



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

D2.4 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), Socorro Bernardos (UPM), Rebeca Varela (UAB), Patricia Martín-Chozas (UPM)
CONTRIBUTORS	Jorge González-Conejero (UAB), Elena Montiel-Ponsoda (UPM)
REVIEWERS	Rebeca Varela (UAB)
VERSION STATUS	V1 Draft
NATURE	EC Open Research Data Pilot
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.3236320
DATE	31/05/2018 (M18)



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
01	Departs from "D2.1 Initial Data Management Plan", a living document. Additions of new items in the ToC	06/05/2019	Víctor Rodríguez-Doncel, Patricia Martín Chozas (UPM)
02	Update of data models, data portal and ethical assessment.	17/05/2019	Víctor Rodríguez-Doncel, Socorro Bernardos (UPM), Rebeca Varela (UAB)
03	Update of data models	30/05/2019	Víctor Rodríguez-Doncel (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 is a second version of the Data Management Plan for Lynx. The document follows the template proposed by the EC (Section 2) and it 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, acting as a catalogue of compliance-related datasets. A strategy for the harmonisation of data models has also been given along with a description of data models of reference (Section 4) and a strategy for minting URIs (Section 5). Finally, a description of the Legal Knowledge Graph is made (Section 6). This document will be superseded by D2.8 *Final report of the data management activities* in November 2020.

TABLE OF CONTENTS

1	INTRODUCTION	8
2	DATA MANAGEMENT PLAN	9
2.1	DATA SUMMARY	9
2.2	FAIR DATA	10
2.3	ALLOCATION OF RESOURCES	16
2.4	DATA SECURITY	16
2.5	LEGAL, ETHICAL AND SOCIETAL ASPECTS	17
2.6	ASSESSMENT OF LEGAL, ETHICAL AND SOCIETAL IMPACT ASPECTS	18
2.6.1	Lynx methodology for the impact assessment	18
2.6.2	General ethical and societal aspects: Ethical and societal impact assessment	19
2.6.2.1	Ethical impact assessment	19
2.6.2.2	Societal impact assessment	20
3	CATALOGUE OF DATASETS	22
3.1	METHODOLOGY FOR CATALOGUING DATASETS	22
3.1.1	Template for data description	23
3.1.2	Lynx Data Portal	24
3.2	TRANSFORMATION OF RESOURCES.....	26
3.3	CATALOGUE OF DATASETS.....	27
3.3.1	Datasets in the regulatory domain.....	27
3.3.2	Datasets in the language domain.....	28
4	DATA MODELS.....	33
4.1	INTRODUCTION.....	33
4.1.1	Existing data models in the regulatory domain	33
4.1.2	Data models in the linguistic domain.....	33
4.2	LYNX DATA MODELS	34
4.2.1	Strategy for the harmonisation of data models.....	34
4.2.2	Definition of Lynx Documents.....	35
4.2.3	Lynx Documents with metadata	36
4.2.4	Lynx Documents with structuring information	37
4.2.5	Lynx document with annotations	39
4.2.6	List of recommended metadata fields and their representation.....	40
5	URI MINTING POLICY.....	42
5.1	BACKGROUND.....	42
5.2	ALTERNATIVE URI MINTING STRATEGIES.....	42
5.2.1	Structured, non-opaque URIs.....	43
5.2.2	Opaque URIs.....	43
5.3	LYNX URI MINTING STRATEGY	44
6	THE MULTILINGUAL LEGAL KNOWLEDGE GRAPH	45
6.1	SCOPE OF THE LEGAL KNOWLEDGE GRAPH.....	45

6.2	KNOWLEDGE GRAPHS.....	46
6.2.1	Legal Knowledge Graphs.....	47
6.2.2	Linguistic Knowledge Graphs	47
6.2.3	The Lynx Multilingual Legal Knowledge Graph	49
	ANNEX I. JSON-LD CONTEXT FOR A LYNX DOCUMENT	52

TABLE OF FIGURES

Figure 1. Schematic description of the Multilingual Legal Knowledge Graph for Compliance	8
Figure 2. Lynx public deliverable at Zenodo.....	12
Figure 3. Deliverables on the Lynx website.....	12
Figure 4. A catalogue of relevant ontologies and vocabularies	15
Figure 5. Datasets in the LKG and out of it.....	22
Figure 6. Screenshot of the Lynx Data Portal	25
Figure 7. Usual activities for publishing linked data. Figure taken from [25].....	26
Figure 8. Structure of Regulatory Datasets	28
Figure 9. Datasets represented by domain.	32
Figure 11. Strategy for the selection of data models in Lynx	35
Figure 12 Original documents and Lynx Documents.....	35
Figure 12 Elements in a Lynx Document	36
Figure 13 Simple example of Lynx Document (JSON-LD)	36
Figure 13 Simple example of Lynx Document (Turtle)	37
Figure 13 Example of Lynx Document with metadata	37
Figure 13 Example of Lynx Document with metadata (Turtle)	37
Figure 13 Example of Lynx Document with language tag (JSON-LD)	37
Figure 18 Example of Lynx Document with structure (JSON-LD)	38
Figure 19 Simple example of Lynx Document (Turtle)	38
Figure 20 UML class diagram representation of Lynx document and Lynx document part.....	39
Figure 20 Annotated Lynx Document (JSON LD).....	39
Figure 19 Annotated Lynx Document (Turtle).....	40
Figure 18. Scope of the multilingual Legal Knowledge Graph.....	45
Figure 19 Lynx LKG and LKGs.....	46
Figure 20. Types of information in the Legal Knowledge Graph	46
Figure 21. Linguistic Linked Open Data Cloud.....	48

LIST OF TABLES

Table 1. Fields describing a data asset	23
Table 2. Fields describing a resource associated to a data asset	24
Table 3. Initial set of resources gathered.....	31
Table 4 List of recommended metadata fields and their representation	40
Table 5 List of some NIF-related properties and their values	41
Table 5 Prefixes used in this document.....	41
Table 6. URI patterns for different resources.....	44

ACRONYMS

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

1 INTRODUCTION

This document is the Data Management Plan (DMP) of the project. The final version of this document will be available as “D2.8 Final report of the data management activities” in M36. This document is complemented by “D7.2 IPR and Data Protection Management”, which was delivered in M6.

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 (a European project supporting Open Science) granting its long term preservation. The description of the most relevant datasets for compliance have been published in a Lynx Data Portal, using the open source data portal CKAN software². 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 *Multilingual Legal Knowledge Graph for Compliance*, or Legal Knowledge Graph for short (Section 6), which is the backbone on when the Lynx services rest (Figure 1).

