



Towards a Linked Open Data Cloud of Language Resources in the Legal Domain

Patricia Martín Chozas

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

OEG weekly meeting
May 31st, 2018

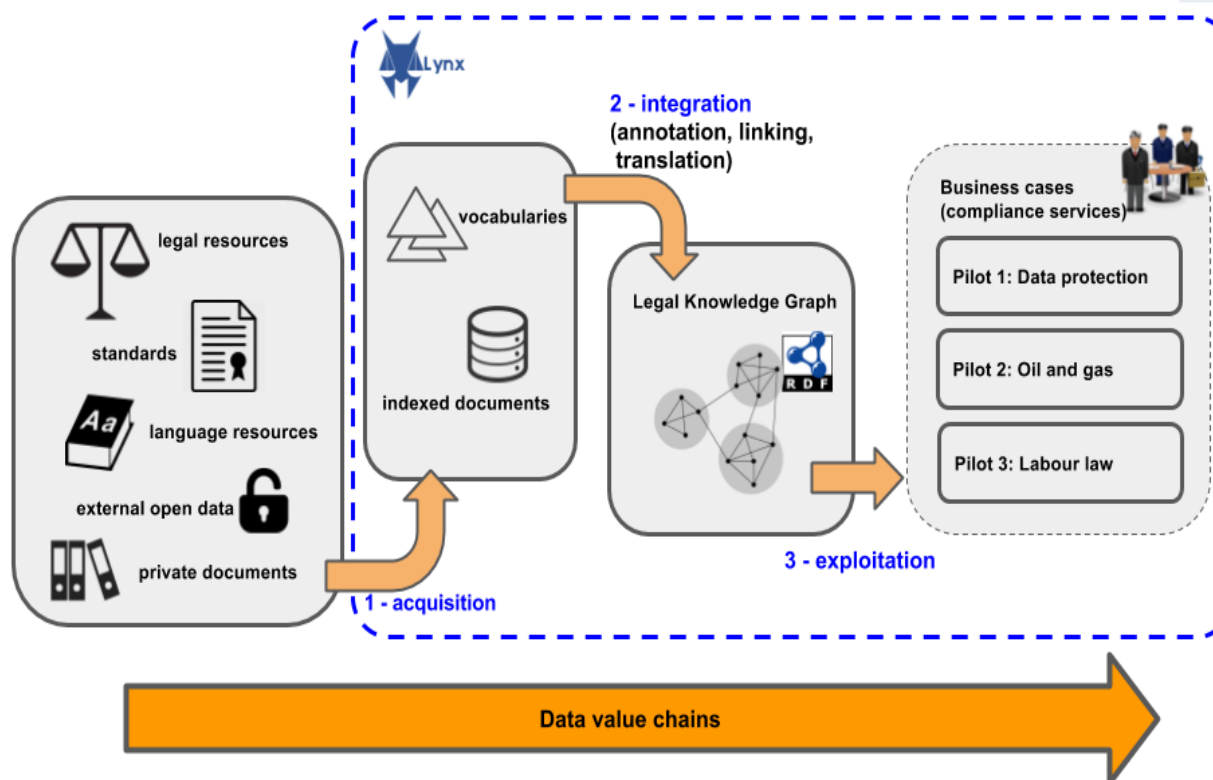
- Lynx Project
 - Goal
 - Needs
- Role in the Project
- Methodology
 - Stage 1: Identification and creation of resources
 - Stage 2: Conversion into RDF
 - Stage 3: Linking process
- Future work

- **Goal:** construction of a Legal Knowledge Graph (LKG) enabling the provision of compliance-related services
- **Needs:**
 - Technical architecture
 - Data acquisition and management
 - Common services infrastructure
 - Content curation services
 - Project management
 - [...]

