



“Language Resources and Linked Data: foundations”

Jorge Gracia, Asunción Gómez-Pérez

Ontology Engineering Group (OEG)

Universidad Politécnica de Madrid (UPM)

jgracia@fi.upm.es

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



NUI Galway
OÉ Gaillimh

UNIVERSITÄT LEIPZIG

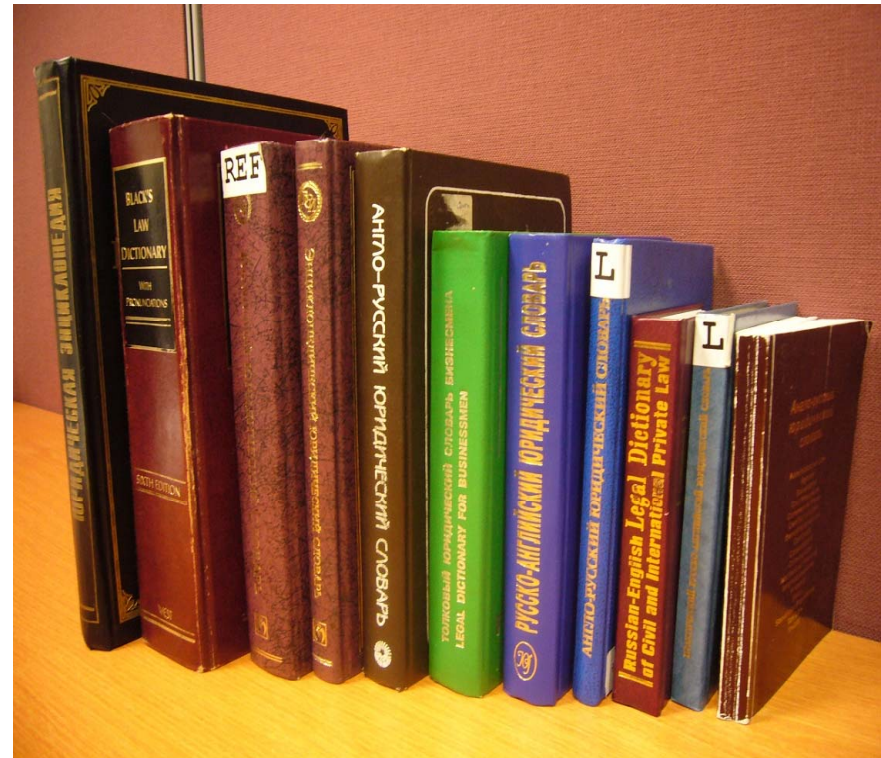


Language Resources on the Web

The current ecosystem

(Electronic) Language Resources

- corpora
- lexica
- dictionaries
- encyclopedic resources
- wordnets
- etc.



Currently...

- Proprietary and heterogeneous formats
- Non-standard APIs
- Disconnected from other resources (silos of LRs)
- Different access levels (from “write me an email” to web services)
- Several LRs repositories with different metadata and schemas

Therefore... discovery and reuse of LRs in third party applications is hard, manual and time consuming





<http://es.wiktionary.org>

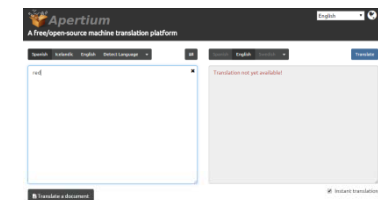
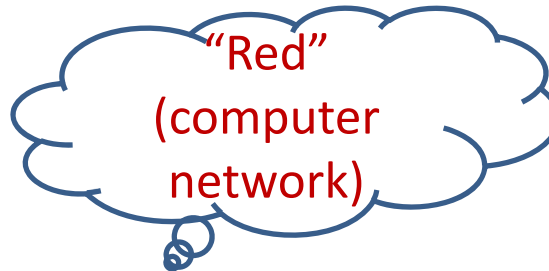


<http://rae.es>



Wikilengua del español

<http://www.wikilengua.org/index.php/Terminesp:red>



<http://www.apertium.org>



WIKIPEDIA

<http://es.wikipedia.org>





REAL ACADEMIA ESPAÑOLA

La institución Obras académicas Biblioteca y Archivo Consultas lingüísticas

Inicio » Recursos » Diccionarios » Diccionario de la lengua española

Diccionario de la lengua española

El *Diccionario de la lengua española (DRAE)* es la obra de referencia de la Academia. La edición actual —la 22.ª, publicada en 2001— incluye más de 88 000 entradas.

á é í ó ú ü ñ

■ Ayuda

red.

(Del lat. *rete*).

1. f. Aparejo hecho con hilos, cuerdas o alambres trabados en forma de mallas, y convenientemente dispuesto para pescar, cazar, cercar, sujetar, etc.
2. f. Labor o tejido de mallas.
3. f. redecilla (ll prenda de malla para el pelo).
4. f. Lugar donde se vende pan u otras cosas que se dan por entre verjas.





<http://es.wiktionary.org>



red

- **Pronunciación:** [red] (AFI)
- **Etimología:** del latín *rete*

Sustantivo femenino [editar]

Singular	Plural
red	redes

1

Malla cosa que fue creada para unir o sostener algo

2

Sistema interconectado

- **Ejemplos:** (...)

Véase también [editar]

- Wikipedia tiene un artículo sobre **red**.

