



Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe

D2.1 Initial Data Management Plan

PROJECT ACRONYM	Lynx
PROJECT TITLE	Building the Legal Knowledge Graph for Smart Compliance Services in Multilingual Europe
GRANT AGREEMENT	H2020-780602
FUNDING SCHEME	ICT-14-2017 - Innovation Action (IA)
STARTING DATE (DURATION)	01/12/2017 (36 months)
PROJECT WEBSITE	http://lynx-project.eu
COORDINATOR	Elena Montiel-Ponsoda (UPM)
RESPONSIBLE AUTHORS	Víctor Rodríguez-Doncel (UPM), Patricia Martín Chozas (UPM), Elena Montiel-Ponsoda (UPM)
CONTRIBUTORS	
REVIEWERS	Christian Sageder (openlaws), Tatjana Gornostaja (TILDE), Georg Rehm (DFKI)
VERSION STATUS	V3 Final
NATURE	Report
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.1256834
DATE	31/05/2018



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780602

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	First version	10/05/2018	Víctor Rodríguez-Doncel, Patricia Martín Chozas (UPM)
02	Second version. Implements changes suggested by reviewers. Sections reorganised.	25/05/2018	Víctor Rodríguez-Doncel (UPM)
03	Third version. Minor changes	31/05/2018	Víctor Rodríguez-Doncel (UPM), Elena Montiel-Ponsoda (UPM)

DISCLAIMER

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. Neither the Lynx consortium as a whole, nor a certain party of the Lynx consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.

EXECUTIVE SUMMARY

This deliverable contains an initial version of the Data Management Plan for Lynx following the template proposed by the EC (Section 2). The document is complemented with a catalogue of datasets belonging to the regulatory and linguistic domains which have been initially identified (Section 3). A methodology for identifying datasets is first described, which includes a template spreadsheet for the metadata description. A CKAN-based Lynx data portal has been also published, with the intention of becoming a reference site for the search of compliance-related datasets. A strategy for the harmonisation of data models has also been given along with a first description of data models of reference (Section 4). Finally, an initial description of the Legal Knowledge Graph is made (Section 5). This document will be superseded in M18 by the final Data Management Plan, Deliverable 2.4.

TABLE OF CONTENTS

EXECUTIVE SUMMARY	2
1 INTRODUCTION	8
2 DATA MANAGEMENT PLAN	9
2.1 DATA SUMMARY	9
2.2 FAIR DATA	9
2.2.1 Making data findable, including provisions for metadata.....	9
2.2.2 Making data openly accessible	9
2.2.3 Making data interoperable	12
2.2.4 Increase data reuse.....	13
2.3 ALLOCATION OF RESOURCES	13
2.4 DATA SECURITY	14
2.4.1 Data Security.....	14
2.5 LEGAL, ETHICAL AND SOCIETAL ASPECTS	14
2.5.1 Legal framework	14
2.5.2 Ethical aspects	15
2.5.3 Societal impact.....	15
3 CATALOGUE OF DATASETS.....	16
3.1 METHODOLOGY FOR CATALOGUING DATASETS	16
3.1.1 Template for data description	17
3.1.2 Lynx Data Portal.....	18
3.2 TRANSFORMATION OF RESOURCES.....	20
3.3 INITIAL CATALOGUE OF DATASETS	21
3.3.1 Datasets in the regulatory domain	21
3.3.2 Datasets in the language domain	21
4 DATA MODELS.....	26
4.1 INTRODUCTION	26
4.1.1 Data models in the regulatory domain.....	26

4.1.2	Data models in the linguistic domain	26
4.2	STRATEGY FOR THE HARMONISATION OF DATA MODELS IN LYNX	27
5	THE MULTILINGUAL LEGAL KNOWLEDGE GRAPH.....	28
5.1	SCOPE OF THE LEGAL KNOWLEDGE GRAPH	28
5.1	KNOWLEDGE GRAPHS	29
5.1.1	Legal Knowledge Graphs.....	29
5.1.2	Linguistic Knowledge Graphs.....	30
5.1.3	The Lynx Multilingual Legal Knowledge Graph.....	31
	REFERENCES.....	32

TABLE OF FIGURES

Figure 1. Schematic description of the Multilingual Legal Knowledge Graph for Compliance.....	8
Figure 2. Lynx public deliverable at Zenodo.	11
Figure 3. Deliverables on the Lynx website	11
Figure 4. A catalogue of relevant ontologies and vocabularies	13
Figure 5. Datasets in the LKG and out of it	16
Figure 6. Screenshot of the Lynx Data Portal	19
Figure 7. Usual activities for publishing linked data. Figure taken from [25].	20
Figure 8. Datasets represented by domain.	25
Figure 9. Datasets represented by format.....	25
Figure 11. Strategy for the selection of data models in Lynx	27
Figure 12. Scope of the multilingual Legal Knowledge Graph.....	28
Figure 13. Types of information in the Legal Knowledge Graph	29
Figure 14. Linguistic Linked Open Data Cloud	30

LIST OF TABLES

Table 1. Fields describing a data asset	17
Table 2. Fields describing a resource associated to a data asset	18
Table 3. Initial set of resources gathered.	24

ACRONYMS

AI	Artificial Intelligence
DCAT-AP	Data Catalogue vocabulary - Application profile for data portals in Europe
DMP	Data Management Plan
EC	European Commission
EU	European Union
FAIR	Findable, Accessible, Interoperable and Reusable
GA	Grant Agreement
GDPR	General Data Protection Regulation
IPR	Intellectual Property Rights
LKG	Legal Knowledge Graph
ORDP	Open Research Data Pilot
OWL	Web Ontology Language
RDF	Resource Description Framework
RDFS	Resource Description Framework Schema
W3C	World Wide Web Consortium

1 INTRODUCTION

This document contains the initial version of the Data Management Plan (DMP). The final version of this document will be available as “D2.4 Data Management Plan” in M18. This document is complemented by “D7.2 IPR and Data Protection Management”, to be delivered by M6 as well.

The Data Management Plan adheres to and complies with the *H2020 Data Management Plan – General Definition* given by the EC online, where the DMP is described as follows:

“A DMP describes the data management life cycle for the data to be collected, processed and/or generated by a Horizon 2020 project. As part of making research data findable, accessible, interoperable and reusable (FAIR), a DMP should include information on:

- the handling of research data during and after the end of the project
- what data will be collected, processed and/or generated
- which methodology and standards will be applied
- whether data will be shared/made open access and
- how data will be curated and preserved (including after the end of the project)”

Section 2 follows the template proposed by the EC¹. Lynx adopts policies compliant with the official FAIR guidelines [1] (findable, accessible, interoperable and re-usable).

Lynx participates Open Research Data Pilot (ORDP) and is obliged to deposit the produced research data in a research data repository. For such effect, the Zenodo repository has been chosen, which exposes the data to OpenAIRE granting its long term preservation. The description of the most relevant datasets for compliance have been published in a Lynx Data Portal, using CKAN technology. Metadata is provided for every relevant dataset, and data is selectively provided whenever it can be republished without license restrictions and relevance for the project is high. This deliverable also describes a catalogue of relevant legal and regulatory data models and a strategy for the homogenisation of the data sources.

Finally, the document describes the concept of a Multilingual Legal Knowledge Graph for Compliance, or Legal Knowledge Graph for short (Section 5), which is the backbone on when the Lynx services will rest (Figure 1).

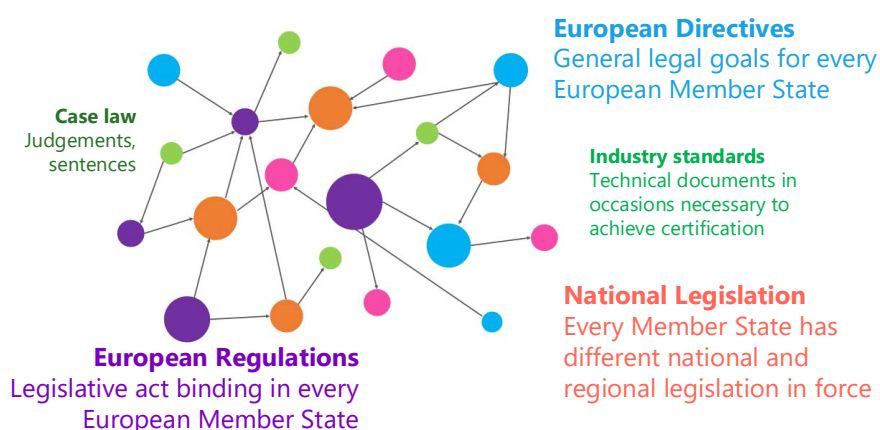


Figure 1. Schematic description of the Multilingual Legal Knowledge Graph for Compliance

¹ http://ec.europa.eu/research/participants/data/ref/h2020/gm/reporting/h2020-tp1-oa-data-mgt-plan_en.docx

2 DATA MANAGEMENT PLAN

This Section is the Initial Data Management Plan. It follows the template proposed by the EC and is applicable to the data used in or generated by Lynx, with the sole exception of pilot-specific data, whose management may be further specified in per-pilot DMPs. If the implementation of the pilots required a different DMP, either new DMP documents or new additions to this document shall be defined by the pilot leaders and the resulting work included in D2.4.

2.1 DATA SUMMARY

Purpose. The main objective of Lynx is “to create an ecosystem of smart cloud services to better manage compliance, based on a legal knowledge graph (LKG) which integrates and links heterogeneous compliance data sources including legislation, case law, standards and other aspects”. In order to deliver these smart services, data will be collected and integrated into a Legal Knowledge Graph, to be described in more detail in Section 3.

Formats. The very nature of this project makes the number of formats too high as to be foreseen in advance. However, the project will be keen on gathering data in RDF format or producing RDF data itself. RDF will be the format of choice for the meta model, using standard vocabularies and ontologies as data models. More details on the initially considered data models are given in Section 5.

Data reuse. The core part of the LKG will be created by reusing existing datasets, either copying them into the consortium servers (only if strictly needed) or using them directly from the sources.

Data origin. Although Lynx will be greedy in gathering and linking as much compliance-related data as possible from any possible source, it can be foreseen that the Eur-Lex portal will be the principal data source. Users of the Pilots may contribute their own data (e.g. private contracts, paid standards), which will be neither included into the LKG nor made publicly available.

Data size. The strong reliance of Lynx in external open data sources minimizes the amount of data that Lynx will have to physically store. No massive data storage infrastructure is foreseen.

Data utility. Data will be useful for SMEs and EU citizens alike through different portals.

2.2 FAIR DATA

2.2.1 Making data findable, including provisions for metadata

Discoverability. Data will be discoverable through a dedicated data portal (<http://data.lynx-project.eu>), further described in Section 4. Data assets will be identified with a harmonized policy to be defined in the forthcoming months.

Naming convention. A specific URI minting policy will be used to identify data assets. The policy will be specified in the forthcoming months after the publication of this deliverable.