■ Data acquisition:

- Legal documents
- Linguistic resources

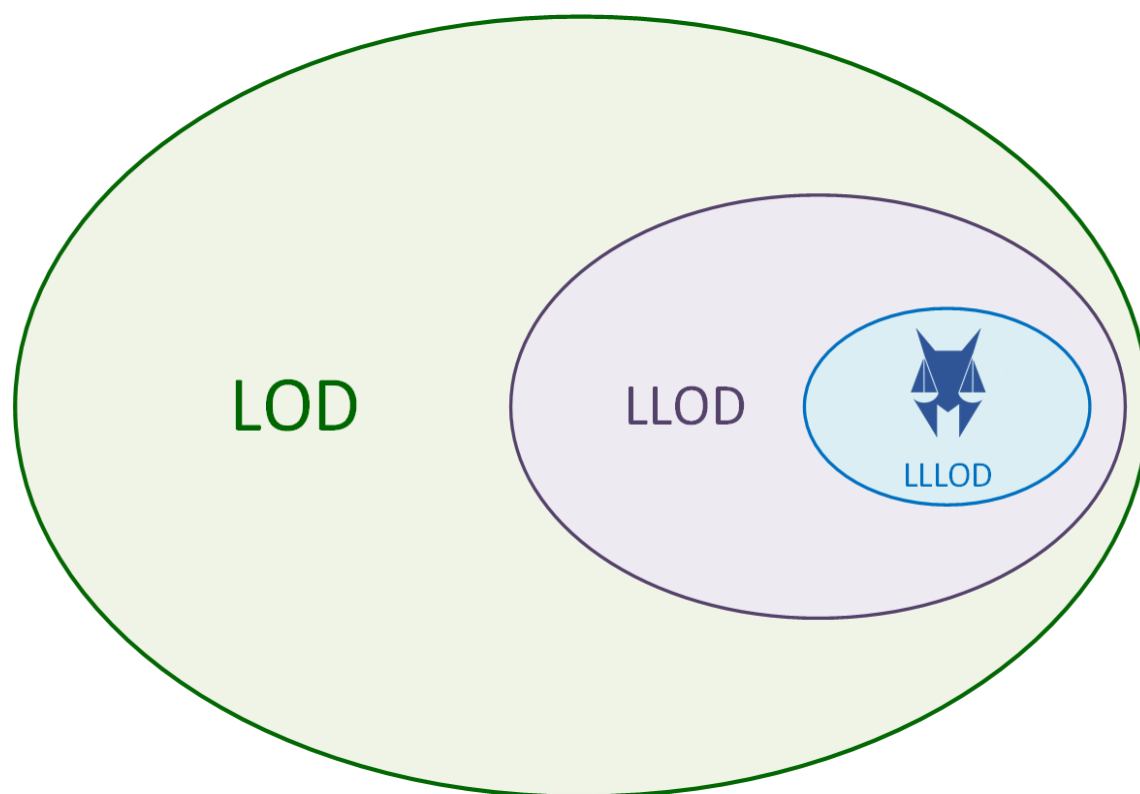
Domain	Language
Labour Law (LL)	English (EN)
Data Protection (DP)	German (DE)
Industrial Standards (IS)	Italian (IT)
	Spanish (ES)



Data Value Chain in Lynx, from Lynx Document of Work

■ Linguistic resources:

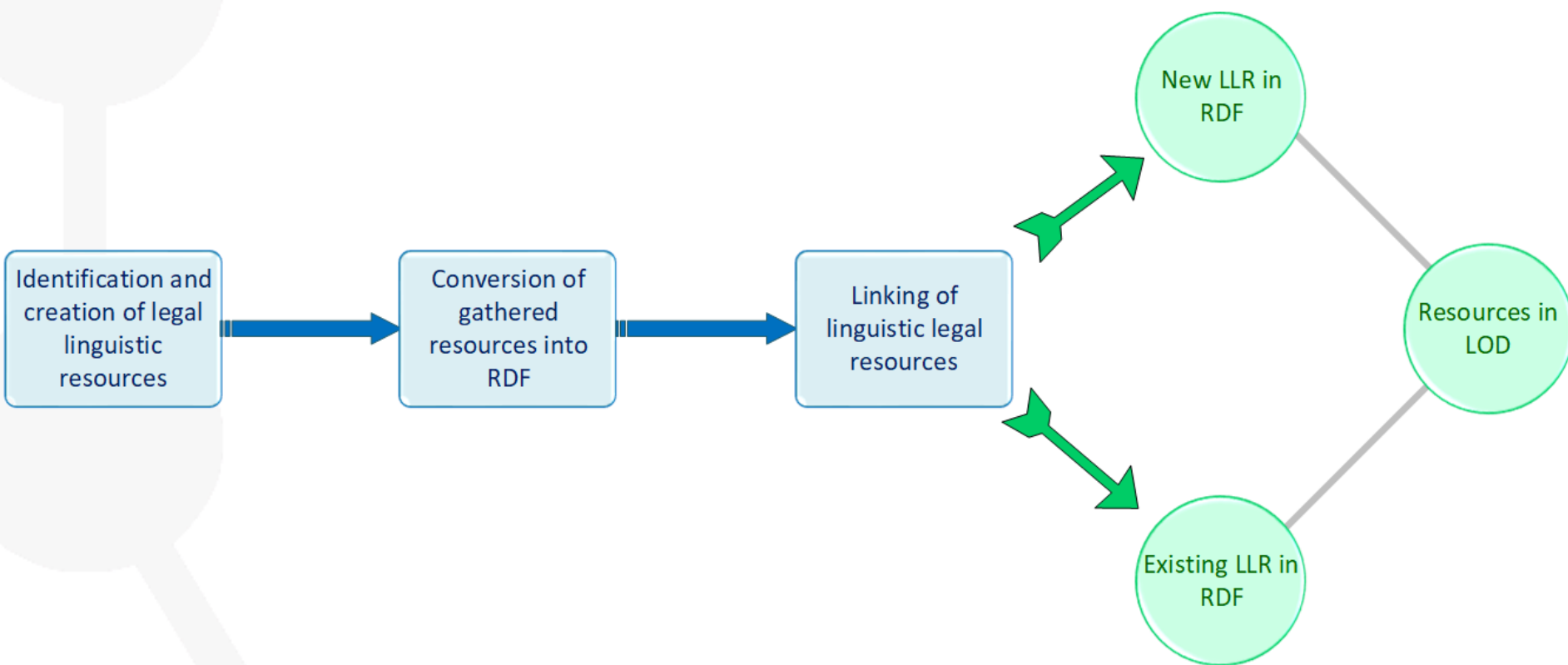
- Glossaries
- Databases
- Dictionaries
- Thesauri
- Lexica
- Terminologies



LOD: *Linked Open Data*

LLOD: *Linguistic Linked Open Data*

LLLOD: *Linguistic Legal Linked Open Data*



Linguistic Legal Linked
Open Data Cloud

■ Search strategies

- General web search
- Resources cited in literature
- Data portals and repositories