Traducción [editar]

- Alemán: **Netz** ^(de)



[1] malla Raschel



[2] la red en 2005



<http://es.wiktionary.org>



Terminesp:red

Wikilengua

Entrada de Terminesp

Este artículo es una entrada de la base terminológica [Terminesp](#). No se puede modificar, pero se pueden hacer comentarios y puede servir de base para un artículo en la *Wikilengua*.

Búsqueda en Terminesp

Norma UNE

[Norma UNE \(Terminesp\)::UNE 21302-131](#)

Título de la norma

Vocabulario electrotécnico: parte 131. Teoría de circuitos

Término en español:

red ([artículo](#))

Término en alemán:

Netzwerk (in der Netzwerktopologie)

Término en inglés:

network

Término en francés:

réseau

Definición

En topología de redes, conjunto de elementos ideales de circuito y de sus interconexiones, considerado como un todo.

<http://es.wiktionary.org>





Apertium

A free/open-source machine translation platform

English


Spanish Icelandic English Detect Language ⇌

Spanish English Swedish ▼

Translate

red

Network

<http://www.apertium.org>

<http://es.wiktionary.org>



Red de computadoras

(Redirigido desde «Redes de computadoras»)

Una **red de computadoras**, también llamada **red de ordenadores**, **red de comunicaciones de datos** o **red informática**, es un conjunto de **equipos informáticos** y **software** conectados entre sí por medio de **dispositivos físicos** que envían y reciben **impulsos eléctricos**, **ondas electromagnéticas** o cualquier otro medio para el transporte de **datos**, con la finalidad de compartir información, recursos y ofrecer **servicios**.¹

Como en todo proceso de **comunicación** se requiere de un **emisor**, un **mensaje**, un **medio** y un **receptor**. La finalidad principal para la creación de una red de computadoras es compartir los recursos y la información en la distancia, asegurar la **confiabilidad** y la disponibilidad de la información, aumentar la **velocidad de transmisión de los datos** y reducir el costo general de estas acciones.² Un ejemplo es **Internet**, la cual es una gran red de millones de computadoras ubicadas en distintos puntos del planeta interconectadas básicamente para compartir información y recursos.

La estructura y el modo de funcionamiento de las redes informáticas actuales están definidos en varios **estándares**, siendo el más importante y extendido de todos ellos el modelo **TCP/IP** basado en el modelo de referencia **OSI**. Este último, estructura cada red en siete capas con funciones concretas pero relacionadas entre sí; en TCP/IP se reducen a cuatro capas. Existen multitud de protocolos repartidos por cada capa, los cuales también están regidos por sus respectivos estándares.³

Índice [ocultar]

- 1 Historia
- 2 Descripción básica
- 3 Componentes básicos de las redes
 - 3.1 Software
 - 3.2 Hardware
 - 3.2.1 Tarjeta de red
 - 3.3 Dispositivos de usuario final
 - 3.3.1 Servidores

<http://www.apertium.org>



<http://es.wikipedia.org>

Jorge Gracia, Asunción Gómez-Pérez

11

attribution: <http://commons.wikimedia.org/wiki/User:Gugere11>



"Red"

Pronunciation: [red]

Grammar category: sustantivo femenino

Singular: "red"

Plural: "redes"



"Red"

Etimologiy Del latin "rete"

Gender: "f"

Definition.: "Conjunto de ordenadores o de equipos informáticos conectados entre sí...."



Wikilengua del español

"Red"

Norm: UNE 21302-131

English: network

German: Netzwerk

**Complementary
but not connected**

"Red_de_computadores"

Category: redes informáticas

Image

"Red"

translations: "xarxa"(ca), "rede"(ga), ...

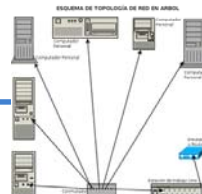


25/11/2014

Jorge Gracia, Asunción Gómez-Pérez

12

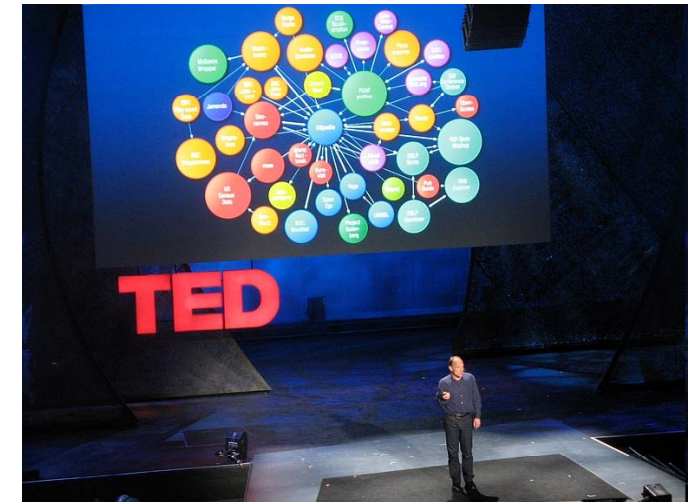
attribution: <http://commons.wikimedia.org/wiki/User:Gugerell>



Linked Data

Linked Data Principles

1. Use **URIs** to identify “things” on the Web
2. Use HTTP URIs so people can **look up** those identifiers and get something
3. When someone looks up a URI, provide useful **information** (in RDF)
4. Include **links** to other URIs, so that they can discover more things.