Search keywords. Open datasets described in the Lynx data portal are findable through standard forms including keyword search.

Versioning. Versioning is an intrinsic part of the URI strategy to be devised.

Metadata. Metadata records describing each dataset will be downloadable as DCAT-AP entries.

2.2.2 Making data openly accessible

Open data: data in the LKG.

The adopted approach is “as open as possible, as closed as necessary”. Data assets produced during the project will preferably be published as open data. Nevertheless, during the project some datasets will be created from existing private resources (e.g. dictionaries by KDictionaries), whose publication would irretrievable damage their business model. These datasets will not be released as open data.

Datasets in the LKG will be in any case published along with a license. This license will be specified as a metadata record in the data catalog, which can also be exported as RDF using the appropriate vocabulary terms (`dtc:license`) and eventually using machine readable licenses.

Open data: research data.

In December 2013, the EC announced their commitment to open data through the Pilot on Open Research Data, as part of the Horizon 2020 Research and Innovation Programme. The Pilot’s aim is to “improve and maximise access to and reuse of research data generated by projects for the benefit of society and the economy”. In the frame of this Pilot on Open Research Data, results of publicly-funded research should be disseminated more broadly and faster, for the benefit of researchers, innovative industry and citizens.

The Lynx project chose to participate in the Open Research Data Pilot (ORDP). Consequently, publishing as “open” the digital research data generated during the project is a contractual obligation (GA Art. 29.3). This provision does not include the pieces of data which are derivative of private data of the partners. Their openness would endanger their economic viability and jeopardize the Lynx project itself (which is sufficient reason not to open the data as per GA Art. 29.3).

Every Lynx partner will ensure Open Access to all peer-reviewed scientific publications relating to its results. Lynx will use Zenodo as the online repository (<https://zenodo.org/communities/lynx/>) to upload public deliverables and possibly part of the scientific production. Zenodo is a research data repository created by OpenAIRE to share data from research projects. Records are indexed immediately in OpenAIRE, which is specifically aimed to support the implementation of the EC and ERC Open Access policies. Nevertheless, in order to avoid fragmentation, the Lynx webpage will act as the central information node.

The following categories of outputs require Open Access to be provided free of charge by Lynx partners, to related datasets, in order to fulfil the H2020 requirements of making it possible for third parties to access, mine, exploit, reproduce and disseminate the results contained therein:

- *Public deliverables* will be available both at Zenodo and the Lynx website at <http://lynx-project.eu/publications/deliverables>. See Figure 1 and Figure 2.
- *Conference and Workshop presentations* may be published at Slideshare under the account <https://www.slideshare.net/LynxProject>.
- *Conference and Workshop papers and articles for specialist magazines* may be also reproduced at: <http://lynx-project.eu/publications/articles>.
- *Research data and metadata* are also available. Metadata and selected data is available in the CKAN data portal, <http://data.lynx-project.eu>, produced research data at Zenodo.

Information will be also given about tools and instruments at the disposal of the beneficiaries and necessary for validating the results.

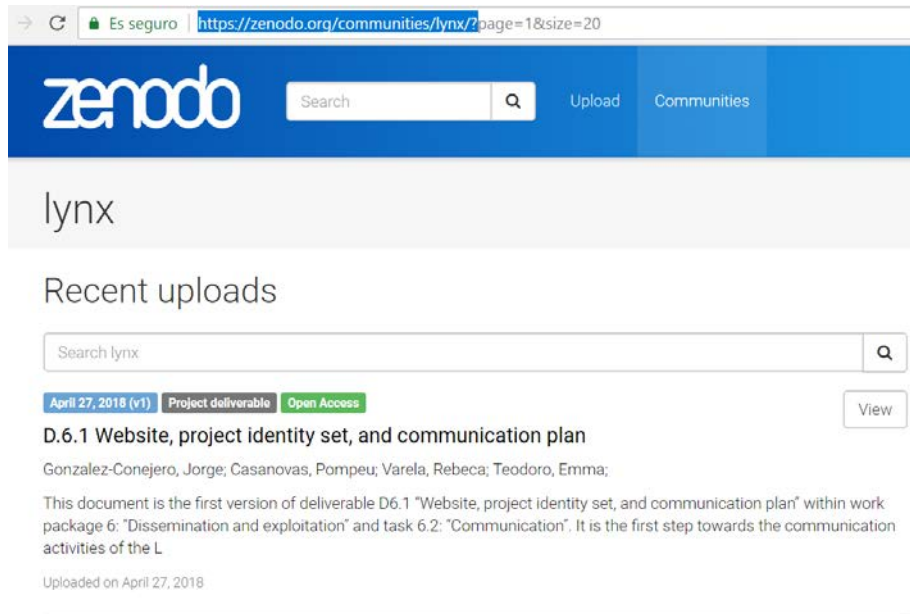


Figure 2. Lynx public deliverable at Zenodo.

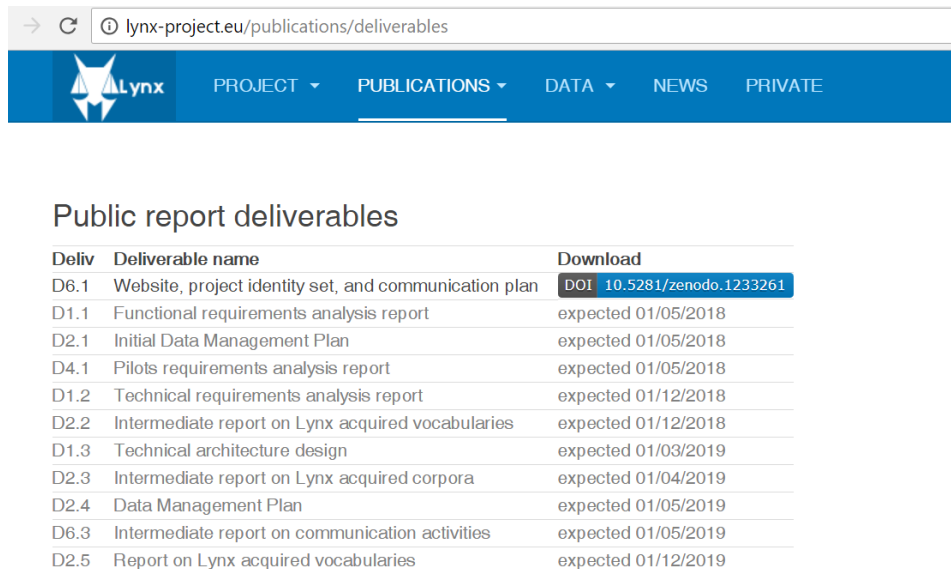


Figure 3. Deliverables on the Lynx website

Accessibility. Data descriptions (metadata) will be accessible through a dedicated data portal, hosted in Madrid and available under <http://data.lynx-project.eu>. Eventually, data from small datasets will be also made available from the web server –where *small* means a file size that does not compromise the web server availability. Eventually the metadata descriptions will be uploaded into other repositories, such as Retele² resources in Spanish language, ELRC-SHARE³ in general and others to be identified. In addition, the cooperation with the CEF eTranslation⁴ TermBank project will be considered, in view of sharing terminological domain-specific resources.

² <http://catalogo.retele.linkeddata.es/>

³ The ELRC-SHARE repository is used for documenting, storing, browsing and accessing Language Resources that are collected through the European Language Resource Coordination. <https://www.elrc-share.eu/>

⁴ The objective of the eTranslation Termbank action, launched by the EC, is to identify and collect terminology resources relevant to national public services, administrations, and governmental institutions across European countries.

Necessary methods and tools to access the data and its documentation. Relevant datasets whose license is liberal will be available as downloadable files. Eventually, a SPARQL endpoint will be set in place for those dataset in RDF form. Also, the CKAN technology in which the portal is based on, offers an API using standard JSON structures to access the data. The CKAN platform provides the documentation on how to use the API (<http://docs.ckan.org/en/ckan-2.7.3/api/>).

Publication of software. Some of the software to be developed in Lynx is expected to be published as Open Source. Other software to be developed in Lynx will be derived from private or non-open source code and, thus, not be made publicly accessible.

Data and code repositories and arrangement. Lynx uses a private source code repository (<https://gitlab.com/superlynx>). Open data will be deposited in the Lynx open data portal; consortium-internal data within the project intranet. The choice of Nextcloud is justified as the information resides within UPM secured servers in Madrid, avoiding third parties and granting the privacy and confidentiality of the data. Gitlab, as a major provider and host of code repositories, is a common choice among developers but if necessary code might be also hosted at UPM.

Data Access Committee. As of today, there is no need for a Data Access Committee⁵.

Conditions for access. Description of data assets include a link to well-known licenses, for which machine readable versions exist. Either Creative Commons Attribution International 4.0 (CC-BY) or Creative Commons Attribution Share-Alike International 4.0 (CC-BY-SA) will be the recommended licenses.

Access control. The Lynx intranet (Nextcloud) provides standard access control functionalities. The servers are located in a secured data centre at UPM. The access point is <https://delicias.dia.fi.upm.es/lynx-nextcloud/>. Access is secured by asymmetric keys or passwords and communications use SSL.

2.2.3 Making data interoperable

Interoperability. The LKG preferred format is RDF, granting interoperability between institutions, organisations and countries. This choice optimally facilitates re-combinations with different datasets from different origins.

Data and metadata vocabularies. Specific data and metadata vocabularies will be defined throughout the entire project. An initial collection has already been edited and will be soon published at <http://lynx-project.eu/data/data-models> (see also Figure 3).

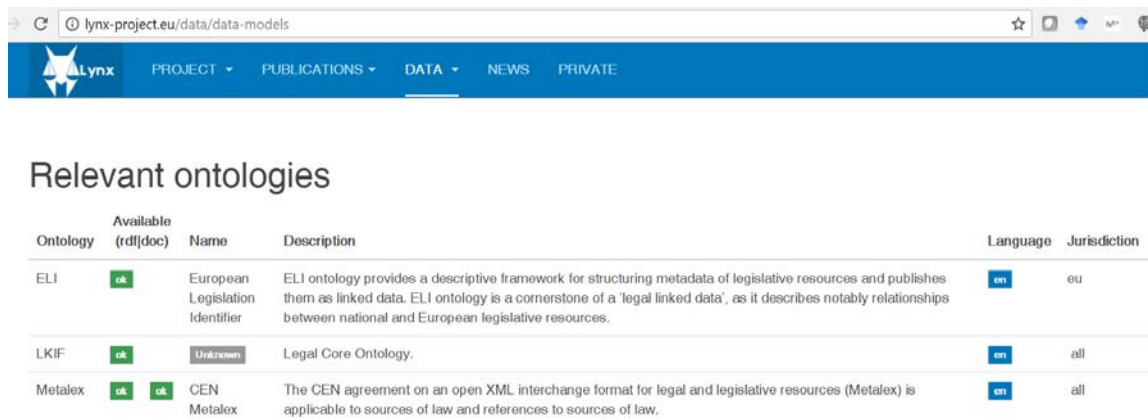
Standard vocabularies and inter-disciplinarity. Standard vocabularies will be used inasmuch as possible, like the ECLI ontology, the Ontolex model and other vocabularies similarly spread. These choices grant inter-disciplinary collaboration. For example, Ontolex⁶ is standard in the language resources and technologies communities, whereas the ELI ontology⁷ (European Law Identifier) is standard in the European legal community.