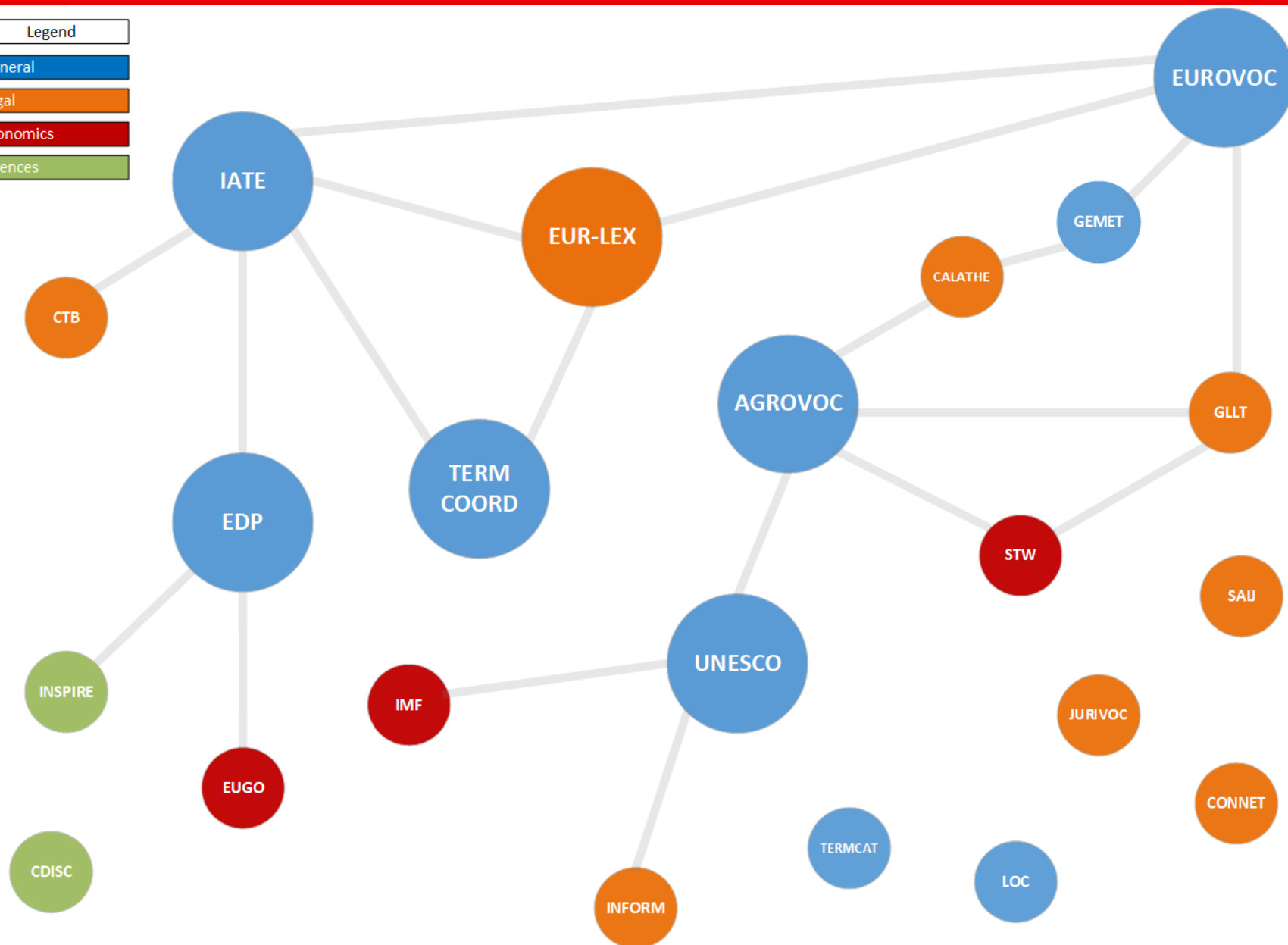
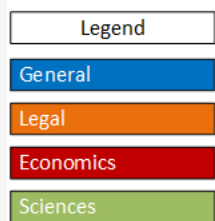
■ First results

- 22 datasets
- General and legal domain
- Various formats
- Multiple languages
- Different typology