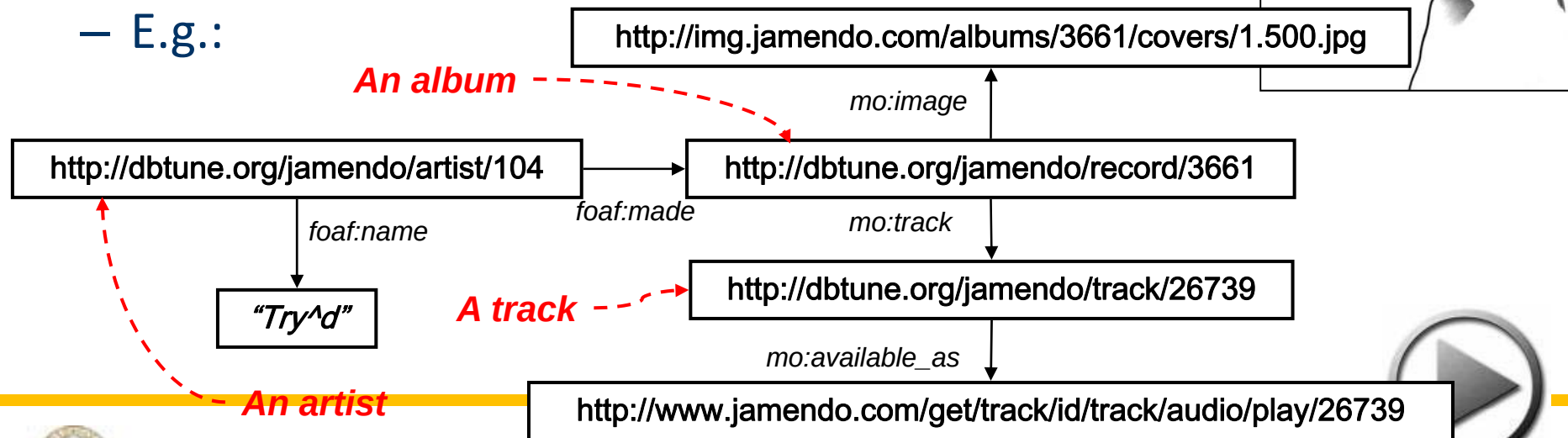
<http://www.w3.org/DesignIssues/LinkedData.html>

URIs (Uniform Resource Identifiers)

- Identify resources on the Web (both “informational” resources and “real-world” ones)

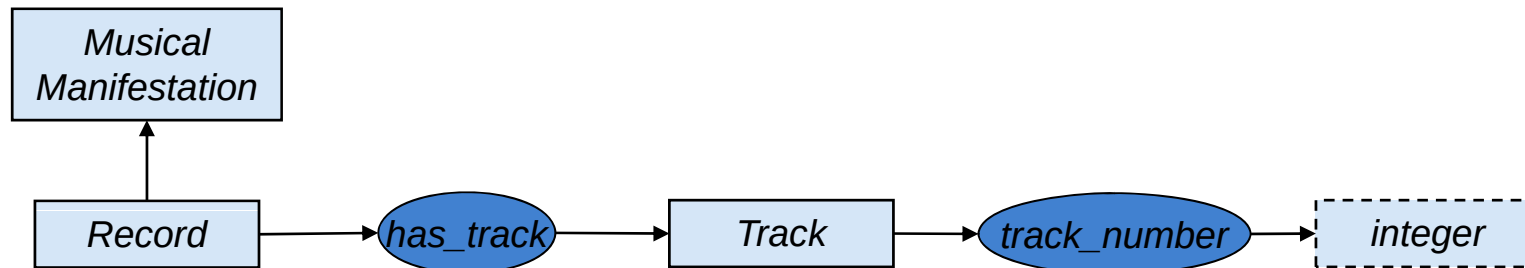
RDF (Resource Description Framework)

- To represent relations between resources
- *subject - predicate - object*
- E.g.:



Web Ontologies (OWL or RDF-S)