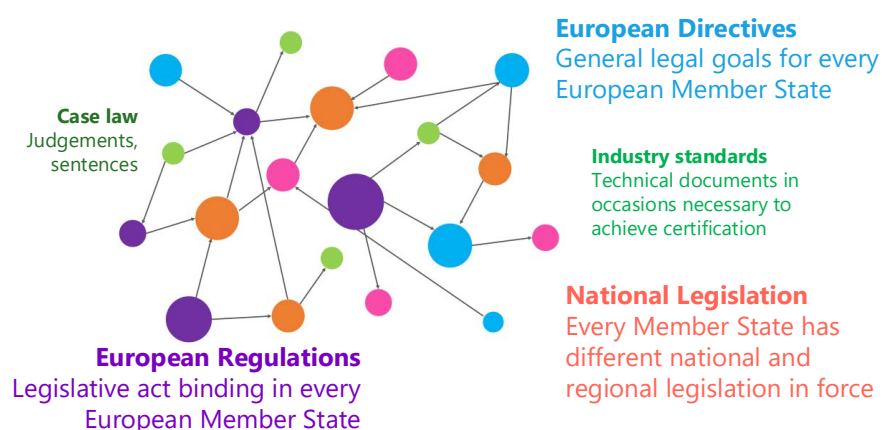


Figure 1. Schematic description of the Multilingual Legal Knowledge Graph for Compliance

¹

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

² <https://ckan.org/>

2 DATA MANAGEMENT PLAN

This Section is the Data Management Plan as of M18. 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 future versions of this document.

The EC promotes the access to and reuse of research data generated by Horizon 2020 projects through the Open Research Data Pilot. This project commit to the rules³ on open access to scientific peer reviewed publications and research data that beneficiaries have to follow in projects funded or co-funded under Horizon 2020 [33]. In particular:

- Lynx has developed and maintains keep up-to-date a Data Management Plan (this version is a snapshot of a continuously evolving document).
- Lynx has deposited the data in a research data repository –Zenodo. Lynx has a community in Zenodo, and CKAN provides a stable repository for data results. The data outcomes of the project live in CKAN.
- Lynx makes sure third parties can freely access, mine, exploit, reproduce and disseminate it – where applicable and not in conflict with any IPR considerations.
- Lynx has made clear what tools will be needed to use the raw data to validate research results – standard formats have been used for data at every moment.

The next sections and the questions are taken from the Horizon 2020 FAIR DMP template, which is recommended by the EU commission but voluntary.

2.1 DATA SUMMARY

1. Data summary	
a) What is the purpose of the data collection / generation and its relation to the objectives of the project?	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 is collected and integrated into a Legal Knowledge Graph, described in more detail in Section 6.
b) What types and formats of data will the project generate / collect?	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 4.
c) Will you re-use any existing data and how?	The core part of the LKG is created by reusing existing datasets, either copying them into the consortium servers (only if strictly needed) or using them directly from the sources.

³ <https://www.openaire.eu/what-is-the-open-research-data-pilot>

d) What is the origin of the data?

Although Lynx is 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 become 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.

e) What is the expected size of the data?

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.

f) To whom might the data be useful ('data utility')?

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

2.2 FAIR DATA

2. FAIR data

2.1 Making data findable, including provisions for metadata

a) Are the data produced and / or used in the project discoverable and identifiable?

Data is discoverable through a dedicated data portal (<http://data.lynx-project.eu>), further described in Section 3. Data assets will be identified with a harmonized policy to be defined in the forthcoming months. Research data may be linked to the corresponding publications and vice versa via their DOIs.

b) What naming conventions do you follow?

A specific URI minting policy has been defined in Section 5 to identify data assets.

c) Will search keywords be provided that optimize possibilities for re-use?

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

d) Do you provide clear version numbers?

Zenodo supports DOI versioning.

e) What metadata will be created?

Metadata records describing each dataset is downloadable as DCAT-AP entries in the CKAN. Assets in Zenodo have also metadata records.

2.2 Making data openly accessible

a) Which data produced and / or used in the project will be made openly available as the default?

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 irremediable 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 (`dct: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.

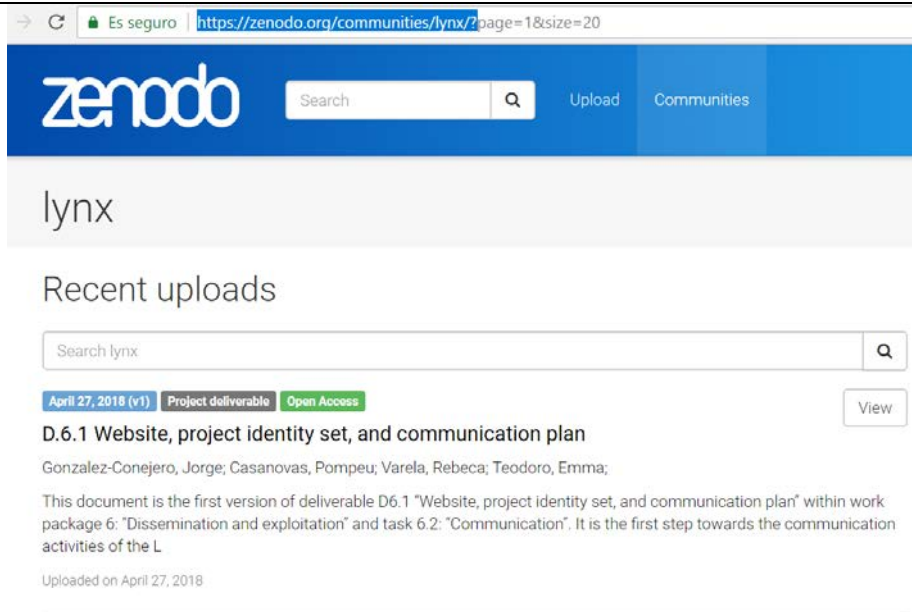
The Lynx project chose to participate in the Open Research Data Pilot (ORDP). Consequently, publishing as "open" the digital research data generated during the project is a contractual obligation (GA Art. 29.3). This provision does not include the pieces of data which are derivative of private data of 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 uses 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:

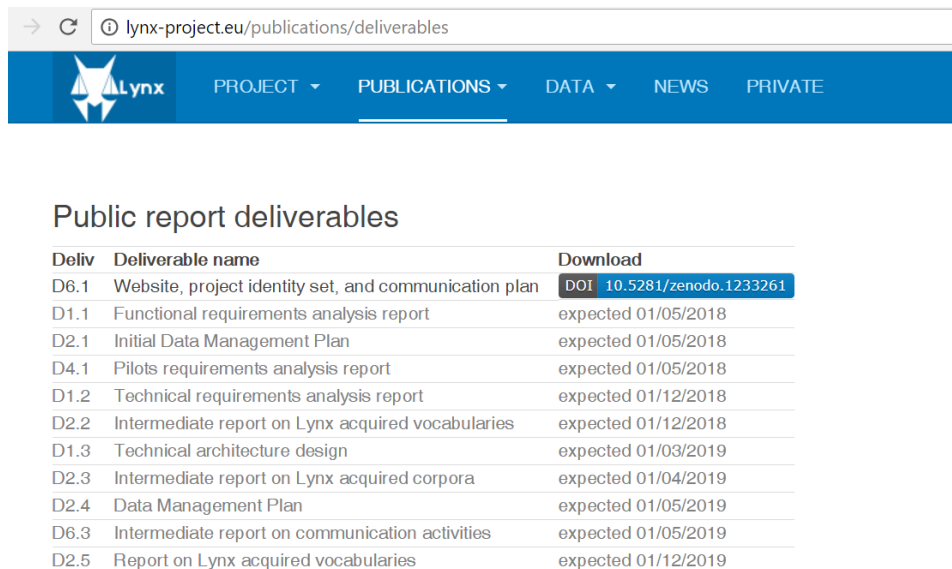
- *Public deliverables* will be available both at Zenodo and the Lynx website at <http://lynx-project.eu/publications/deliverables>. See Figure 2 and Figure 3.
- *Conference and Workshop presentations* may be published at Slideshare under the account <https://www.slideshare.net/LynxProject>.
- *Conference 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.



zenodo Search Upload Communities
 lynx
 Recent uploads
 Search lynx
 April 27, 2018 (v1) Project deliverable Open Access View
D.6.1 Website, project identity set, and communication plan
 Gonzalez-Conejero, Jorge; Casanovas, Pompeu; Varela, Rebeca; Teodoro, Emma;
 This document is the first version of deliverable D6.1 "Website, project identity set, and communication plan" within work package 6: "Dissemination and exploitation" and task 6.2: "Communication". It is the first step towards the communication activities of the L
 Uploaded on April 27, 2018

Figure 2. Lynx public deliverable at Zenodo.



lynx-project.eu/publications/deliverables

PROJECT PUBLICATIONS DATA NEWS PRIVATE

Public report deliverables

Deliv	Deliverable name	Download
D6.1	Website, project identity set, and communication plan	DOI: 10.5281/zenodo.1233261
D1.1	Functional requirements analysis report	expected 01/05/2018
D2.1	Initial Data Management Plan	expected 01/05/2018
D4.1	Pilots requirements analysis report	expected 01/05/2018
D1.2	Technical requirements analysis report	expected 01/12/2018
D2.2	Intermediate report on Lynx acquired vocabularies	expected 01/12/2018
D1.3	Technical architecture design	expected 01/03/2019
D2.3	Intermediate report on Lynx acquired corpora	expected 01/04/2019
D2.4	Data Management Plan	expected 01/05/2019
D6.3	Intermediate report on communication activities	expected 01/05/2019
D2.5	Report on Lynx acquired vocabularies	expected 01/12/2019

Figure 3. Deliverables on the Lynx website

b) How will the data be made accessible (e.g. by deposition in a repository)?

Data descriptions (metadata) are accessible through a dedicated data portal, hosted in Madrid and available under <http://data.lynx-project.eu>. Data from small datasets is also 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 Retele⁴ resources in Spanish language, ELRC-SHARE⁵ in general and others to be identified. In addition, the cooperation with the CEF eTranslation⁶ TermBank project will be considered, in view of sharing terminological domain-specific resources.

c) What methods or software tools are needed to access the data?

Relevant datasets whose license is liberal is 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/>).

d) Is documentation about the software needed to access the data included?

Yes, tools to visualize RDF and JSON are given.

e) Is it possible to include the relevant software (e.g. in open source code)?

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.

f) Where will the data and associated metadata, documentation and code be deposited?

Lynx uses a private source code repository (<https://gitlab.com/superlynx>). Open data is 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.

g) Have you explored appropriate arrangements with the identified repository?

Zenodo already foresees the existence of H2020 consortiums.

h) If there are restrictions on use, how will access be provided?

All metadata in Zenodo are openly accessible as soon as the record is published, even if there are restrictions like an embargo on the publications or research data themselves. In this way, it is always possible to contact the author of the data to ask for individual agreements on accessing the data, even if there are general restrictions.

i) Is there a need for a data access committee?

⁴ <http://catalogo.retele.linkeddata.es/>

⁵ 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/>

⁶ 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.

As of today, there is no need for a Data Access Committee⁷.

j) Are there well described conditions for access (i.e. a machine readable license)?

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.

k) How will the identity of the person accessing the data be ascertained?

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

2.3 Making data interoperable

a) Are the data produced in the project interoperable?

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. Zenodo uses standard interfaces, protocols, metadata, etc. CKAN implements standard api access.

b) What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Specific data and metadata vocabularies will be defined throughout the entire project. An initial collection has already been edited and has been published at <http://lynx-project.eu/data2/data-models> (see also Figure 4).

c) Will you be using standard vocabularies for all data types present in your data set, to allow interdisciplinary interoperability?

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, Ontolex⁸ is standard in the language resources and technologies communities, whereas the ELI ontology⁹ (European Law Identifier) is standard in the European legal community.

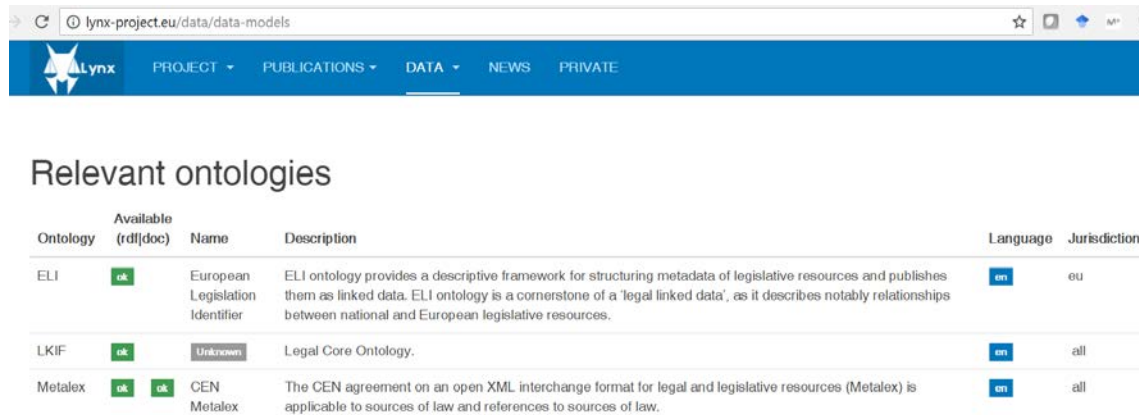
d) In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

⁷A Data Access Committee is a body of one or more named individuals who are responsible for data release to external requestors.