Mappings of vocabularies and ontologies developed by Lynx. If vocabularies or ontologies are further defined, they will be published online, documented and mapped to other standard ontologies. Figure 3 illustrates a possible visualization for the data models.

⁵ A Data Access Committee is a body of one or more named individuals who are responsible for data release to external requestors.

⁶ <http://lemon-model.net/>

⁷ <http://publications.europa.eu/mdr/eli/>



Ontology	Available (rdl/doc)	Name	Description	Language	Jurisdiction
ELI	ok	European Legislation Identifier	ELI ontology provides a descriptive framework for structuring metadata of legislative resources and publishes them as linked data. ELI ontology is a cornerstone of a 'legal linked data', as it describes notably relationships between national and European legislative resources.	en	eu
LKIF	ok	Unknown	Legal Core Ontology.	en	all
Metalex	ok ok	CEN Metalex	The CEN agreement on an open XML interchange format for legal and legislative resources (Metalex) is applicable to sources of law and references to sources of law.	en	all

Figure 4. A catalogue of relevant ontologies and vocabularies

2.2.4 Increase data reuse

Embargoes. No data embargoes are foreseen. Public data will be published as soon as possible, but private data will remain private as long as the interested parties, rightsholders of the data, decide.

Data after the project. Lynx aims at building a LKG towards compliance. In the long term, the LKG may be repurposed and the data portal may become a reference entry point to find open, linguistic legal information as RDF.

Data validity after time. Some of the datasets require maintenance (e.g. legislation and case law must be kept up to date). Whereas a core of information may still be of interest even with no maintenance, those datasets directly used by services under exploitation will be maintained. In any case, metadata records describing the datasets will include a field informing on the last modification date.

Data quality assurance. Only formal aspects of data quality are expected to be assured. In particular, the 5-stars⁸ will be considered, and the data portal will describe this quality level in due time.

2.3 ALLOCATION OF RESOURCES

Costs. The cost of publishing FAIR data includes (a) maintenance of the physical servers; (b) time devoted to the data generation and (c) long term preservation of the data.

Coverage of the costs. Resources to maintain and generate data are covered by the project. Long term preservation of data is free by uploading the research data at Zenodo.

Responsibility of the Data Management. UPM is responsible for managing data in the data portal, and for managing private data in the intranet. UPM is not responsible of keeping personal data collected to provide the pilot services but the directly involved partners (openlaws, Cuatrecasas, DNV GL).

Long term preservation. Public deliverables and research data will be uploaded to Zenodo, which grants the long term preservation. A specific community has been created in Zenodo⁹. Alternatively, if difficulties are found with Zenodo, datasets may also be uploaded to Figshare¹⁰ or B2Share¹¹ where a permanent DOI is retrieved. Other sites such as META-SHARE, ELRC-SHARE or the European Language

⁸ <http://www.w3.org/DesignIssues/LinkedData.html>

⁹ <https://zenodo.org/communities/lynx/>

¹⁰ <https://figshare.com/>

¹¹ <https://b2share.eudat.eu/>

Grid may be considered in addition to grant long term preservation and maximize the impact and dissemination.

2.4 DATA SECURITY

2.4.1 Data Security

Data security. UPM is physically storing data on their servers: webpage, files and data in the Nextcloud system, the CKAN data catalogue and mailing lists. These pieces of data are both digitally and physically secured in a data centre. Backups are made of these systems, to external hard disks or other machines. In principle, no personal data will be kept at UPM, and the pilot leaders will define specific DMP with specific data protection provisions and specific data security details.

Long term preservation. Relevant data which is open, shall be uploaded to Zenodo. In addition, relevant language datasets produced in the course of Lynx will be uploaded to catalogues of language resources.

2.5 LEGAL, ETHICAL AND SOCIETAL ASPECTS

2.5.1 Legal framework

EU citizens are granted the rights of privacy and data protection by the Charter of Fundamental rights of the EU. In particular, Art. 7 states that *“everyone has the right respect for private and family life, home and communications”*, whereas Art. 8 regulates that *“everyone has the right to the protection of personal data concerning him or her”* and that processing of such data must be *“on the basis of the consent of the person concerned or some other legitimate basis laid down by law.”*

These rights are developed in detail by the General Data Protection Regulation (GDPR), Regulation 2016/679/EC, which is in force in every Member State since 25th of May of 2018. This regulation imposes obligations to the Lynx consortium, which is also reminded by Art. 39 of the Lynx Grant Agreement (GA): *“the beneficiaries must process personal data under the Agreement in compliance with applicable EU and national law on data protection”* The same GA also reminds that beneficiaries *“may grant their personnel access only to data that is strictly necessary for implementing, managing and monitoring the Agreement”* (GA Art. 39.2).

Personal data is, according to GDPR art. 4.1 *“any information relating to an identified or identifiable natural person (‘data subject’); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person”*, whereas *data processing* is (art. 4.2): *“any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction”*. With these definitions, Pilot 1 (Compliance Assurance Services in Data Protection) will most likely have to collect and process personal data, and possibly other Pilots as well.

The purposes for which personal data will be collected are justified in compliance with art.5.b, and the processing of personal data is legitimate in compliance with art. 6. The implementation of the Pilot 1 and other pilots processing personal data will have to implement the necessary legal provisions to respect the rights of the data subjects.

Several internal communication channels have been established for Lynx: mailing lists, a website and an intranet. The three servers are hosted at UPM and comply with the Spanish legislation.

The Lynx web site (<http://lynx-project.eu>) is compliant regarding the management of cookies with *Ley 34/2002, de 11 de julio, de servicios de la sociedad de la información y de comercio electrónico*. Lynx will most likely handle datasets with personal data (Pilot 1), as users will be registered in the Lynx platform to enjoy personalised services and to upload contracts with personal data. The consortium will adopt any measure to comply with the current legislation.

2.5.2 Ethical aspects

The ethical aspect of greatest interest is the processing of personal data. The processing of personal data may become a possibility in the framework of Pilot 1. GA Article 34 “Ethics and research integrity” is binding and shall be respected. Ethical and privacy related concerns are fully addressed in Section 3.2 of Deliverable 7.2 “*IPR and Data Protection management documents*”.

Besides, the ethics issues identified are already being handled by the pilot organisations during their daily operation activities, as they confront with national laws and EU directives regarding the use of information in their daily services, as clearance for the processing, storing methods, data destruction, etc. has been provided to such organisation a priori and is not case specific. The research to be done during Lynx does not raise any other issues, and the project will make sure that it will follow the same patterns and rules used by the pilot organisations, that will guarantee the proper handling of ethical issues and the adherence to national, EU wide and international law and directives that do not violate the terms of the programme.

2.5.3 Societal impact

The societal impact of this project is expected to be positive, enhancing the access of EU citizens to legislation and contributing towards a fairer Europe. In addition to the best effort made by the project partners, members of the Advisory Board may be requested to issue a statement on the ethical and societal impact of the Lynx project.

3 CATALOGUE OF DATASETS

This section describes a catalogue of relevant legal, regulatory and linguistic datasets. Datasets in the Legal Knowledge Graph are those necessary to provide compliance related services that also meet the requirement of being published as linked data. The purpose of Lynx Task 2.1 is twofold:

- a) Identify as many as possible open dataset possibly relevant to the problem in question (either in RDF or not)
- b) Build the Legal Knowledge Graph by identifying existing linked data resources or by transforming existing datasets into linked data whenever necessary

Figure 5 represents the Legal Knowledge Graph as a collection of dataset published as linked data. The LKG lies amidst another cloud of datasets, in various formats either structured or not (such as PDF, XLS or XML). The section contains: (a) the methodology followed to describe datasets of interest; (b) the methodology to transform existing resources into LKG datasets; (c) a description of the Lynx data portal and the related technology and (d) an initial list of relevant datasets.

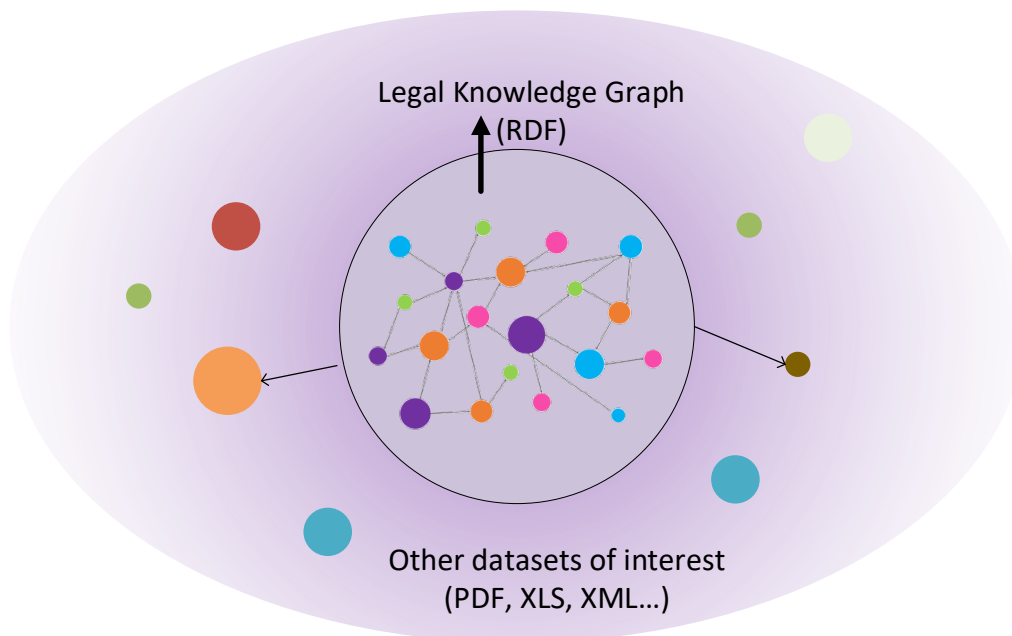


Figure 5. Datasets in the LKG and out of it

3.1 METHODOLOGY FOR CATALOGUING DATASETS

Data assets potentially relevant to the Lynx project are those that might help providing multilingual compliance services. They might be referenced by datasets in the LKG as external references.

