useful for language resources in general and terminologies and dictionaries in particular
- Towards unified/linked **graph** dictionaries on the Web
- **Enrichment** of data by linking to other linked data resources
- What Will come next?
  - More dictionaries/terminologies/LRs converted into LD
  - New language resources built as LD from scratch
  - New generation of LD-aware NLP techniques

# Thanks!

**Jorge Gracia del Río**

**jogracia@unizar.es**  
**<http://jogracia.url.ph/web/>**



**Departamento de  
Informática e Ingeniería  
de Sistemas  
Universidad Zaragoza**