⁸<http://lemon-model.net/>

⁹ <http://publications.europa.eu/mdr/eli/>

If vocabularies or ontologies are further defined, they will be published online, documented and mapped to other standard ontologies. Figure 4 illustrates a possible visualization for the data models.



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

2.4 Increase data re-use (through clarifying licences)

a) How will the data be licensed to permit the widest re-use possible?

Data in Zenodo is openly licensed.

b) When will the data be made available for re-use?

Guidance: *If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.*

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

c) Are the data produced and / or used in the project useable by third parties, in particular after the end of 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.

d) How long is it intended that the data remains re-usable?

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.

e) Are data quality assurance processes described?

Only formal aspects of data quality are expected to be assured. In particular, the 5-stars¹⁰ paradigm is considered, and the data portal describes this quality level in due time.

2.3 ALLOCATION OF RESOURCES

3 Allocation of resources

a) What are the costs for making data FAIR in your project?

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. Zenodo is free. Maintaining the hosting for CKAN costs money, but this has been foreseen in the budget.

b) How will these be covered?

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.

c) Who will be responsible for data management in your project?

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

UPM is responsible for the Zenodo account, and must approve (*curate*) every upload.

d) Are the resources for long term preservation discussed?

Public deliverables and research data are being uploaded to Zenodo, which grants the long term preservation. A specific community has been created in Zenodo¹¹. Alternatively, if difficulties are found with Zenodo, datasets may also be uploaded to Figshare¹² or B2Share¹³ where a permanent DOI is retrieved. Other sites such as META-SHARE, ELRC-SHARE or the European Language Grid may be considered in addition to grant long term preservation and maximize the impact and dissemination.

2.4 DATA SECURITY

4 Data security

a) Is the data safely stored in certified repositories for long term preservation and curation?

UPM is physically storing data on their servers: webpage, files and data in the Nextcloud system, the CKAN data catalogue and mailing lists. Source code is hosted at Gitlab on a Dutch data center.

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.

b) What provisions are in place for data security?

¹¹<https://zenodo.org/communities/lynx/>

¹²<https://figshare.com/>

¹³<https://b2share.eudat.eu/>

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

5 Ethical aspects

a) Are there any ethical or legal issues that can have an impact on data sharing?

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 May 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 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.

Ethical and societal 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.

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. An more detailed internal assessment of the Legal, Ethical and Societal impact of this project is made in Section 2.6.

Finally, the Lynx websites will try to comply with the W3C recommendations on accessibility, such as the Web Content Accessibility Guidelines (WCAG) 2.0 –which covers a wide range of recommendations for making Web content more accessible.

b) Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Whenever the operation of the pilos start, pilot leaders will report these consent documents.

2.6 ASSESSMENT OF LEGAL, ETHICAL AND SOCIETAL IMPACT ASPECTS

2.6.1 Lynx methodology for the impact assessment

The Lynx strategy for dealing with legal, ethical and societal aspects was initially included in *D2.1 Initial Data Management Plan* and *D7.2 IPR and Data Protection Management Documents*. The main issue identified as posing potential risks in terms of ethical, legal and societal impact was the potential affection of some Human Rights, and in particular, the right to privacy and data protection. To manage this risks the Consortium put in place a series of measures as a result of the Initial Recommendations.

At this stage of the project, as part of the Ongoing Monitoring devised in Paragraph 3.3.4 of D7.2, the UAB partner has proceeded to review the status of implementation of the risk management strategy. Furthermore an ethical and societal impact assessment has been conducted to verify that no other issues have arisen now that the project has advanced in the development of the Lynx solution.

Paragraph 2.2 below contains the Ethical and Societal impact assessment. This assessment has been conducted following the methodology developed by the H2020 e-SIDES project.¹⁴ In particular, Deliverable 2.2. of the e-SIDES project contains a list of ethical, legal societal and economic issues of Big Data technologies. This list has been verified against the Lynx project, explaining how Lynx deals with avoiding each one of the issues on the lists.

Paragraph 2.3 presents the review of the status of implementation of the Initial Recommendations. The original strategy for the management of privacy and data protection presented in *D7.2 IPR and Data Protection Management Documents*, included a two-fold perspective: recommendations for the requirements elicitation techniques to be deployed in Tasks 1.1. and 4.1., and recommendations for the Lynx Solution. Since then, Tasks 1.1. and 4.1. have finished and been reported in the corresponding deliverables (*D1.1 Functional requirements analysis report* and *D4.1 Pilots requirements analysis report*). Therefore, an update is necessary only in relation to the recommendations for the Lynx solution. Below we have included a review of the status of implementation of each of this recommendations at this

¹⁴ <https://e-sides.eu/e-sides-project>

stage of the project, as well as the indication of whether there are still some concerns related to some of them, in the form of mid-term recommendations.

2.6.2 General ethical and societal aspects: Ethical and societal impact assessment

2.6.2.1 Ethical impact assessment

- **Human welfare:** Discrimination of humans by big data-mediated prejudice can occur. Detrimental implications can emerge in the contexts of employment, schooling or travelling by various forms of big data-mediated unfair treatment of citizens.

Lynx: Personal data is not the type of data relevant for Lynx. Lynx integrates and links heterogeneous compliance data sources including legislation, case law, standards and other private documents such as contracts. Within this sources personal data may be contained. However, personal data *per se* is not analysed or processed in order to extract patterns, trends, decisions or connexions related to humans and human behaviour. Therefore Lynx will not impact in human welfare.

- **Autonomy:** Big data-driven profiling practices can limit free will, free choice and be manipulative in raising awareness about, for instance, news, culture, politics and consumption.

Lynx: Lynx does not entail automated decision making nor profiling, therefore autonomy is preserved.

- **Non-maleficence:** Non-transparent data reuse in the world of big data are vast and could have diverse detrimental effects for citizens. This puts non-maleficence as a value under pressure.

Lynx: The only foreseen reuse is that of personal data contained in case-law. However, this is openly available data and therefore can be used as part of the legal documents to provide compliance services. The reuse is therefore transparent and there is no risk of maleficence.

- **Justice (incl. equality, non-discrimination, digital inclusion):** Systematic unfairness can emerge, for instance, by generating false positives during preventative law enforcement practices or false negatives during biometric identification processes. (Such instances put constant pressure on the value of justice.)