The identification and description of these datasets is being made during the project in a cooperative way, during the entire project lifespan. The methodology has consisted of the following steps:

1. *Identification of datasets of possible interest*
 - Identification of relevant datasets by the partners;
 - Discovery of relevant datasets by browsing data portals, reviewing literature and making general searches;
2. *Description of resources*
 - Description of the resources identified in Step 1 using an agreed template (spreadsheet) with metadata records (see Section 4.2.1).
3. *Publication of dataset descriptions*
 - Publication of the dataset description in the CKAN Open Data Portal via CKAN form

— Transformation of the metadata records to RDF using the vocabulary DCAT-AP (to be an automated task from the spreadsheet)

This process is being iteratively carried out throughout the project.

3.1.1 Template for data description

Every partner of Lynx, within their domain of expertise, has described an initial list of data sources of interest for the project. In order to homogeneously describe the data assets, a template with metadata records has been created with the due consensus among the partners.

The template for data description contains two main blocks: one with general information about the dataset and another with information about the resource. Within this context, “dataset” makes reference to the whole asset, while “resource” defines each one of the different formats in which the dataset is published. For instance, the UNESCO thesaurus is a single dataset which can be found as two different resources: as a SPARQL Endpoint and as a downloadable file in RDF.

Thereby, the metadata records in Table 1 describe information about the dataset as a whole.

Field	Description
Title	the name of the dataset given by the author or institution that publishes it.
URI	identifier pointing to the dataset.
Type in the LKG	type of dataset in the legal knowledge graph (language, data, etc.).
Type	type of dataset (term bank, glossary, vocabulary, corpus, etc.).
Domain	topic covered by the dataset (law, education, culture, government, etc.).
Identifiers	other type of identifiers assigned to the dataset (ISRN, DOI, Standard ID, etc.).
Description	a brief description of the content of the dataset.
Availability	if the dataset is available online, upon request or not available.
Languages	languages in which the content of the dataset are available.
Creator	author or institution that created the dataset.
Publisher	institution publishing the dataset.
License	license of the dataset (Creative Commons, or others).
Other rights	if the dataset contains personal information.
Jurisdiction	jurisdiction where the dataset applies (if necessary).
Date of this entry	date of registration of the dataset in the CKAN.
Proposed by	Lynx partner or Lynx organisation proposing the dataset.
Number of entries	number of terms, triplets or entries that the dataset contains.
Last update	date in which the last modification of the dataset took place.
Dataset organisation	name of the Lynx organisation registering the dataset.

Table 1. Fields describing a data asset

The second block of metadata (whose fields are listed in Table 2) gives additional information about the resource in which the metadata can be accessed. This section is repeated as many times as needed (depending on the number of formats of the metadata).

Field	Description
Description	description of the type of resource (i.e. downloadable file, SPARQL endpoint, website search application, etc.).
Data format	the format of the resource (RDF, XML, SKOS, CSV, etc.).
Data access	technology used to expose the resource (relational database, API, linked data, etc.).
Open format	if the format of the resource is open or not.
URI	the URI pointing to the different resources.

Table 2. Fields describing a resource associated to a data asset

The template was materialized as a spreadsheet distributed among the partners.

3.1.2 Lynx Data Portal

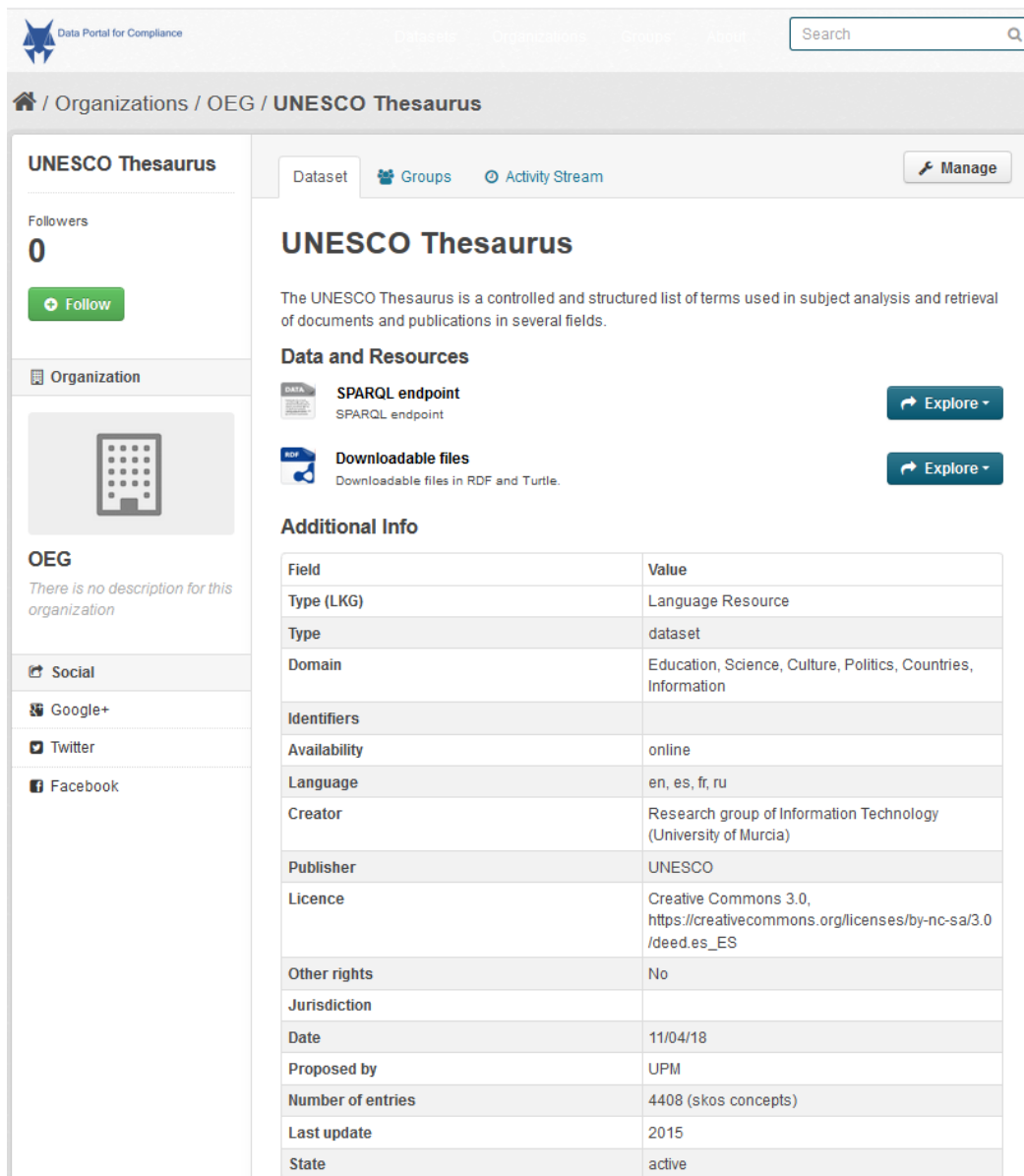
With the aim of publishing the metadata of the harvested datasets, a data portal has been made available under <http://data.lynx-project.eu>.

This data portal uses the technology of CKAN. The Comprehensive Knowledge Archive Network (CKAN) is a web-based management system for the storage and distribution of open data. The system is open source¹², and it has been deployed on the UPM servers using containerization technologies –Rancher¹³, a leading solution to deploy Docker containers in a Platform as a Service (PaaS).

The CKAN open data portal gives access to the resources gathered by all the members of the Lynx project. In the same way, members are able to register and describe their harvested resources to jointly create the Lynx Open Data Portal. To correctly display the relevant information about the datasets, CKAN application uses the metadata described in Section 4.2.1. As a result, each dataset presents the interface as shown by Figure 5 .

¹² <https://github.com/ckan/ckan>

¹³ <https://rancher.com/>



UNESCO Thesaurus

Followers: 0

Organization: OEG

There is no description for this organization

Social: Google+, Twitter, Facebook

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: [Explore](#)
- Downloadable files: [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

Figure 6. Screenshot of the Lynx Data Portal

The “Data and Resources” section corresponds to the “Resource information” metadata block and “Additional Info” contains the metadata of the “Dataset information” table.

The CKAN data portal allows faceted browsing, with filters such as language, format and jurisdiction. At this moment, there are 26 datasets classified in the CKAN, but this number will grow. For the metadata records to be correctly displayed on the website, it was required to establish a correspondence between the metadata in the spreadsheet and the structure in the JSON file that gives shape to the CKAN platform.

In the Lynx Data Portal, each dataset can be accessed through their own URI, that is built by using the ID of each resource. Datasets IDs are shown in Table 3, contained in the next section. As a result, dataset URIs look like the example below, where the ID would be unesco-thesaurus:

<http://data.lynx-project.eu/dataset/unesco-thesaurus>

The CKAN API enables a direct access to the metadata records. The API is intended for developers who want to write code that interacts with CKAN sites and their data, and it is documented online¹⁴. For example, the method:

```
http://data.lynx-project.eu/api/rest/dataset/unesco-thesaurus
```

will return the following answer:

```
{
  "license_title": null,
  "maintainer": null,
  "private": false,
  "maintainer_email": null,
  "num_tags": 0,
  "id": "efaf72c9-f8da-4257-b77e-c1f90952d71a",
  "metadata_created": "2018-04-11T08:35:41.813169",
  "relationships": [],
  "license": null,
  "metadata_modified": "2018-04-11T08:39:59.429186",
  "author": null,
  "author_email": null,
  "download_url": "http://skos.um.es/sparql/",
  "state": "active",
  "version": null,
  "creator_user_id": "3b131ddc-4bbf-42ff-9c33-ee1c4f7adb5c",
  "type": "dataset",
  "resources": [
    {
      "Distribuciones": "SPARQL endpoint",
      "hash": "",
      "description": "SPARQL endpoint",
      "format": "SKOS",
      "package_id": "efaf72c9-f8da-4257-b77e-c1f90952d71a",
      "mimetype_inner": null,
      "url_type": null,
      "formatoabierto": "",
      "id": "2a610dc8-15cd-4f17-ae0-149201c427cd",
      "size": null,
      "mimetype": null,
      "cache_url": null,
      "name": "SPARQL endpoint",
      "created": "2018-04-11T08:39:13.979840",
      "url": "http://skos.um.es/sparql/",
      "cache_last_updated": null,
      "last_modified": null,
      "position": 0,
      "resource_type": null,
      "Downloadable files": "Downloadable files",
      "hash": "",
      "description": "Downloadable files in RDF and Turtle.",
      "format": "RDF",
      "package_id": "efaf72c9-f8da-4257-b77e-c1f90952d71a",
      "mimetype_inner": null,
      "url_type": null,
      "formatoabierto": "",
      "id": "81ddd071-4018-4850-b5d8-04b4f5badd7d",
      "size": null,
      "mimetype": null,
      "cache_url": null,
      "name": "Downloadable files",
      "created": "2018-04-11T08:39:59.170137",
      "url": "http://skos.um.es/unescothes/downloads.php",
      "cache_last_updated": null,
      "last_modified": null,
      "position": 1,
      "resource_type": null
    }
  ],
  "num_resources": 2,
  "tags": [],
  "groups": [],
  "license_id": null,
  "organization": {
    "description": "",
    "title": "OEG",
    "created": "2018-04-05T08:10:35.821305",
    "approval_status": "approved",
    "is_organization": true,
    "state": "active",
    "image_url": "",
    "revision_id": "66f3c9c3-9bdf-4ebe-8ed2-54b4aea30375",
    "type": "organization",
    "id": "d4250a6e-d1d4-4a2d-8e40-b663271d8404",
    "name": "oeg",
    "name": "unesco-thesaurus",
    "isopen": false,
    "notes_rendered": "<p>The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in several fields.</p>",
    "url": null,
    "ckan_url": "http://data.lynx-project.eu/dataset/unesco-thesaurus",
    "notes": "The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in several fields.\r\n",
    "owner_org": "d4250a6e-d1d4-4a2d-8e40-b663271d8404",
    "ratings_average": null,
    "extras": {
      "lkg_type": "language",
      "domain": "Education, Science, Culture, Politics, Countries, Information",
      "total_number": "4408 (skos concepts)",
      "language": "en, es, fr, ru",
      "creator": "Research group of Information Technology (University of Murcia)",
      "publisher": "UNESCO",
      "jurisdiction": "",
      "other_rights": "no",
      "last_update": "2015",
      "licence": "Creative Commons 3.0, https://creativecommons.org/licenses/by-nc-sa/3.0/deed.es_ES",
      "date": "11/04/18",
      "partner": "UPM",
      "identifier": "",
      "availability": "online",
      "ratings_count": 0,
      "title": "UNESCO Thesaurus",
      "revision_id": "67553ea8-aa13-4dfe-905d-eb499d2d78e9"
    }
  }
}
```