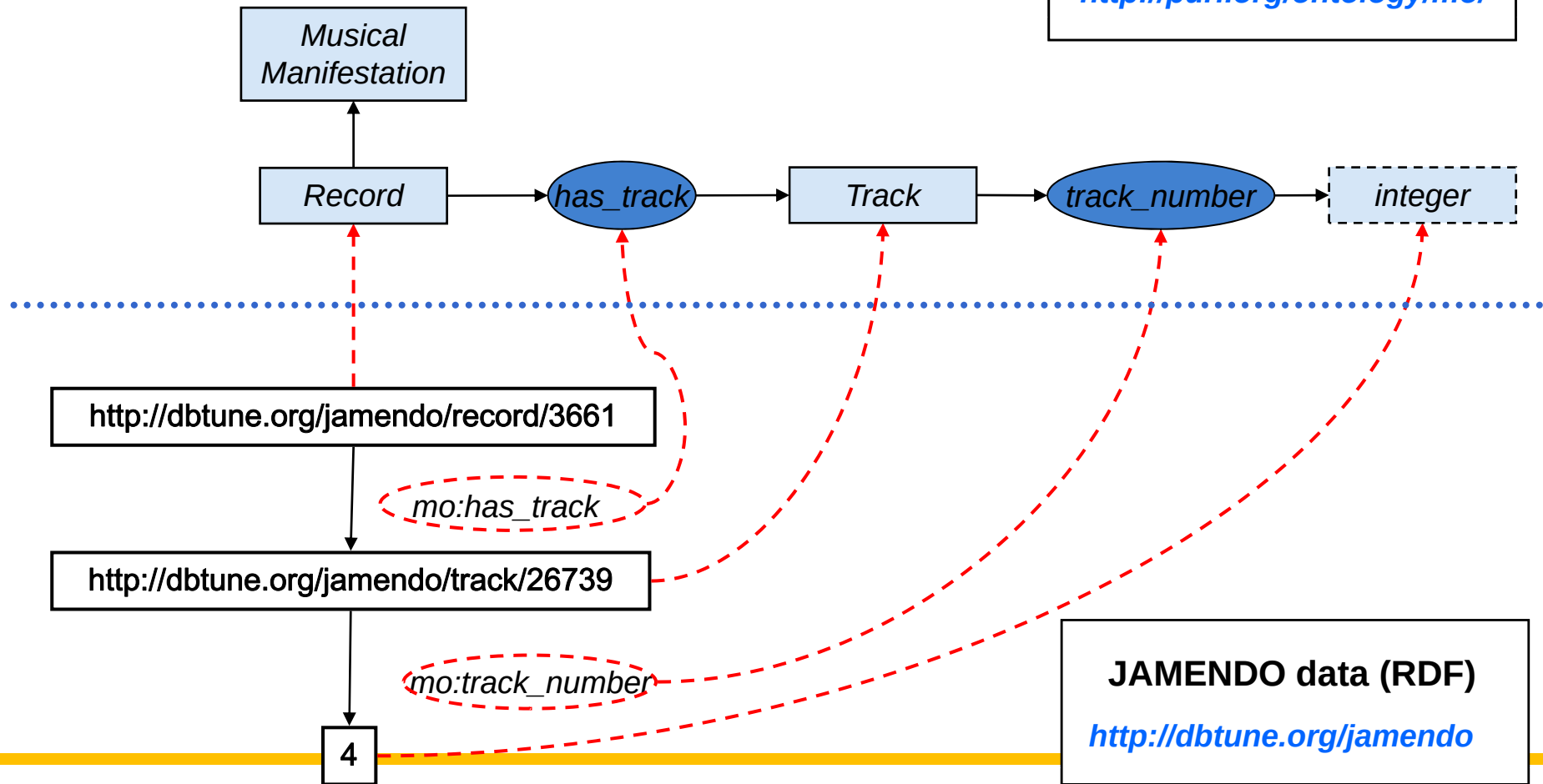
- Classes of objects
- Properties, or relations among them
- Individuals, or instances of these classes
- Allow use of inference rules
- E.g.: from MUSIC ontology <http://purl.org/ontology/mo/>



Putting things together

MUSIC ontology (OWL)

<http://purl.org/ontology/mo/>





Data accessible in a standardised way (SPARQL):

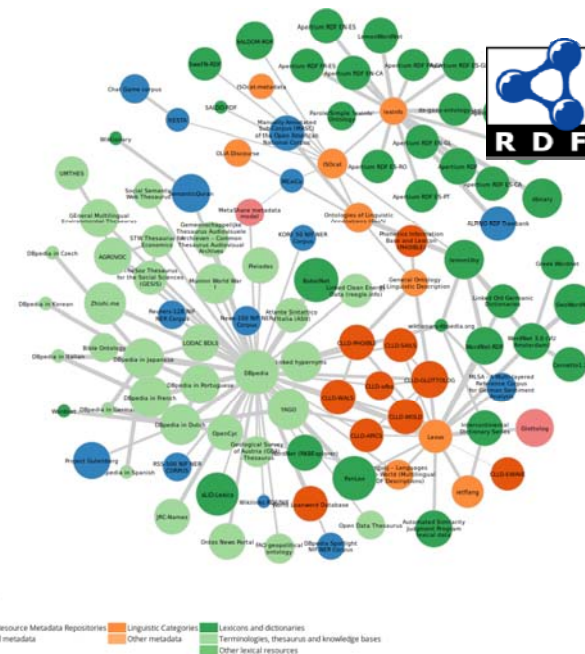
- *“Give me the title of the albums made by Tryad along with their publication date”*