Lynx: Lynx does not entail automated decision making nor profiling. The aim of Lynx is not to identify, characterize or give access to services to individuals.

- **Accountability (incl. Transparency):** For instance, in the healthcare domain patients or in the marketing domain consumers often do not know what it means and who to turn to when their data is shared via surveys for research and marketing purposes.

Lynx: As part of their Data Protection Policy, users of the Lynx technology should disclose to their clients that their personal data may be processed by the Lynx technology.

- **Trustworthiness (including honesty and underpinning also security):** Citizens often do not know how to tackle a big data-based calculation about them or how to refute their digital profile, in case there are falsely accused, e.g.: false negatives during biometric identification, false positives during profiling practices. Their trust is then undermined. The technology operators trust at the same time lies too much in the system.

Lynx: Lynx does not entail automated decision making nor profiling. It does not generate any type of conclusion on individuals or individual's behaviours.

- **Privacy:** Simply the myriad of correlations between personal data in big data schemes allows for easy identifiability, this can lead to many instances for privacy intrusion.
Lynx: Privacy and data protection implication of Lynx are described in further detail in the List of legal issues.
- **Dignity:** For instance, when revealing too much about a user, principles of data minimization and design requirements of encryption appear to be insufficient. Adverse consequences of algorithmic profiling, such as discrimination or stigmatization also demonstrate that dignity is fragile in many contexts of big data.
Lynx: Lynx does not entail automated decision making nor profiling, therefore autonomy is preserved.
- **Solidarity:** Big data-based calculations in which commercial interests are prioritized rather than non-profit- led interests, are examples of situations in which solidarity is under pressure. For instance, immigrants are screened by big data-based technologies, they may not have the legal position to defend themselves from potential false accusations resulting from digital profiling which can be seen as a non-solidary treatment.
Lynx: Lynx does not entail automated decision making nor profiling, therefore autonomy is preserved.
- **Environmental welfare:** Big data has rather indirect effects on the environment. But for instance, lithium mining for batteries is such. (But extending the life-expectancy of batteries and, for instance, using more sun-energy for longer-lasting batteries could be helpful.)

2.6.2.2 Societal impact assessment

- **Unequal access:** People are not in the same starting position with respect to data and data-related technologies. Certain skills are needed to find one's way in the data era. Privacy policies are usually long and difficult to understand. Moreover, people are usually not able to keep their data out of the hands of parties they don't want to have them.
Lynx: Lynx technologies are foreseen to be used by experienced, trained professionals. No personal data will be processed other than that contained in case law (openly available data) and private documents such as contracts (consent and privacy policy of user). The users of the Lynx technologies will make sure that their clients understand when their personal data may be processed by the Lynx technologies. However, it is important to remember that personal data per se will not be analysed or processed in order to extract patterns, trends, decisions or connexions related to humans and human behaviour.
- **Normalisation:** The services offered to people are selected on the basis of comparisons of their preferences and the preferences of people considered similar to them. People are put into categories whose characteristics are determined by what is most common. There is pressure toward conformity: the breadth of choices is restricted, and pluralism and individuality are pushed back.
Lynx: Lynx does not collect nor process any data on preferences and or characteristics of individuals. It is important to remember that personal data per se will not be analysed or processed in order to extract patterns, trends, decisions or connexions related to humans and human behaviour.
- **Discrimination:** People are treated differently based on different individual characteristics or their affiliation to a group. The possibility to reproach people with things they did years ago or to hold

people accountable for things they may do in the future affects people's behaviour. The data as well as the algorithms may be incorrect or unreliable, though.

Lynx: Lynx does not process any data on characteristics of individuals or behaviours. It is important to remember that personal data per se will not be analysed or processed in order to extract patterns, trends, decisions or connexions related to humans and human behaviour.

- **Dependency:** People depend on governmental policy for security and privacy purposes. It is considered a misconception that people can be self-governing in a digital universe defined by big data.

People choosing not to disclose personal information may be denied critical information, social support, convenience or selection. People also depend on the availability of services provided by companies. It is considered a risk if there are no alternatives to services that are based on the collection or disclosure of personal data.

Lynx: Lynx does not determine access to public services. As for private companies Lynx adds value to the service provided by their users to their clients. If a client rejects the processing of his/her personal data by the Lynx technologies the company will provide the service nonetheless, just without the improvement in efficiency

- **Intrusiveness:** Big data has integrated itself into nearly every part of people's online life and to some extent also in their offline experience. There is a strong sentiment that levels of data surveillance are too intimate but nevertheless many press 'agree' to the countless number of 'terms and conditions' agreements presented to them.

Lynx: Lynx does not request personal data from its users or third parties. It does not intrude individual's private lives.

- **Non-transparency:** Algorithms are often like black boxes to people, they are not only opaque but also mostly unregulated and thus perceived as incontestable. People usually cannot be sure who is collecting, processing or sharing which data. Moreover, there are limited means for people to check if a company has taken suitable measures to protect sensitive data.

Lynx: Lynx users will make sure that their privacy policy includes all the relevant information on the Lynx platform, the processing of personal data, the data controller and processors, etc. More information on this can be found in the list of legal issues.

- **Abusiveness:** Even with privacy regulations in place, large-scale collection and storage of personal data make the respective data stores attractive to many parties including criminals. Simply anonymised data sets can be easily attacked in terms of privacy. The risk of abuse is not limited to unauthorised actors alone but also to an overexpansion of the purposes of data use by authorised actors (e.g. law enforcement, social security).

Lynx: Lynx does not entail large-scale collection and storage of personal data. Minor amounts of personal data may be processed as part of some of the sources used by Lynx, namely case-law and private documents such as contracts.

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.

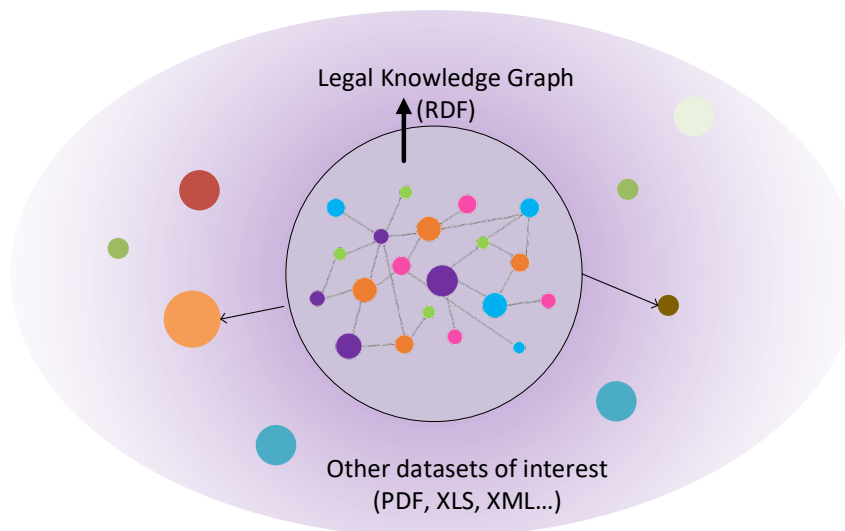


Figure 5. Datasets in the LKG and out of it

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 3.1.1).