3.2 TRANSFORMATION OF RESOURCES

The minimum content of the LKG is the collection of datasets necessary for the execution of the Lynx pilots that are published as linked data. Whereas transformation of resources to linked data is not a central activity of Lynx, the project foresees that some resources will exist but not as linked data, and a transformation process will be necessary.

The cycle of activities usually made when publishing linked data (see Figure 7) is

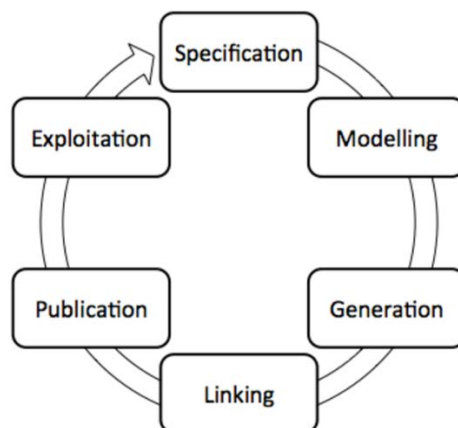


Figure 7. Usual activities for publishing linked data. Figure taken from [25].

¹⁴ <http://docs.ckan.org/en/ckan-2.7.3/api/>

Whereas the specification is derived from the pilots and the use case needs, the modelling process will lean on existing data models, to be harmonized as described in Section 4.2. The generation of linked data will be the transformation of existing resources. These transformation will be different depending on the source format:

- From unstructured text, extraction tools (PoolParty, OpenCalais, SketchEngine etc.) will be used, before creating the entities.
- From relational databases, technologies such as R2RML exist, but no relation database is expected to be necessary.
- For tabular data, Open Refine and similar tools will be used.

The publication means is to be decided but it will be made using either PoolParty or Open Link Virtuoso in local servers.

3.3 INITIAL CATALOGUE OF DATASETS

This section contains only preliminary information, and it will be completed by M18 with D2.4.

3.3.1 Datasets in the regulatory domain

These are the initially identified datasets in the regulatory domain:

- Eur-Lex: Database of legal information containing: EU law (EU treaties, directives, regulations, decisions, consolidated legislation, etc.) preparatory acts (legislative proposals, reports, green and white papers, etc.), EU case-law (judgments, orders, etc.), international agreements, etc. A huge database updated daily with some texts dating back to 1951.
- Openlaws: Austrian laws (federal laws and of the 9 regions) and rulings (from 10 different courts), German federal laws, European laws (regulations, directives) and rulings (general court, European Court of Justice). It includes Eur-Lex, 11k national acts and 300k national cases in a neo4j graph.
- DNV-GL: Standards, regulations and guidelines to the public, usually in PDF.

3.3.2 Datasets in the language domain

Using the methodology described in Section 4.2, several sites and repositories have been surveyed. One of the sources of most interest for linguistic open data is the Linked Open Data Cloud¹⁵ or LOD cloud, due to its open nature and its adequate format as linked data or RDF. In particular, the Linguistic Linked Open Data Cloud¹⁶ is a subset of the LOD cloud which provides exclusively linguistic resources sorted by typology. Different types of datasets in the Linguistic Linked Open Data Cloud are:

- Corpora
- Terminology, thesauri and Knowledge Bases
- Lexicons and Dictionaries
- Linguistic Resource Metadata
- Linguistic Data Categories
- Typological Databases

Within this project, the three first types of resources have been shortlisted as the most useful.

Besides consuming linked data or RDF in general, other valuable non-RDF resources can be included in the graph, possibly once converted to RDF. Many non-RDF resources of interest in this context can be

¹⁵ <http://lod-cloud.net/clouds/lod-cloud.svg>

¹⁶ <http://linguistic-lod.org/>

found in data portals like the European Data Portal, the Library of Congress or the Termcoord public portal, which is of particular interest for the multilingual glossaries in the domain of law.

Due to the huge amount of information and open data available nowadays, it is essential to establish these limits to gather only the relevant resources. In the case that more types of datasets are required, they will be harvested at a later stage. Thus, some of the resources already published as linked data and that have been identified as of interest for Lynx are listed below:

- STW Thesaurus for Economics: a thesaurus that provides a vocabulary on any economic subject. It also contains terms used in law, sociology and politics (monolingual in English) [30].
- Copyright Termbank: a multilingual term bank of copyright-related terms that has been published connecting WIPO definitions, IATE terms and definitions from Creative Commons licenses (multilingual).
- EuroVoc: a multilingual and multidisciplinary thesaurus covering the activities of the EU. It is not specifically legal, but it contains pertinent information about the EU and their politics and law (multilingual).
- AGROVOC: a controlled vocabulary covering all the fields of the Food and Agriculture Organization (FAO) of the United Nations. It contains general information and it has been selected since it shares many structures with other important resources (multilingual).
- IATE: a terminological database developed by the EU which is constantly being updated by translators and terminologists. Amongst other domains, the terms are related with law and EU governments (multilingual). A transformation to RDF was made in 2015.

Resources published in other formats have been considered as well. Structured formats include TBX (used for term bases), CSV and XLS. Exceptionally, resources published in non-machine-readable formats might be considered.

Consequently, the following resources published by the EU have also been listed as usable, although they are not included in the Linguistic Linked Open Data Cloud:

- INSPIRE Glossary: a term base developed by the INSPIRE Knowledge Base of the European Union. Although this project is related with the field of spatial information, the glossary contains general terms and definitions that specify the common terminology used in the INSPIRE Directive and in the INSPIRE Implementing Regulations (monolingual, en).
- EUGO Glossary: a term base addressed to companies and entrepreneurs that need to comply with administrative or professional requirements to perform a remunerated economic activity in Spain. This glossary is part of a European project and contains terms about regulations that are valuable for Lynx purpose (monolingual in Spanish).
- GEMET: a general thesaurus, conceived to define a common general language to serve as the core of general terminology for the environment. This glossary is available in RDF and it shares terms and structures with EuroVoc (multilingual).
- Termcoord: a portal supported by the European Union that contains glossaries developed by the different institutions. These glossaries cover several fields including law, international relations and government. Although the resources are available in PDF, at some point these documents could be treated and transformed into RDF if necessary (multilingual).

In the same way, the United Nations also counts with consolidated terminological resources. Given their intergovernmental domain, the following resources have been selected:

- UNESCO Thesaurus: a controlled list of terms intended for the subject analysis of texts and document retrieval. The thesaurus contains terms on several domains such as education, politics, culture and social sciences. It has been published as a SKOS thesaurus and can be accessed through a SPARQL endpoint (multilingual).
- InforMEA Glossary: a term bank developed by the United Nations and supported by the European Union with the aim of gathering terms on Environmental Law and Agreements. It is available as RDF and it will be upgraded to a thesaurus during the following months (multilingual).
- International Monetary Fund Glossary: a terminology list containing terms on economics and public finances related with the European Union. It is available as a PDF downloadable file; however, it may be transformed as a future work (multilingual).

On the other hand, other linguistic resources (not supported by the EU nor the UN) have been spotted. Some of them are already converted into RDF:

- Termcat (Terminologia Oberta): a set of terminological databases supported by the government of Catalonia. They contain term equivalents in several languages. Part of these terminological databases were converted into RDF previously and are part of the TerminotecaRDF project. They can be accessed through a SPARQL endpoint (multilingual).
- German Labour Law Thesaurus: a thesaurus that covers all main areas of labour law, such as the roles of employee and employer; legal aspects around labour contracts. It is available through a SPARQL endpoint and as RDF downloadable files (monolingual, de).
- Jurivoc: a juridical thesaurus developed by the Federal Supreme Court of Switzerland in cooperation with Swiss legal libraries. It contains juridical terms arranged in a monohierarchic structure (multilingual).
- SAIJ Thesaurus: a thesaurus that organises legal knowledge through a list of controlled terms which represent concepts. It is available in RDF and intended to ease users' access information related to the argentine legal system that can be found in a file or in a documentation centre (monolingual, es).
- CaLaThe: a thesaurus for the domain of cadastre and land administration that provides a controlled vocabulary. It is interesting because it shares structures and terms with AGROVOC and the GEMET thesaurus, and it can be downloaded as an RDF file (monolingual, en).
- CDISC Glossary: a glossary contains definitions of terms and abbreviations that can be relevant for medical laws and agreements It is available in several formats, including OWL (monolingual, en).

Finally, one last resource available in other PDF has also been considered due to different facts:

- Connecticut Glossary: a glossary that contains legal terms published by the Judicial Branch of the State of Connecticut. It can be transformed into a machine-readable format and from there into RDF since it provides with equivalences of legal terms from English into Spanish (bilingual).

Table 3 lists all the resources as a review of the information presented above. On the other hand, the set of the identified linguistic resources has also been represented in an interactive graph, in which each dataset is coloured as per the domain it covers (Figure 8). A second version of the graph has also been created in order to make a distinction between those datasets in RDF (green) and those in different formats (grey) (Figure 9). The graph also represents the relations between each asset, since most of those in RDF share structures and terms.