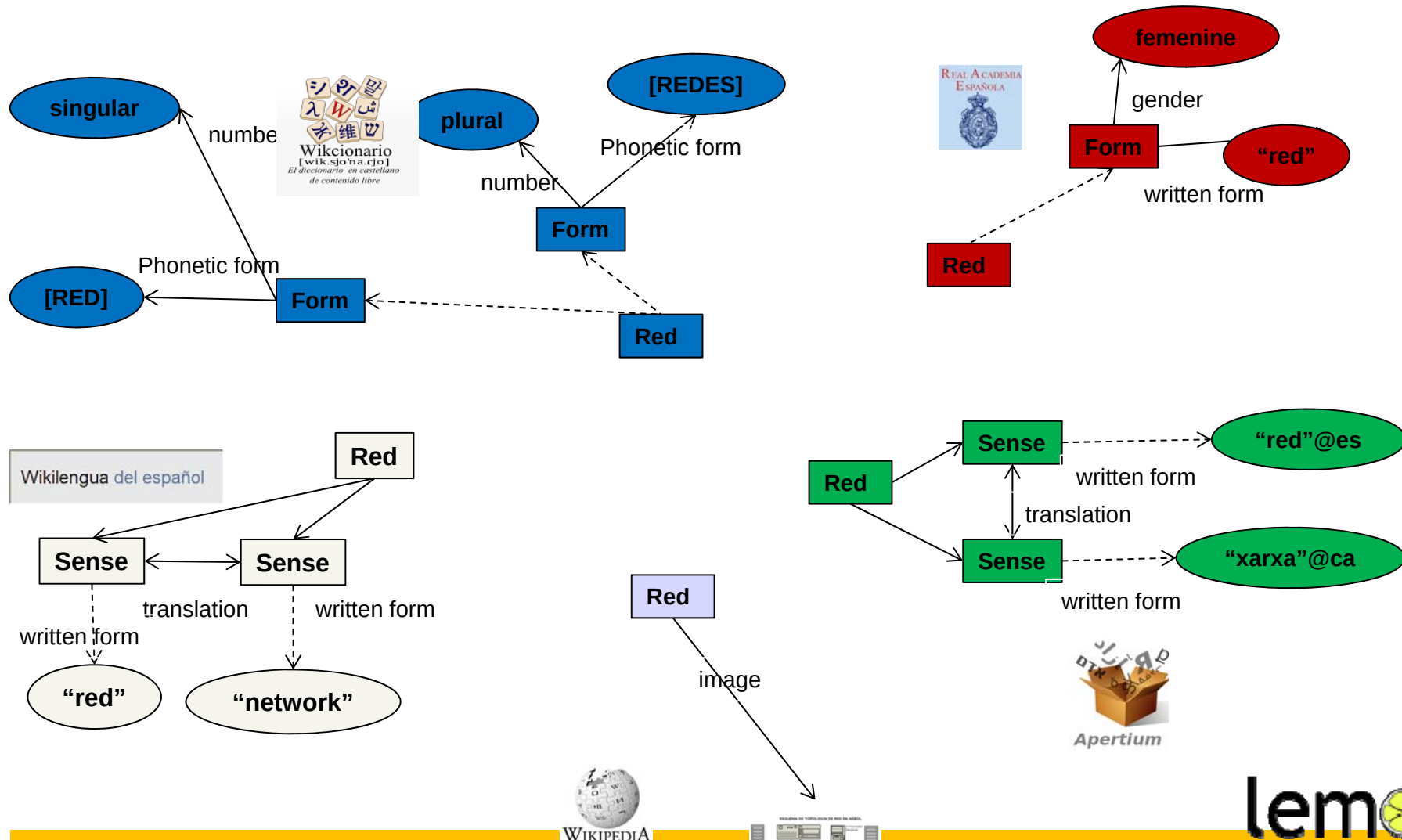
```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX mo: <http://purl.org/ontology/mo/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT DISTINCT ?title ?date
WHERE
{ ?artist rdf:type mo:MusicArtist;
  foaf:made ?album;
  foaf:name "t r y ^ d";
  ?album dc:title ?title;
  dc:date ?date }
```

Title	Date
“Listen”	29-12-2006
“Public Domain”	20-01-2006
“Tryad Demo (Public Domain)”	14-04-2005