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 Lynx partner, 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.

As the project progressed, it was required to add a new property to the first metadata selection reported in the D2.1, *Initial Data Management Plan*.

At this stage of the project, Lynx Data Portal collects a wide amount of resources; however, not all of them are included in the Legal Knowledge Graph. Such external resources are present in the portal since they can be useful in further processes. Therefore, a classification between those datasets in the LKG and the external resources is performed by the use of the Boolean parameter “Directly LKG Link”.

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.
Direct LKG Link [NEW]	indicates whether a dataset is directly represented in the LKG or if it is an external resource.

Table 1. Fields describing a data asset

The second set of metadata records, 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

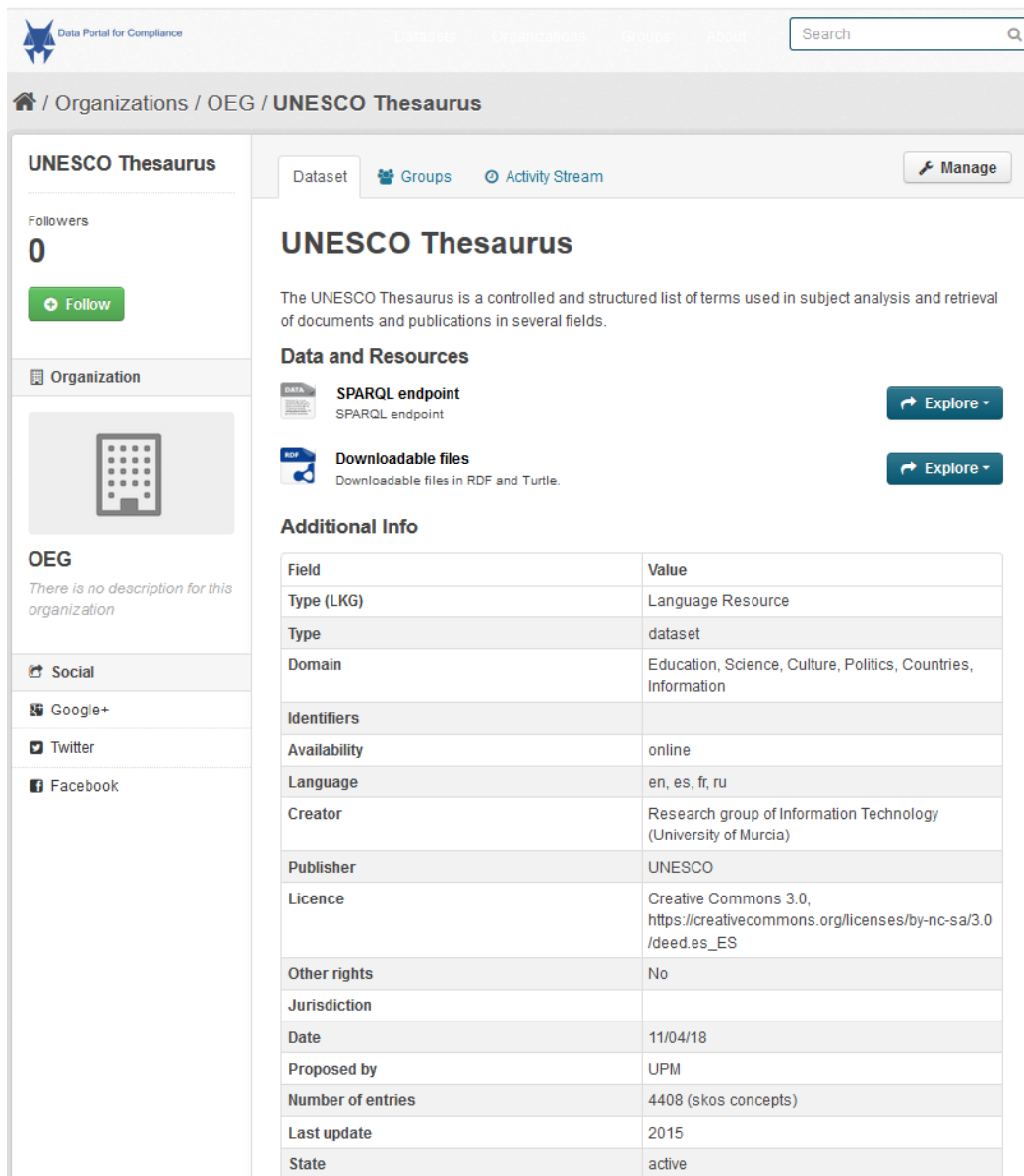
With the aim of publishing the metadata of the harvested datasets, a data portal has been made available under <http://data.lynx-project.eu>.

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 source¹⁵, and it has been deployed on the UPM servers using containerization technologies –Rancher¹⁶, 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 6 .

¹⁵ <https://github.com/ckan/ckan>

¹⁶ <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 67 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 online¹⁷. For example, the REST GET 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,
      "Distribuciones": "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 is shown in Figure 7.

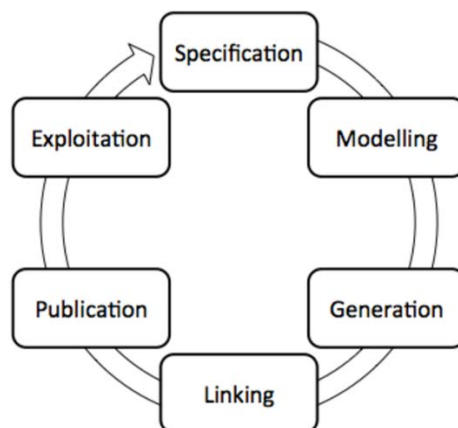


Figure 7. Usual activities for publishing linked data. Figure taken from [25].

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

- From unstructured text, extraction tools (PoolParty, OpenCalais, SketchEngine etc.) and dedicated harvesters to create resources in the LKG.
- From relational databases, technologies such as R2RML exist and its use is foreseen, but as of M18 no use of them has been made.
- For tabular data, Open Refine and similar tools have been used.