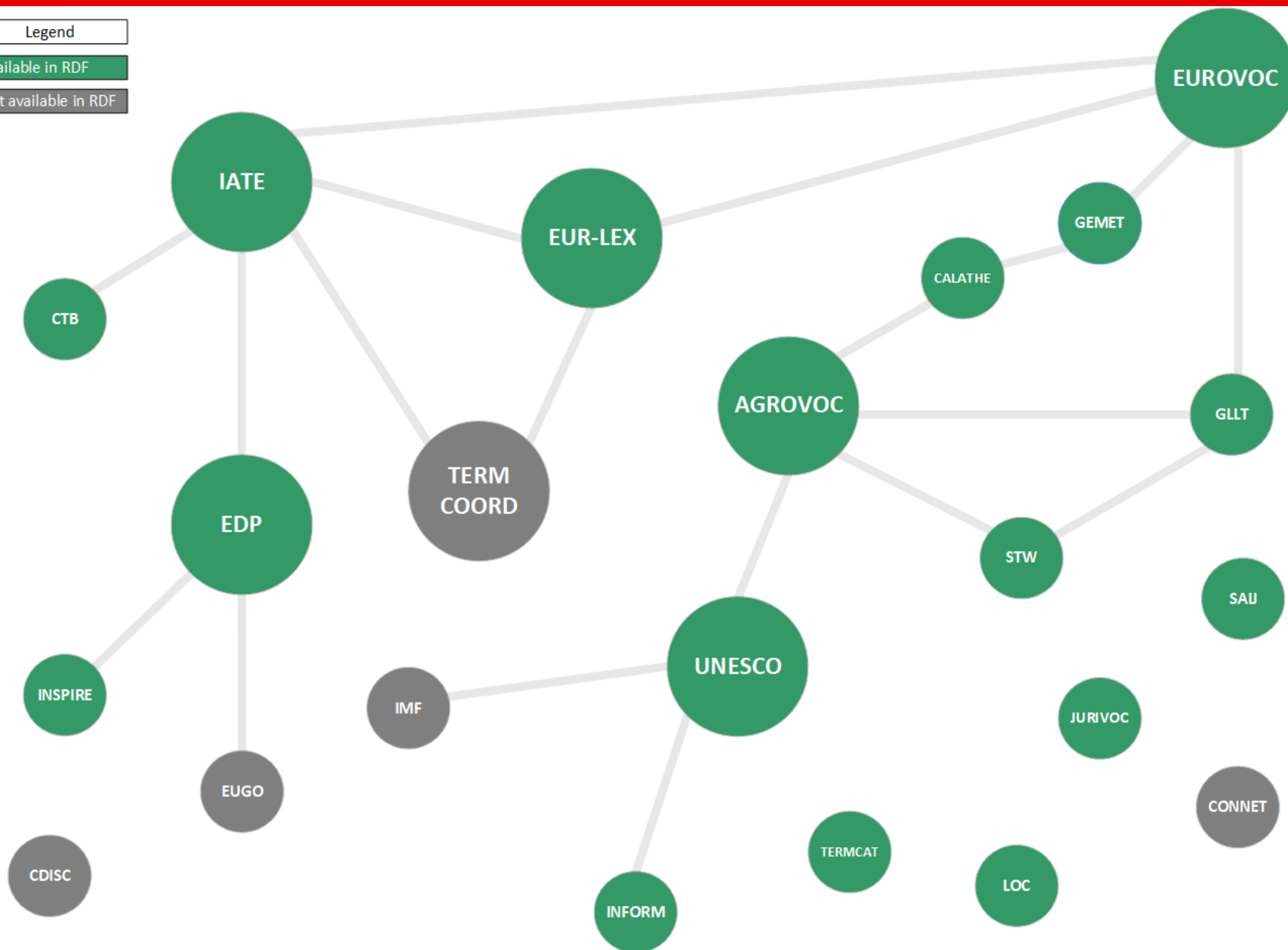
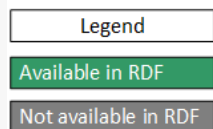
Stage 1.1: Identification of reusable resources

ID	Name	Description	Language
iate	IATE	EU terminological database.	EU languages
eurovoc	Eurovoc	EU multilingual thesaurus.	EU languages
eur-lex	EUR-Lex	EU legal corpora portal.	EU languages
connecticut-legal-glossary	Connecticut Legal Glossary	Bilingual legal glossary.	en, es
unesco-thesaurus	UNESCO Thesaurus	Multilingual multidisciplinary thesaurus.	en, es, fr, ru
library-of-congress	Library of Congress	Legal corpora portal.	en
imf	International Monetary Fund	Economic multilingual terminology.	en, de, es
eugo-glossary	EUGO Glossary	Business monolingual dictionary.	es
cdisc-glossary	CDISC Glossary	Clinical monolingual glossary.	en
stw	STW Thesaurus for Economics	Economic monolingual thesaurus.	en
edp	European Data Portal	EU datasets.	EU languages
inspire	INSPIRE Glossary (EU)	General terms and definitions in English.	en
saij	SAIJ Thesaurus	Controlled list of legal terms.	es
calathe	CaLaThe	Cadastral vocabulary.	en
Gemet	GEMET	General multilingual thesauri.	en, de, es, it
informea	InforMEA Glossary (UNESCO)	Monolingual glossary on environmental law.	en
copyright-termbank	Copyright Termbank	Multi-lingual termbank of copyright-related terms.	en, es, fr, pt
gllt	German labour law thesaurus	Thesaurus with labour law terms.	de
jurivoc	Jurivoc	Juridical terms from Switzerland.	de, it, fr
termcat	Termcat	Terms from several fields including law.	ca, en, es, de, fr, it
termcoord	Termcoord	Glossaries from EU institutions and bodies.	EU languages
agrovoc	Agrovoc	Controlled general vocabulary.	29 languages