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
conneticut-legal-glossary	Conneticut Glossary	Legal	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	en
stw	STW Thesaurus for Economics		Economic monolingual thesaurus.	en
edp	European Portal	Data	EU datasets.	EU languages
inspire	INSPIRE (EU)	Glossary	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 (UNESCO)	Glossary	Monolingual glossary on environmental law.	en
copyright-termbank	Copyright Termbank		Multi-lingual term bank 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

Table 3. Initial set of resources gathered.

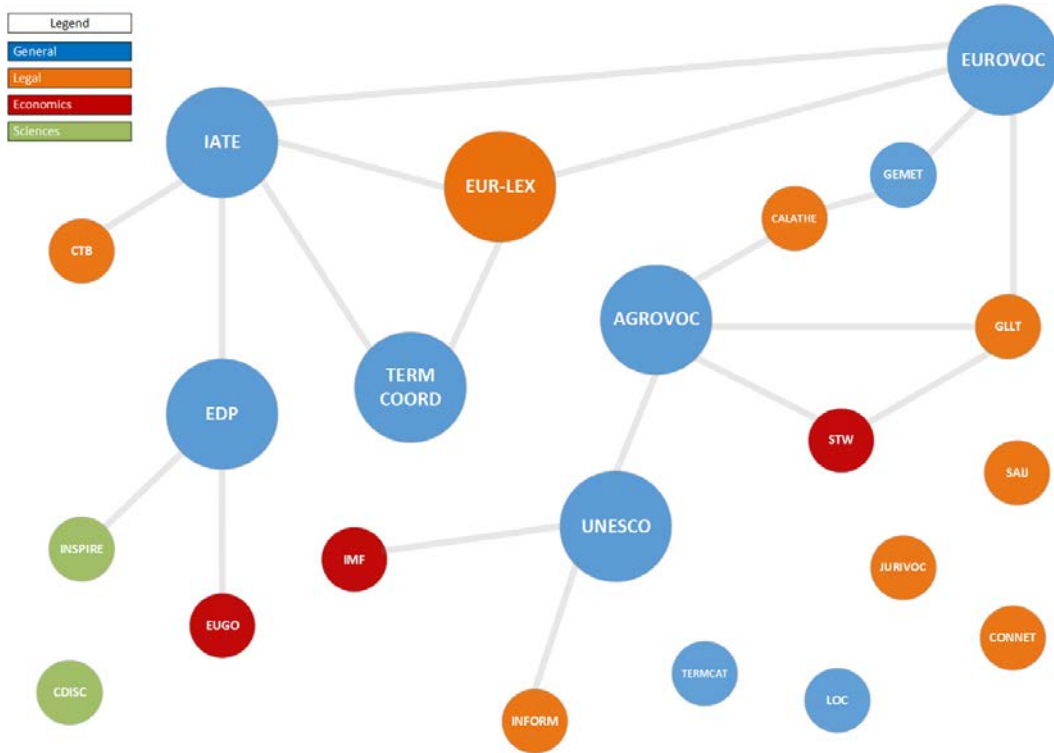


Figure 8. Datasets represented by domain.

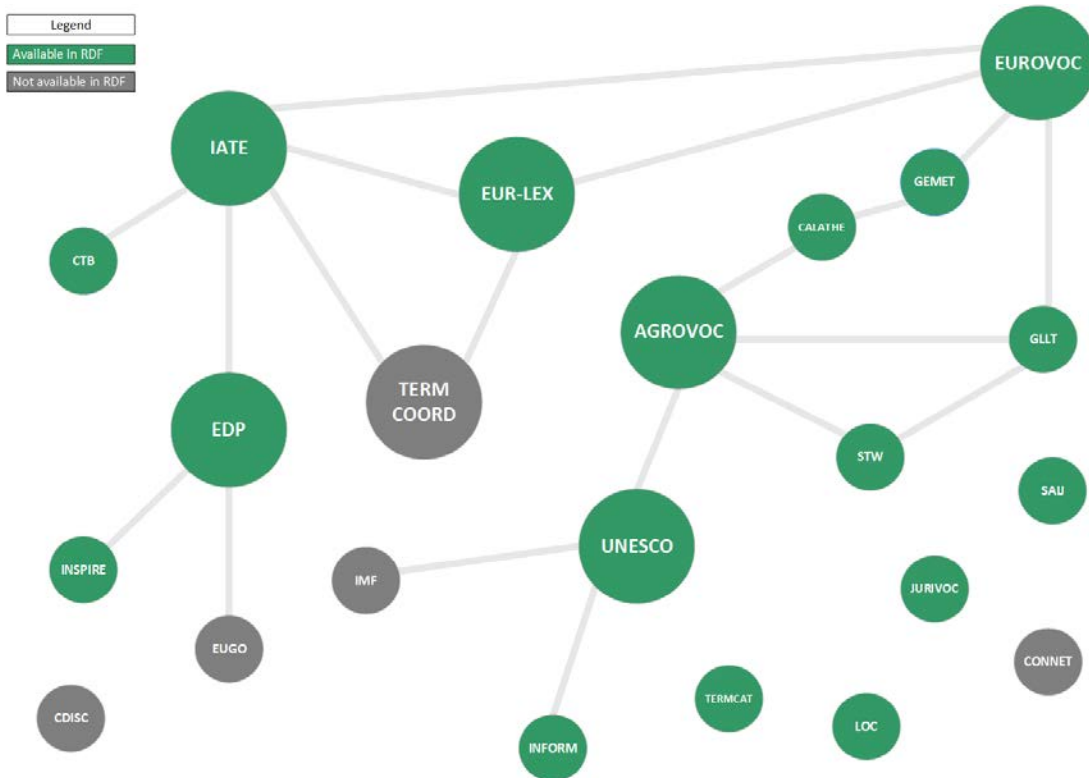


Figure 9. Datasets represented by format.

4 DATA MODELS

4.1 INTRODUCTION

4.1.1 Data models in the regulatory domain

A number of vocabularies and ontologies for documents in the legal domain has been published in the last few years. Núria Casellas surveyed 52 legal ontologies in 2011 [18], and in the meantime many other new ontologies have appeared, but in practice, only a few of them have direct interest for the LKG, as not every published legal ontology is created with the intention of supporting data models. Some ontologies had the intent of formalizing abstract conceptualizations. For example, ontology design patterns in the legal domain have been explored [17] –but these works have little interest for supporting data publication.

The XML schema Akoma Ntoso¹⁷ was initially funded by the United Nations to become some years later an OASIS specification as Legal RuleML¹⁸. MetaLex [12] was an XML vocabulary for the encoding of the structure and content of legislative documents, which included in newer versions functionality related to timekeeping and version management. The European Committee for Standardization (CEN) adopted MetaLex and evolved the schema to an OWL ontology. MetaLex was extended in the context of the FP6 ESTRELLA project (2006-2008) which developed a network of ontologies known as Legal Knowledge Interchange Format (LKIF). The LKIF ontologies are still available and a reference in the area¹⁹ [14]. Licenses used for the publication of copyrighted work have been modelled with the ODRL (Open Digital Rights Language) language [27].

The European Legislation Identifier (ELI) is a system to make legislation available online in a standardised format, so that it can be accessed, exchanged and reused across border [13]. ELI describes a new common framework to unify and link national legislation with European legislation. ELI, as a framework, proposes a URI template for the identification of legal resources on the web and it also provides an OWL ontology for supporting the representation of metadata of legal events and documents. The European Case Law Identifier (ECLI), much like ELI, was introduced recently for modelling case laws. The BO-ECLI project, funded under the Justice Programme of the European Union (2015-2017), aimed to broaden the use of ECLI and to further improve the accessibility of case law.

4.1.2 Data models in the linguistic domain

Similarly, a large amount of language resources can already be found across the Semantic Web. Such datasets are represented with various schemas, depending on given factors such as the inner structure of the dataset, language, content or the objective of its publication, to mention but a few. *Simple Knowledge Organization System (SKOS)* is aimed to represent the structure of organization systems such as thesauri and taxonomies, since they share many similarities. It is widely used within the Semantic Web context, since it provides an intuitive language and can be combined with formal representation languages such as the Web Ontology Language (OWL). *SKOS XL* works as an extension of SKOS to represent lexical information [23].

With regard to multilingualism in ontologies, *Linguistic Information Repository (LIR)* was proposed as model for ontology localisation: it grants the localisation of the ontology terminological layer, without modifying the ontology conceptualisation. LIR allows enriching ontology entities with the linguistic information necessary for the localisation and cultural adaptation of the ontology [24].

¹⁷ <http://www.akomantoso.org/>

¹⁸ <https://www.oasis-open.org/committees/legalruleml/>

¹⁹ <https://github.com/RinkeHoekstra/lkif-core>

Another model intended for the representation of linguistic descriptions associated to ontology concepts is *Lexinfo* [20]. It contains a complete collection of linguistic categories. Currently, it is used in combination with other models such as *Ontolex* (described in the next paragraph), to describe the properties of the linguistic objects that describe ontology entities. Other repositories of linguistic categories are *ISOcat*²⁰, *OLiA*²¹ or *GOLD*²².

The *Lexicon Model for Ontologies* or *lemon* [26] was especially created to represent lexical information in the Semantic Web, covering some needs that previous models did not. This model has evolved in the context of a W3C Community Group into *lemon-Ontolex* first, now better known as *Ontolex*²³. In this model, linguistic descriptions are as well separated from the ontology, and point to the corresponding concept in the ontology. The structure of this model is divided into a core set of classes and different modules containing various types of linguistic information that range from morpho-syntactic properties of lexical entries, lexical and terminological variation and translation, decomposition of phrase structures, syntactic frames and mappings to the ontological predicates, and morphological decomposition of lexical forms. Linguistic annotations such as data categories and linguistic descriptors are not captured in the model but referred to by pointing to models that contain them (see *LexInfo* model above).

4.2 STRATEGY FOR THE HARMONISATION OF DATA MODELS IN LYNX

The LKG needs a uniform collection of data models in order to integrate heterogeneous resources. The definition of these data models will be provided in Deliverable 2.4.

In order to select the data models, a simultaneous top down and bottom up approaches will be conducted, as illustrated by Figure 11. A parallel work is carried out, where in the one hand a top down approach is conducted, extracting a list of formats, vocabularies and ontologies which can be chosen to satisfy the functional requirements of the pilots, whereas in the other hand a bottom up approach is followed, exploring every possible format, vocabulary or ontology of interest, with special attention to the most widely spread ones.