3.3 CATALOGUE OF DATASETS

This section contains the datasets catalogued as of M18.

3.3.1 Datasets in the regulatory domain

Within the initial version of this document (D2.1), three datasets in the regulatory domain were identified:

- 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.

As the project has progressed, many other datasets have been collected and a new structure has been accordingly defined. Pilot 1 changed the focus from “Data Protection” into “Contracts”. Therefore, harvested legal corpora has been organised accordingly: Contracts, Labour Law and Industrial Standards.

Regarding Pilot 1, contract corpora is provided by openlaws. Most of documents are in Austrian German containing personal data that is to be disclosed. Thus, this kind of files are private and not published in the Data Portal.

Nevertheless, openlaws will provide more contracts in future stages and some of them are expected to be bilingual, combining German and English information. Hence, they will need to be processed ad hoc.

Since Labour Law, Pilot 2, is a huge field itself, these specific corpora is, in turn, divided into three subtopics:

- Collective agreements, official documents about conditions of work for a specific sector at the same level as ordinary laws.
- Judgements, case law related to labour law in the different jurisdictions.
- Legislation, at European Union level and Member State Level.

Finally, each corpus is accordingly separated as per the four languages of the project: English, German, Spanish and Dutch. See Figure 8 to get a clear idea of the structure of Lynx datasets in the regulatory domain.

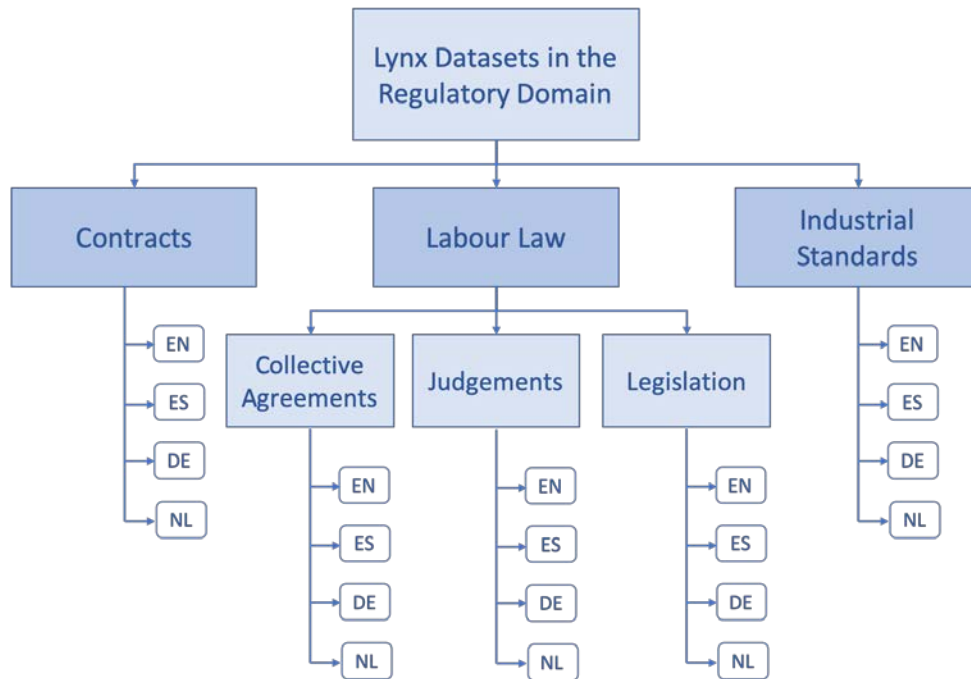


Figure 8. Structure of Regulatory Datasets

Finally, since Pilot 3, Industrial Standards, is led by DNV, most of the documents cover Dutch language. However, a few of them are also in English. Just like Pilot 1, at this moment, Industrial Standards corpus is for private use only.

3.3.2 Datasets in the language domain

Using the methodology described in Section 3.1, several sites and repositories have been surveyed. One of the sources of most interest for linguistic open data is the Linked Open Data Cloud¹⁸ or LOD cloud, due to its open nature and its adequate format as linked data or RDF. In particular, the Linguistic Linked Open Data Cloud¹⁹ 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 found in data portals like the European Data Portal, the Library of Congress or the Termcoörd 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,

¹⁸ <http://lod-cloud.net/clouds/lod-cloud.svg>

¹⁹ <http://linguistic-lod.org/>

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).
- SAII 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 9).

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 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 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.

4 DATA MODELS

4.1 INTRODUCTION

4.1.1 Existing 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 Ntoso²⁰ was initially funded by the United Nations to become some years later an OASIS specification as Legal RuleML²¹. 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 area²² [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].

²⁰<http://www.akomantoso.org/>

²¹ <https://www.oasis-open.org/committees/legalruleml/>

²² <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*²³, *OLiA*²⁴ or *GOLD*²⁵.

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*²⁶. 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 LYNX DATA MODELS

4.2.1 Strategy for the harmonisation of data models

Users of the LKG need a uniform collection of data models in order to integrate heterogeneous resources, which is initially provided in this Deliverable but which will be in constant maintenance until the end of the project.

In order to select the data models, a simultaneous top down and bottom up approaches has been conducted, as illustrated by Figure 10. A parallel work has been carried out, where in the one hand a top down approach has been 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 has been followed, exploring every possible format, vocabulary or ontology of interest, with special attention to the most widely spread ones.