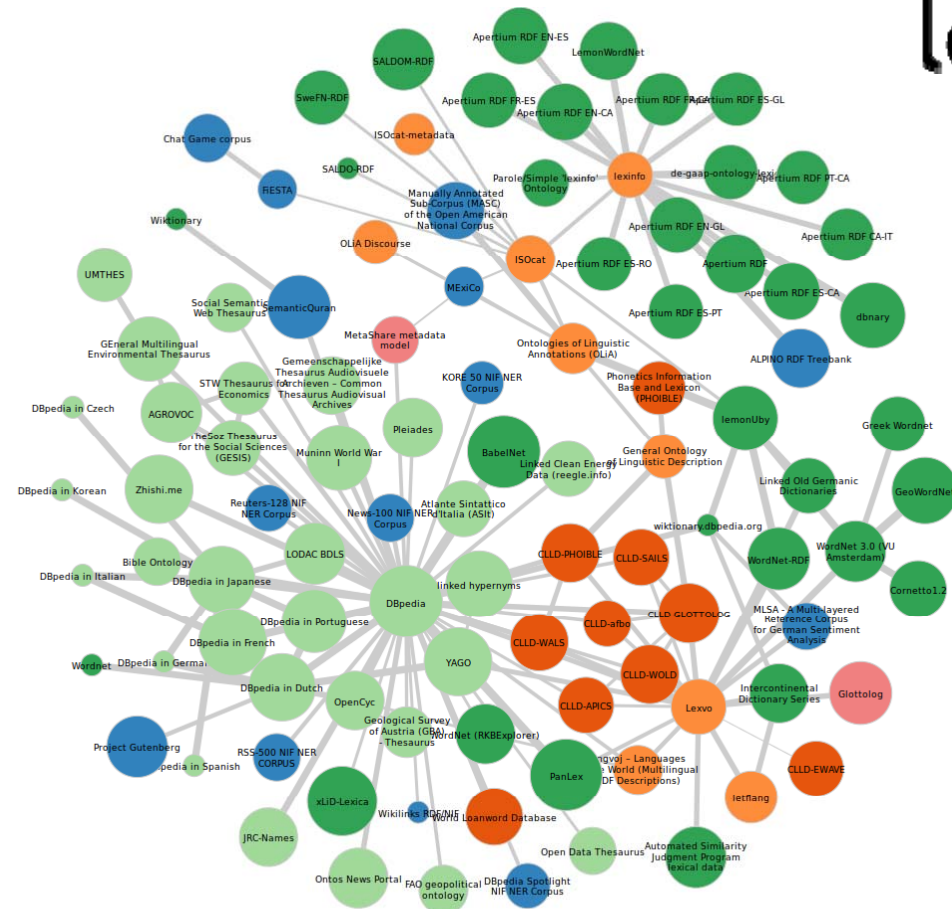
Linked Data for Language Resources

Linked Data allows Language Resources to interoperate





Linguistic Linked Data Cloud



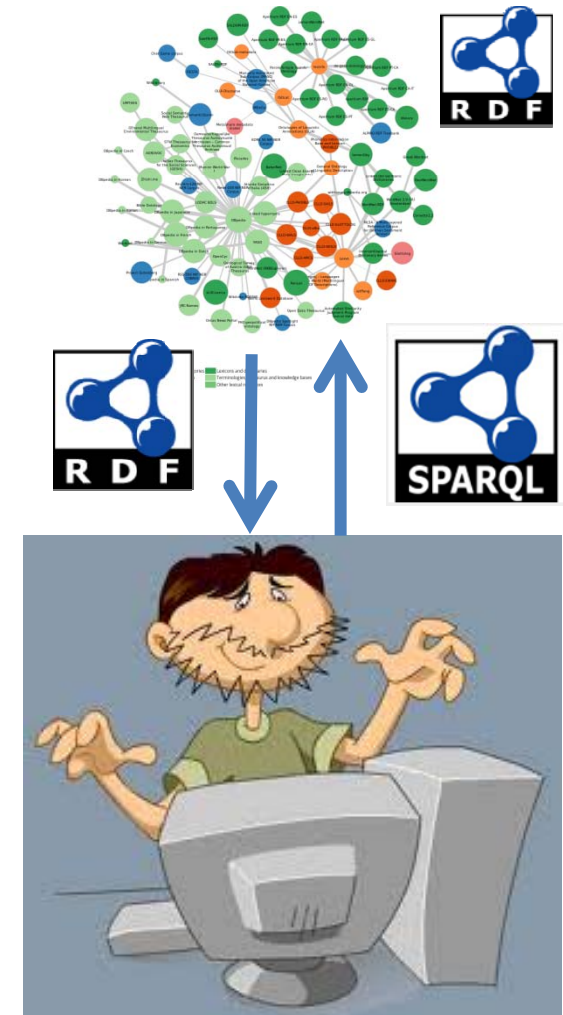
lemon

NIF

NLP Interchange Format



Linked Data
allows uniform
access to
Language
Resources and
Services



Linguistic Data accessible in a standardised way (SPARQL):

“List Language Resources by language and type of resource”

```
SELECT ?l COUNT(?l) ?type
WHERE { ?s ms:language ?l; a ?type.
FILTER((REGEX(STR(?type), "pur|")))}
GROUP BY ?l ?type
```

l	[.1]	type
eng	I 110	ms:LexicalConceptualResource
spa	I 89	ms:LexicalConceptualResource
EN	I 49	ms:LexicalConceptualResource
fre/fra	I 39	ms:LexicalConceptualResource
en	I 36	ms:CorpusText
ger/deu	I 25	ms:LexicalConceptualResource
LV	I 24	ms:LexicalConceptualResource
DE	I 23	ms:LexicalConceptualResource
RU	I 23	ms:LexicalConceptualResource
fi	I 21	ms:CorpusText
FR	I 19	ms:LexicalConceptualResource
fi	I 17	ms:LexicalConceptualResource
IS	I 17	ms:LexicalConceptualResource



**sample data
(855 records)**

Acknowledgements: Marta Villegas and Nuria Bel

“Give me available translations of the Spanish term ‘red’ along with their part of speech”

```
SELECT DISTINCT ?translated_written_rep ?pos
WHERE {
  ?form_a lemon:writtenRep "libro"@es .
  ?lex_entry_a lemon:lexicalForm ?form_a .
  ?sense_a lemon:isSenseOf ?lex_entry_a .
  ?trans tr:translationSense ?sense_a .
  ... .
```



translated_written_rep	pos
"filet" @fr	<http://www.lexinfo.net/ontology/2.0/lexinfo#noun>
"réseau" @fr	<http://www.lexinfo.net/ontology/2.0/lexinfo#noun>
"network" @en	<http://www.lexinfo.net/ontology/2.0/lexinfo#noun>
"net" @en	<http://www.lexinfo.net/ontology/2.0/lexinfo#noun>

BENEFITS of LRs as Linked Data

1. **Short term** (it's happening now!) [1]

- **Simpler** and formal model (respect to XSD counterpart).
- **Merging** of different schemas is straightforward
- Resources are explicitly **linked** among themselves and to external data (with stable, resolvable identifiers)
- Data **exposed** in a standardized way (SPARQL)



[1] "Metadata as Linked Open Data: mapping disparate XML metadata registries into one RDF/OWL registry." Marta Villegas, Maite Melero and Núria Bel. In Proc. of LREC'2014, Reykjavik, Iceland, May 2014