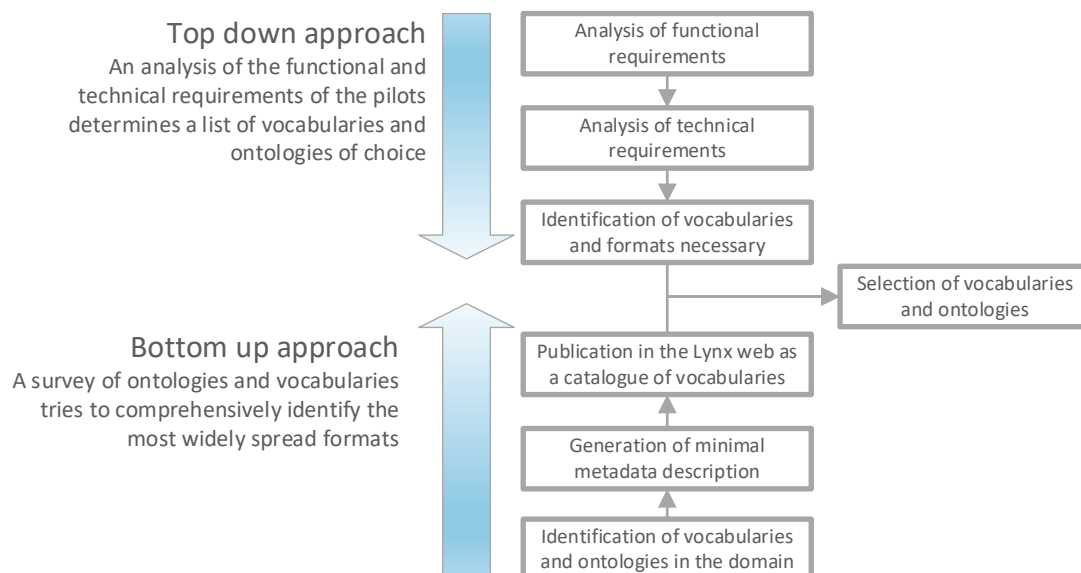


Figure 10. Strategy for the selection of data models in Lynx

²⁰ <http://www.iso.org/sites/dcr-redirect/dcr.html>

²¹ <http://www.acoli.informatik.uni-frankfurt.de/resources/olia/>

²² <http://linguistics-ontology.org/>

²³ https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

5 THE MULTILINGUAL LEGAL KNOWLEDGE GRAPH

As stated in the introduction, a secondary goal of this document is to define the Legal Knowledge Graph that will be developed during the Lynx project with a linguistic regulatory Linked Open Data Cloud.

5.1 SCOPE OF THE LEGAL KNOWLEDGE GRAPH

The amount of legal data made accessible either in open or under payment modalities by legal information providers can be hardly imagined. Lexis Nexis claimed²⁴ to have 30 Terabytes of content, WestLaw accounted for more than 40,000 *databases*. Their value can be roughly estimated: as of 2012, the four big players (WestLaw, Lexis Nexis, Wolters Kluwer and Bloomberg Legal) totalled about \$10,000M in revenues. Language data (e.g. resources with any kind of linguistic information) belongs to a much smaller domain, but still, unmanageable as a whole.

The Lynx project is interested in a small fraction of the information belonging to these domains. In particular, Lynx is in principle interested only in using the data necessary to provide the compliance services described in the pilots. Data of interest is regulatory data (legal and standards-related) and language data (to cover the multilingual aspects of the services). The intersection of these domains is of the utmost interest and Lynx will try to comprehensively identify every possible open dataset in this core category. These ideas are represented in Figure 4.

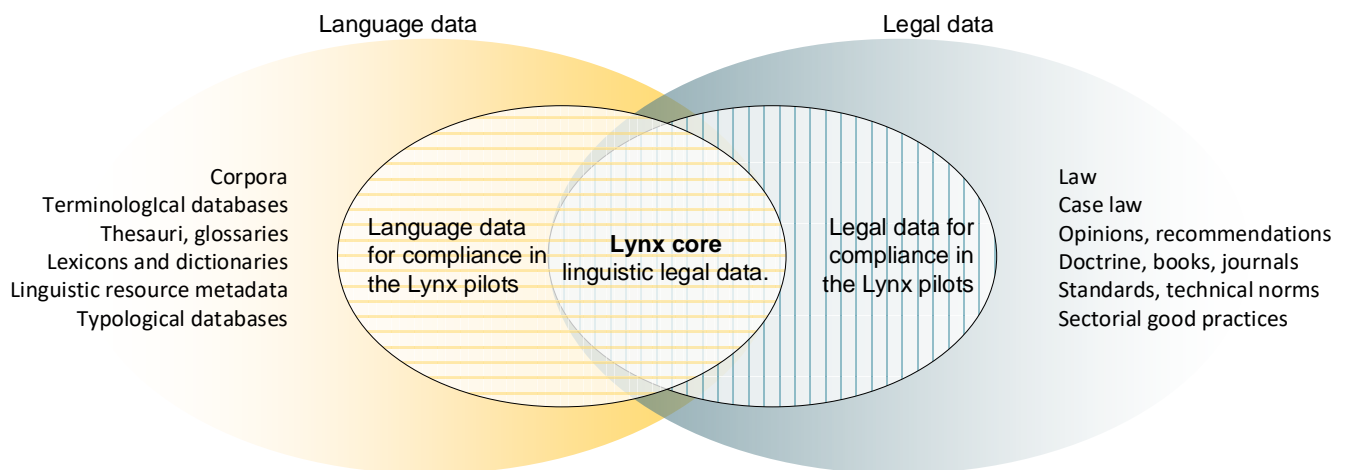


Figure 11. Scope of the multilingual Legal Knowledge Graph

The definitions of both *language data* and *regulatory data* are indeed fuzzy, but flexible as to introduce data of many different kinds whenever necessary (geographical data, user information, etc.). Because data in the Semantic Web is indissociable from the data models, and data models are accessed in the same manner as data is, ontologies and vocabularies are part of the LKG as well. Moreover, any kind of metadata (describing documents, standards etc.) is also part of the LKG, as well as the description of the entities producing the documents (courts, users, jurisdictions). In order to provide the compliance services, and with different degree of interest, both primary and secondary law are of use, and any relevant document in a wide sense may become part of the Legal Knowledge Graph. This is illustrated in Figure 5.

²⁴ Welcome to LexisNexis Legal & Professional". Lexisnexis.com. 2014-03-19.

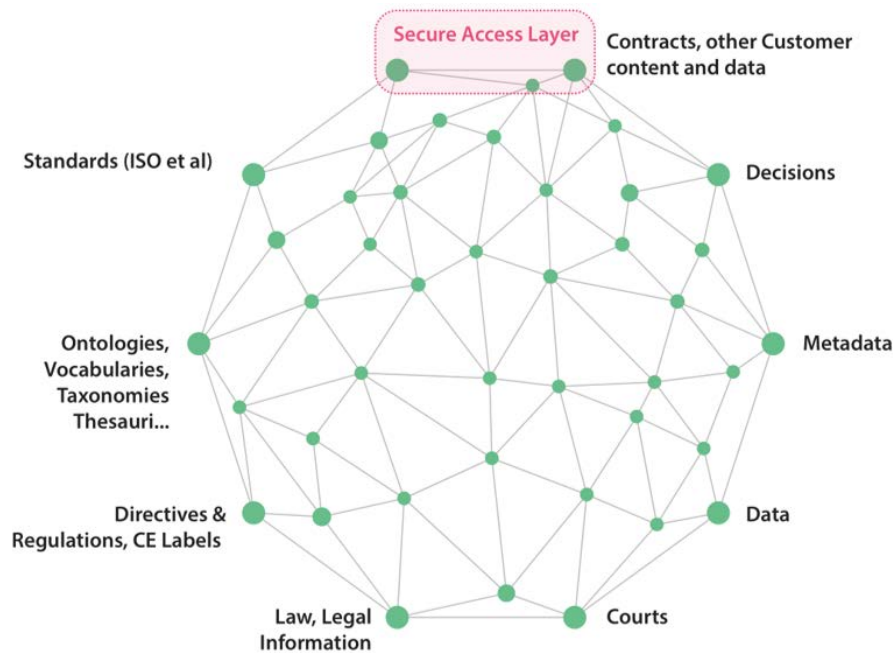


Figure 12. Types of information in the Legal Knowledge Graph

5.1 KNOWLEDGE GRAPHS

In the realm of Artificial Intelligence, a knowledge graph is a data structure to represent information, where entities are represented as nodes, their attributes as node labels and the relationship between entities are represented as edges. Knowledge graphs such as Google's²⁵, Freebase [2] and WordNet [3] turn data into knowledge, and they have become important resources for many AI and NLP applications such as information search, data integration, data analytics, question answering or context-sensitive recommendations.

Large knowledge graphs include millions of concepts and billions of relationships. For example, DBpedia describes about 30M entities connected through 10,000M relationships. Entities belong to classes described in ontologies. There are different manners of representing knowledge graphs, not the least important being the one using W3C specifications of the Semantic Web: RDF, RDFS, OWL. RDF data is accessible online in different forms: as file dumps, through a SPARQL endpoints or dedicated APIs or simply published online as Linked Data [4].

5.1.1 Legal Knowledge Graphs

In the last few years, a number of Legal Knowledge Graphs have been created in different applications. The MetaLex Document Server offers legal documents as versioned Linked Data [10], including Dutch national regulations. Finnish [9] and Greek [8] legislation are also offered as Linked Data.

The Publications Office of the EU maintains the central content and metadata CELLAR repository for storing official publications and bibliographic resources produced by the institutions of the EU [11]. The content of CELLAR, which includes EU legislation, is made publicly available by the Eur-Lex service and it offers also an SPARQL endpoint.

The FP7 EUCases project (2013-2015) offered European and national case law and legislation linked in an open data stack (<http://eucases.eu>).

²⁵ <https://www.google.es/intl/es/insidesearch/features/search/knowledge.html>

Finally, Openlaws offers a platform based on linked open data, open source software and open innovation processes [5][6][7]. Lynx will benefit from the expertise of Openlaws, which will be the preferred source for the data models, methods and algorithms. New H2020 projects in the area of data protection are also using semantic web technologies, such as the H2020 Special²⁶, devoted to ease the collection of user consents and represent policies as RDF or the H2020 Mirel²⁷ (2016-2019), with a network of experts to define a formal framework and to develop tools for mining and reasoning with legal texts, or e-Compliance, an FP7 project (2013-2016), focused on using semantic web technologies for regulatory compliance in the maritime domain.

5.1.2 Linguistic Knowledge Graphs

In the last few years, the language technology community has shaped the Linguistic Linked Open Data Cloud: the graph with those language resources available in RDF and published as Linked Data [16]. The graph represented in Figure 6, resembles the one of the Linked Data Cloud, but limited to the language domain.