Stage 1.1: Identification of reusable resources



Stage 1.1: Identification of reusable resources



■ Automatic term extraction from legal corpora

- Labour Law Corpora
- Data protection Corpora
- Industrial Standards Corpora

■ Term extraction tool selection

- Evaluation of 9 TE tools
- Several tests per domain and language
- Evaluation factors: format, input, access, language filter, output, compounds, stop words and additional services.

Stage 1.2: Creation of new resources

	URL	Format	Document format (input)	Access	Language filter	Results (output)	Comp. terms	Stop words	Additional services
TE Translated.net Labs	https://labs.translated.net/terminology-extraction/	Online	Plain text on website	Free	EN, IT, FR	HTML	Yes	No	Link terms with Google
VocabGrabber	https://www.visualthesaurus.com/vocabgrabber/	Online	Plain text on website	Under payment	No filter	HTML (if free use)	No	Yes	Graphical representation
TermoStat Web	http://termostat.ling.umontreal.ca/index.php	Online	Plain text from PC	Free	EN, ES, FR, IT, PT	HTML, TXT	Yes	No	-
Sketch Engine	https://www.sketchengine.co.uk/	Online	Multiformat from PC, URL	Under payment	Multilingual	HTML, TBX, CSV	Yes	No	Corpus creation, training and search
Five Filters	http://fivefilters.org/term-extraction/	Online	Plain text on website	Free	No filter	HTML, JSON, XML, TEXT, PHP	Yes	Yes	-
Termine	http://www.nactem.ac.uk/software/termine	Online	Plain text or PDF/TXT from PC	Free	No filter	HTML, TXT	Yes	No	-
Pootle	https://pootle.translatehouse.org/	Download	-	-	-	-	-	-	Python version not compatible for the moment
TBXTools	https://sourceforge.net/projects/tbxtools/	Download	TXT	Free (with permissions for the Wiki)	EN	TXT	Yes	No	It needs files to train and compare
TermSuite (TreeTagger)	http://termsuite.github.io/	Download	TXT	Free	Multilingual	TXT	No	Yes	Word Analysis (Grammatical Category) (not really a term extractor)

- Language filter
- Single and compound terms
- No stop words
- Creation of corpus for extraction
- Different corpora depending on domain
- Several input formats including PDF
- CSV and TBX output



<https://www.sketchengine.eu/>

■ Reference corpora evaluation

- Comparison of results from extraction using General and Legal corpora

Result
Labour Law Glossary (ES)
Data Protection Glossary (EN)
Industrial Standards Glossary (EN)

■ Lynx CKAN

- Displays different resources
- Gathers metadata
- Filters by jurisdiction, language, format, etc.

The screenshot shows the 'UNESCO Thesaurus' dataset page on the 'Data Portal for Compliance'. The page has a header with navigation links: 'Data Portal for Compliance', 'Datasets', 'Organizations', 'Groups', and 'About'. A search bar is located on the right. The breadcrumb trail is 'Home / Organizations / OEG / UNESCO Thesaurus'. The main content area has tabs for 'Dataset', 'Groups', and 'Activity Stream', with a 'Manage' button on the right. The 'Dataset' tab is active, showing the title 'UNESCO Thesaurus' and a description: 'The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in several fields.' Below this, there are two resource sections: 'Data and Resources'. The first section is 'SPARQL endpoint' with a 'DATA' icon and an 'Explore' button. The second section is 'Downloadable files' with an 'RDF' icon and an 'Explore' button. On the left sidebar, there is a 'Followers' section showing '0' and a 'Follow' button, and an 'Organization' section with a building icon.