2. Medium term

- Improved **discovery** of dataset and services
- Aggregation and **integration** of linguistic resources

3. Long term

- Composition and **chaining** of LRs

Examples of Language Resources as Linked Data

<http://wordnet-rdf.princeton.edu/wn31/108952341-n.rdf>

WordNet
A lexical database for English



City of Light, French capital, Paris, capital of France

What is WordNet?
People
News
Use WordNet online
WordNet RDF

Gloss	the capital and largest city of France; and international center of culture and commerce	
Instance	108709407-n	
Hypernym	Gloss	the capital city of a nation
Lexical Domain	noun.location	

BabelNet 2.0
A very large multilingual encyclopedic dictionary and semantic network

[search](#) · [explore](#) · [publications](#) · [download](#)

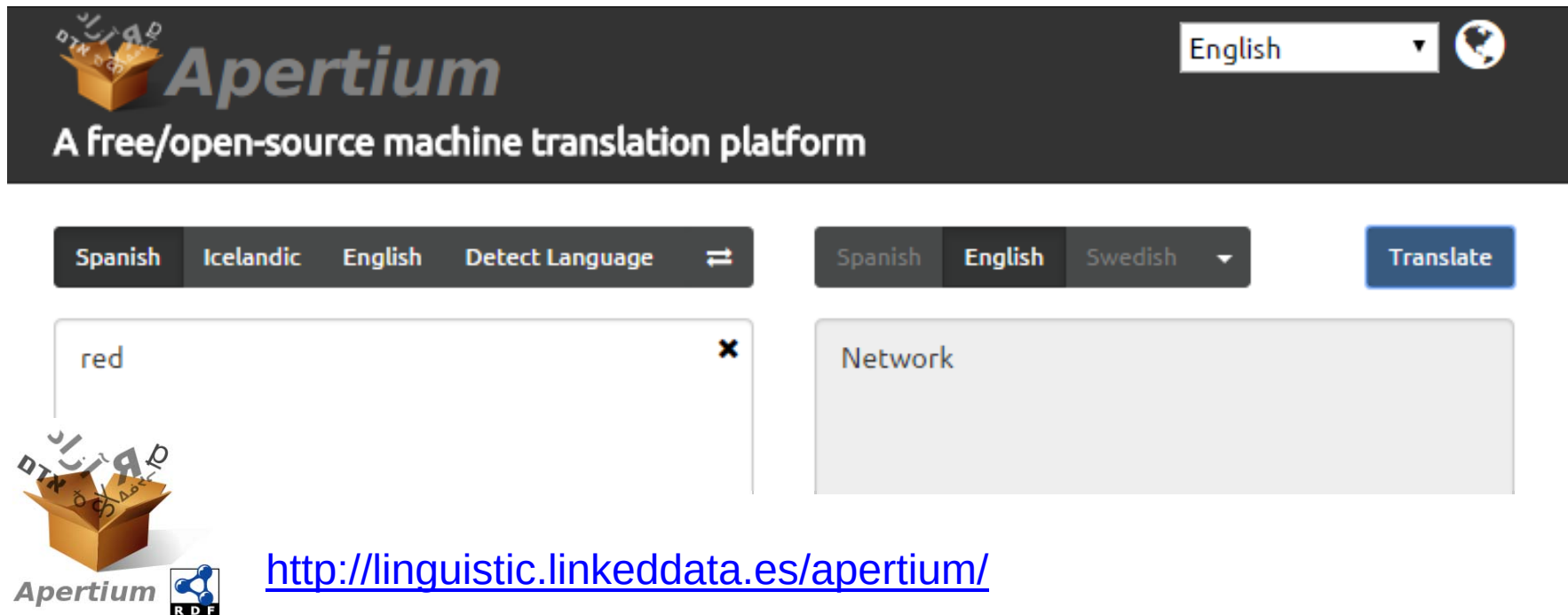
BabelNet 2.0.1: General statistics

Number of languages:	60
Total number of Babel synsets:	9,348,287
Total number of Babel senses:	50,282,923
Total number of concepts:	3,684,512
Total number of Named Entities:	5,663,775
Total number of lexico-semantic relations:	262,687,848
Total number of glosses (textual definitions):	17,961,157
Total number of images:	7,764,270

RDF dumps for the Linguistic Linked Open Data cloud (for Semantic Web fans)

-  BabelNet in Lemon RDF format (IRIs) version 2.0 (March 2014) **NEW!**
-  BabelNet in Lemon RDF format (URIs) version 2.0 (March 2014) **NEW!**
-  SPARQL endpoint for version 2.0 (March 2014) **NEW!**