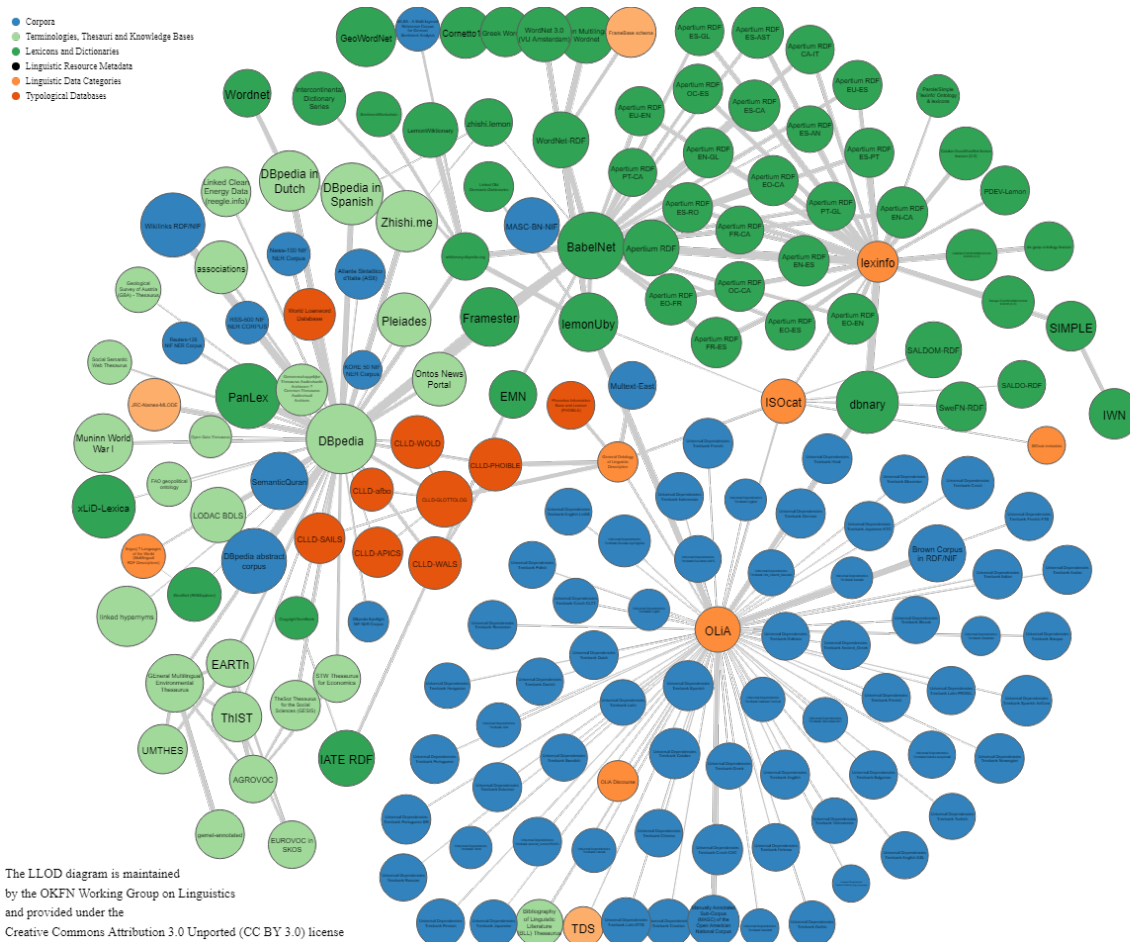


Figure 13. Linguistic Linked Open Data Cloud²⁸

A major resource contained in this graph is *DBpedia*, a vast network that structures data from Wikipedia and links them with other datasets available on the Web [3]. The result is published as Open Data

²⁶ <https://www.specialprivacy.eu>

²⁷ <http://www.mirelproject.eu>

²⁸ <http://linguistic-lod.org/llod-cloud>

available for the consumption of both humans and machines. Different versions of DBpedia exist for different languages.

Another core resource in the LOD Cloud is *BabelNet* [15], a huge multilingual semantic network, generated automatically from various resources and integrating the lexicographical information of *WordNet* and the encyclopaedic knowledge of Wikipedia. BabelNet also applies Machine Translation to get information from several languages. As a result, BabelNet is considered an encyclopaedic dictionary that contains concepts and named entities connected thanks to a great amount of semantic relations.

Wordnet, is one of the best known Linguistic Knowledge Graphs, since it is a large online lexical database that contains nouns, verbs, adjectives and adverbs in English [3]. These words are organised in sets of synonyms that represent concepts, known as *synsets*. WordNet uses these synonyms to represent word senses; thus, synonymy is WordNet's most important relation. Four additional relations are also used by this network: antonymy (opposing-name), hyponymy (sub-name), meronymy (part-name), troponymy (manner-name) and entailment relations. Other resources equivalent to WordNet have been published for different languages, such as EuroWordNet [29].

However, there are other semantic networks (considered linguistic knowledge graphs) that do not appear in the LOD Cloud but are also worth to mention. This is the case of *ConceptNet* [28], a semantic network designed to represent common sense and support textual reasoning about documents in the real word. It represents part of human experiences and tries to share this common-sense knowledge with machines. ConceptNet is often integrated with natural language processing applications to speed up the enrichment of AI systems with common sense [4].

5.1.3 The Lynx Multilingual Legal Knowledge Graph

Building on these previous experiences, we are in the position to define the Lynx Multilingual Legal Knowledge Graph.

The **Lynx Multilingual Legal Knowledge Graph (LKG)** is a knowledge graph using W3C specifications with the necessary information to provide multilingual compliance services. The Lynx LKG builds on previous initiatives reusing open data and will evolve adding new resources whenever needed to provide compliance services. The LKG preferred form of publication is Linked Data, although other access mechanisms will be provided.

REFERENCES

- [1] H2020 Programme Guidelines on FAIR Data Management in Horizon 2020
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
- [2] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). Freebase: a collaboratively created graph database for structuring human knowledge. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 1247-1250). ACM.
- [3] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.
- [4] Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International journal on semantic web and information systems*, 5(3), 1-22.
- [5] Wass, C., Dini, P., Eiser, T., Heistracher, T., Lampoltshammer, T. J., Marcon, G., ... & Winkels, R. (2013, February). OpenLaws. eu. In Proceedings of the 16th International Legal Informatics Symposium IRIS (Vol. 292, pp. 21-23).
- [6] Winkels, R. (2015). The OpenLaws project: Big Open Legal Data. In Proceedings of the International Legal Informatics Symposium (IRIS 2015) (pp. 189-196).
- [7] Lampoltshammer, T. J., Sageder, C., & Heistracher, T. (2015). The openlaws platform—An open architecture for big open legal data. In Proceedings of the 18th International Legal Informatics Symposium IRIS (Vol. 309, pp. 173-179).
- [8] Chalkidis, I., Nikolaou, C., Soursos, P., & Koubarakis, M. (2017). Modeling and querying greek legislation using semantic web technologies. In European Semantic Web Conference (pp. 591-606). Springer, Cham.
- [9] Frosterus, M., Tuominen, J., Wahlroos, M., & Hyvönen, E. (2013). The Finnish law as a linked data service. In Extended Semantic Web Conference (pp. 289-290). Springer, Berlin, Heidelberg.
- [10] Hoekstra, R. (2011). The MetaLex document server. In International Semantic Web Conference (pp. 128-143). Springer, Berlin, Heidelberg.
- [11] Francesconi, E., Küster, M. W., Gratz, P., & Thelen, S. (2015). The ontology-based approach of the publications office of the EU for document accessibility and open data services. In International Conference on Electronic Government and the Information Systems Perspective (pp. 29-39). Springer, Cham.
- [12] Boer, A., Hoekstra, R., Winkels, R., Van Engers, T., & Willaert, F. (2002). Metalex: Legislation in xml. *Legal Knowledge and Information Systems (Jurix 2002)*, 1-10.
- [13] Force, E. T. (2015). ELI: A Technical Implementation Guide. Publications Office of the European Union.
- [14] Hoekstra, R., Breuker, J., Di Bello, M., & Boer, A. (2007). The LKIF Core Ontology of Basic Legal Concepts. *LOAIT*, 321, 43-63.
- [15] Navigli, R., & Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In Proceedings of the 48th annual meeting of the association for computational linguistics (pp. 216-225). Association for Computational Linguistics.
- [16] Chiarcos, C., McCrae, J., Cimiano, P., & Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In *New Trends of Research in Ontologies and Lexical Resources* (pp. 7-25). Springer, Berlin, Heidelberg.
- [17] Gangemi, A. (2007). Design Patterns for Legal Ontology Constructions. *LOAIT*, 2007, 65-85.
- [18] Casellas, N. (2011). *Legal ontology engineering: Methodologies, modelling trends, and the ontology of professional judicial knowledge* (Vol. 3). Springer Science & Business Media.
- [19] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., & Ives, Z. (2007). DBpedia: A Nucleus for a Web of Open Data. *The semantic web*.
- [20] Cimiano, P., Buitelaar, P., McCrae, J., & Sintek, M. (2011). LexInfo: A Declarative Model for the Lexicon-Ontology Interface. *Web Semantics: Science, Services and Agents on the World Wide Web*.
- [21] Liu, H., & Singh, P. (2004). ConceptNet - a practical commonsense reasoning tool-kit. *BT technology journal*.
- [22] McCrae, J., Spohr, D., & Cimiano, P. (2011). Linking Lexical Resources and Ontologies on the Semantic Web with lemon. *Extended Semantic Web Conference*.
- [23] Miles, A., & Bechhofer, S. (2009). SKOS Simple Knowledge Organization System reference. Recuperado el 13 de 05 de 2018, de <https://www.w3.org/TR/skos-reference/>
- [24] Montiel-Ponsoda, E., de Cea, G. A., Gómez-Pérez, A., & Peters, W. (2011). Enriching ontologies with multilingual information. *Natural language engineering*, 17(3), 283-309.
- [25] Villazón-Terrazas, B., Vilches-Blázquez, L. M., Corcho, O., & Gómez-Pérez, A. (2011). Methodological guidelines for publishing government linked data. In *Linking government data* (pp. 27-49). Springer, New York, NY.

- [26] McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E. Spohr, D. & Wunner, T. (2012). Interchanging lexical resources on the Semantic Web. *Language Resources and Evaluation*, 46(4), 701-719.
- [27] Rodríguez Doncel, V., Gómez-Pérez, A., & Villata, S. (2014). A dataset of RDF licenses. In *Proc. of the 27th Int. Conf. on Legal Knowledge and Information System (JURIX)*, R. Hoekstra (Ed.), ISBN 978-1-61499-467-1, pp. 187-189, IOS Press. DOI 10.3233/978-1-61499-468-8-187
- [28] Liu, H., & Singh, P. (2004). ConceptNet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4), 211-226.
- [29] Vossen, P. J. T. M. (1997). EuroWordNet: a multilingual database for information retrieval.
- [30] Neubert, J. (2009). Bringing the "Thesaurus for Economics" on to the Web of Linked Data. LDOW, 25964.
- [31] Rodríguez-Doncel, V.; Casanovas, P. (2015). A Linked term bank of copyright-related terms. *Inn Legal knowledge and information systems*. 2015, p. 91-100. Amsterdam: IOS Press. DOI 10.3233/978-1-61499-609-5-91