Data Portal for Compliance

Datasets Organizations Groups About

Search

Home / Organizations / OEG / **UNESCO Thesaurus**

UNESCO Thesaurus

Followers
0

[+ Follow](#)

Organization

UNESCO Thesaurus

The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in several fields.

Data and Resources

SPARQL endpoint
SPARQL endpoint [Explore](#)

Downloadable files
Downloadable files in RDF and Turtle. [Explore](#)

Additional Info

Field	Value
Type (LKG)	Language Resource
Type	dataset
Domain	Education, Science, Culture, Politics, Countries, Information
Identifiers	
Availability	online
Language	en, es, fr, ru
Creator	Research group of Information Technology (University of Murcia)
Publisher	UNESCO
Licence	Creative Commons 3.0, https://creativecommons.org/licenses/by-nc-sa/3.0/deed.es_ES
Other rights	No
Jurisdiction	
Date	11/04/18
Proposed by	UPM
Number of entries	4408 (skos concepts)
Last update	2015
State	active

■ Converted resources

- New LL glossary (ES; TBX)
- New DP glossary (EN; TBX)
- New IS glossary (EN; TBX)
- Existing LL Termcat glossary (CA, EN, ES, FR; XML)

■ Vocabularies applied: fist approach

SKOS
skos:prefLabel
skos:altLabel
skos:definition
skos:note
skos:broader
skos:topConcept

DublinCore
dc:creator
dc:date
dc:title
dc:description

- Ontolex
Vartrans Module

Stage 2: Conversion into RDF

conceptId	prefLabel	definition
controller-n-en	controller	The chief accounting officer of a business enterprise or an institution (such as a college)
derogation-n-en	derogation	Provision in an EU legislative measure which allows for all or part of the legal measure to be applied differently, or not at all, to individuals, groups or organisations
member-n-en	member	A person or legal entity having an interest as a partner or shareholder in a firm or other enterprise

altLabel	altLabel	broader	topConcept	note
		annulment		
				E.g. "Spain is member of the European Union"

- conceptId to build the URI of each entry:

<http://linguistic.linkeddata.es/terminoteca/lynx/controller-n-en>

■ Resources to link

- New LL glossary (ES; RDF)
- New DP glossary (EN; RDF)
- New IS glossary (EN; RDF)
- Existing LL Termcat glossary (ES; RDF)

■ OpenRefine linking tests (ongoing)

- DBpedia
- Eurovoc
- Babelnet

■ Current status:

- Labour Law Glossary in Spanish linked to Babelnet
- Data Protection Glossary in English linked to Babelnet

■ Linking issues


- Semi-automatic process



■ Linking issues

- Unreliable

```
<rdf:Description rdf:about="http://linguistic.linkeddata.es/terminoteca/lynx/parliament-n-en">  
  <skos:inScheme rdf:resource="http://linguistic.linkeddata.es/terminoteca/lynx"/>  
  <skos:prefLabel>Parliament</skos:prefLabel>  
  <owl:sameAs rdf:resource="http://babelnet.org/rdf/parliament_n_MS"/>  
</rdf:Description>
```



This is an automatic match with probability 1

■ Next steps

- Link the two remaining glossaries

■ Future work: current glossaries

- Translation of terms in the glossaries
- Representation of translations with vartrans Ontolex module
- Explore other properties to add more information to the datasets

■ Future work: linking process

- Evaluation of other tools: Karma, Poolparty
- Fully-automatic linking process

■ Future work: general [Lynx]

- Transformation and linking of existing resources into RDF
- Widen reusable resources search
- Explore other extraction tools: Poolparty
- Automatic creation of figures (low priority)

- Thanks so much!!
- Questions?



Towards a Linked Open Data Cloud of Language Resources in the Legal Domain

Patricia Martín Chozas

Ontology Engineering Group (OEG)
Universidad Politécnica de Madrid (UPM)

OEG weekly meeting
May 31st, 2018