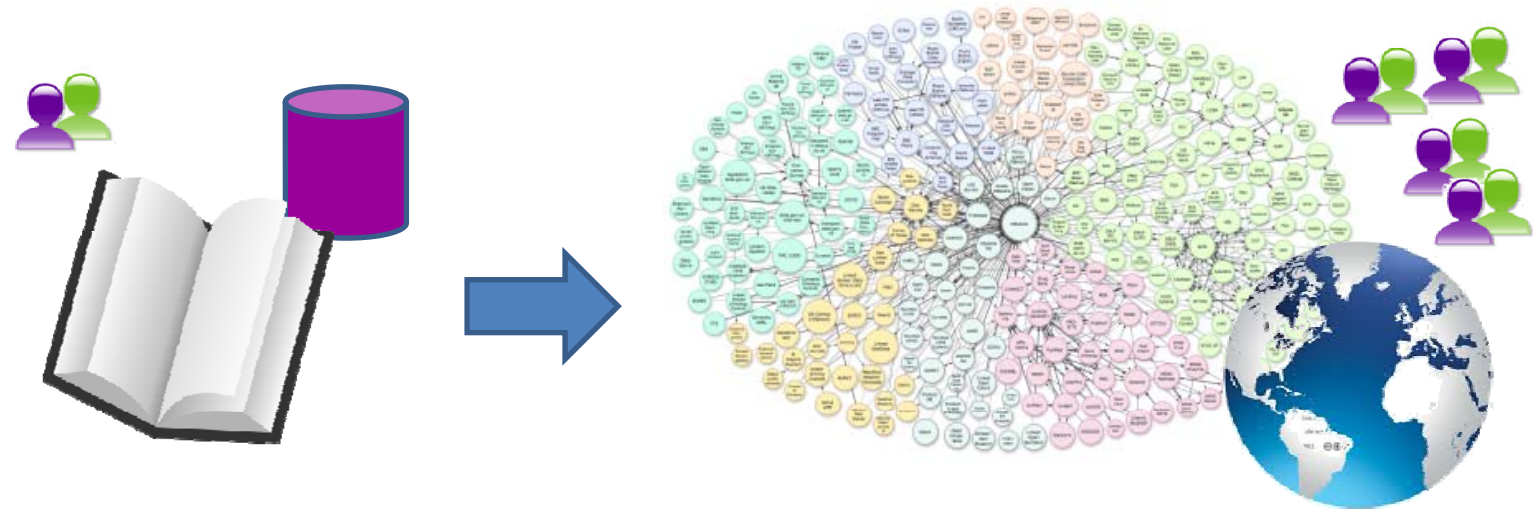
Apertium [<http://www.apertium.org>] is an open source platform for automatic machine translations. Its (>30) bilingual dictionaries are being converted into RDF



The screenshot shows the Apertium web interface. At the top, there's a dark header with the Apertium logo (a box with letters) and the text "Apertium A free/open-source machine translation platform". To the right of the header is a language selector dropdown set to "English" and a globe icon. Below the header, there are two rows of language buttons. The first row has "Spanish", "Icelandic", "English", and "Detect Language" with a double-headed arrow. The second row has "Spanish", "English", and "Swedish" with a dropdown arrow. To the right of these buttons is a blue "Translate" button. Below the buttons, there are two text input fields. The left field contains the word "red" and has a small 'x' icon to its right. The right field contains the word "Network". At the bottom left of the screenshot, there is a small Apertium logo and an RDF icon. To the right of these icons is the URL <http://linguistic.linkeddata.es/apertium/>.

Goals of the tutorial

GOAL: to allow language resources to expose their **Data as Linked Data** on the **Web** for their consumption by semantic enabled applications in a direct manner



Motivating questions

1. How to represent rich **lexical information** (beyond rdfs:label) and associate it to ontologies and linked data?
2. How to **generate** multilingual linguistic linked data from data silos?

3. How to represent and publish multilingual texts, annotations and **corpora** as linked data?
4. How to perform **word sense disambiguation** and **entity linking** of multilingual linked data?



MORNING

- 9:00 - 9:10 Welcome and introduction
- 9:10 - 9:30 **Session 1: foundations**
- 9:30 – 10:00 **Session 2.** Modelling lexical resources on the Web of Data: the **lemon** model (intro)
- 10:30 - 11:00 *Coffee break*
- 11:00 - 11:45 **Session 2.** Modelling lexical resources on the Web of Data: the **lemon** model (hands-on)
- 11:45 - 12:30 **Session 3. Methodology** and tools for Multilingual Linguistic Linked Data generation (intro + hands-on)
- 12:30 - 13:30 *Lunch break*
-

AFTERNOON

13:30 - 14:00 **Session 3. Methodology** and tools for Multilingual Linguistic Linked Data generation (hands-on)

14:00 – 15:00 **Session 4.** Integrating NLP with Linked Data and RDF: the **NIF** format (intro + hands-on)

15:00 - 15:30 *Coffee break*

15:30 - 15:45 **Session 4.** Integrating NLP with Linked Data and RDF: the **NIF** format (hands-on)

15:45 - 17:00 **Session 5.** Multilingual Word Sense Disambiguation and Entity Linking on the Web based on **BabelNet** (intro + hands-on)

Your laptop...



and our pendrive

- Tutorial's web page http://www.lider-project.eu/ekaw14_LRLDTutorial
- LIDER project <http://www.lider-project.eu/>
- lemon model <http://lemon-model.net/>
- NIF Interchange Format <http://persistence.uni-leipzig.org/nlp2rdf/>
- BabelNet <http://babelnet.org/>
- W3C Best Practises for Multilingual Linked Open Data community group <http://www.w3.org/community/bpmlod/>
- W3C Ontology Lexica community group <http://www.w3.org/community/ontolex/>



www.lider-project.eu



Join the community
“Linked Data for Language Resources”

www.w3c.org/community/ld4lt



twitter.com/multilingweb

Hashtag: #LiderEU