²³ <http://www.iso.org/sites/dcr-redirect/dcr.html>

²⁴ <http://www.acoli.informatik.uni-frankfurt.de/resources/olia/>

²⁵ <http://linguistics-ontology.org/>

²⁶ https://www.w3.org/community/ontolex/wiki/Final_Model_Specification

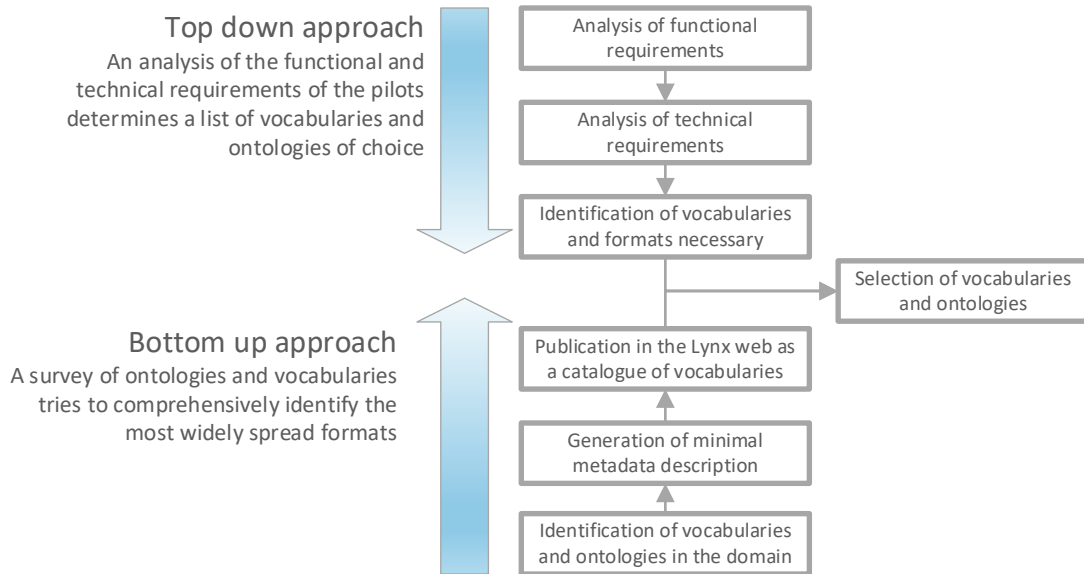


Figure 10. Strategy for the selection of data models in Lynx

4.2.2 Definition of Lynx Documents

The added value of the Lynx services revolves around a better processing of heterogenous, multilingual documents in the legal domain. Hence, the most important data structure is the *Lynx Document*. Lynx Documents may be grouped in *Collections*, and may be enriched with *Annotations*.

The main entities to deal with can be defined as follows:

- **Lynx Documents** are the basic information units in Lynx: identified pieces of text, possibly with structure, metadata and annotations. A **Lynx Document Part** is a part of Lynx documents.
- **Collections** are groups of Lynx Documents with any logical relation. There may be one collection per use case, per jurisdiction, etc.
- **Annotations** are enrichments of Lynx Documents, such as summaries, translation, recognized entities, etc.

Because most of AI algorithms dealing with documents focus on text -manipulation of images, videos or tables is less developed-, the essence of a Lynx Document is its text version. Thus, the key element in a Lynx Document is an identified piece of text. This document can be annotated with an arbitrary number of metadata elements (creation date, author, etc.), and eventually structured for a minimally attractive visual representation.

Original documents are transformed as represented in Figure 11: first, they are acquired by harvesters from their heterogeneous sources and formats, being structured and represented in a uniform manner. Then, they are enriched with annotations (such as named entities like persons, organisations, etc.).

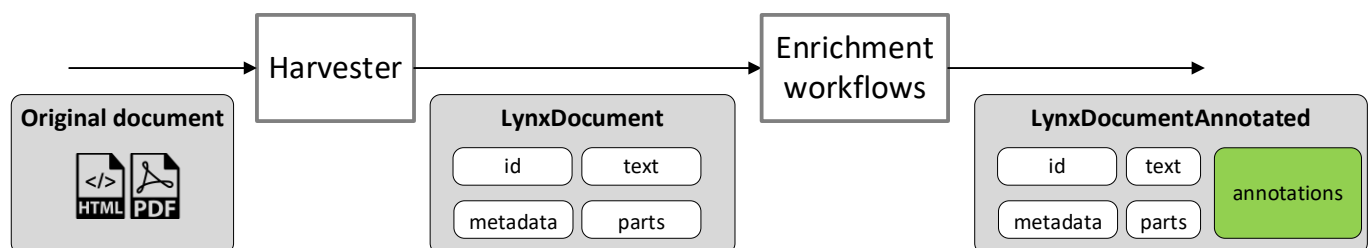


Figure 11 Original documents and Lynx Documents

The elements in a complete Lynx Document, with annotations, are depicted in Figure 12. Metadata is defined as a list of pairs attribute-values. Parts are defined as text fragments delimited by two offsets,

possibly with a title and a parent, so that they can be nested. Annotations also refer to text fragments delimited by two offsets, and describe in different manners such a fragment (e.g. ‘it refers to a Location which is Madrid, Spain’).

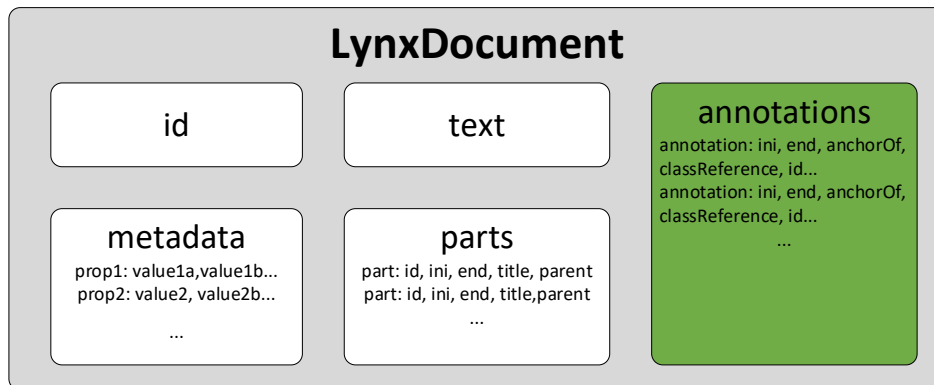


Figure 12 Elements in a Lynx Document

Lynx Documents can be serialized as RDF documents. Explicit support is given to its serialization as JSON-LD version 1.0, and a JSON-LD context is available at:

<http://lynx-project.eu/doc/jsonld/lynxdocument.json>

The format of a Lynx Document is shared among the three pilots and is valid for every type of documents. Refinements of this schema are possible –for example, even if an initial table of metadata records is described, new fields can be added as they become necessary for the pilot implementation.

4.2.3 Lynx Documents with metadata

The simplest possible Lynx Document as a JSON file is shown in the listing below.

```
{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
  "@id": "doc001",
  "@type": "http://lynx-project.eu/def/lkg/LynxDocument",
  "text" : "This is the first Lynx document, a piece of identified text."
}
```

Figure 13 Simple example of Lynx Document (JSON-LD)

The first line declares the context (@context), which describes how to interpret the rest of the JSON LD document. It references an external file. The second one (@id) declares the identifier of the element. The complete URI to identify the document is created from this string and also from the @base declared in the context. The @type declares what is the type of the document, and finally the text element represents the text of the document.

The text is not repeated in the fragments, in order to save space. Alternative transformations of this JSON structure are possible and recommended for every specific implementation need (e.g. OLS in Pilot 1).

The JSON-LD version can, however, be automatically converted into other RDF syntaxes. For example, the Turtle version of the same document follows.

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

```
<http://lkg.lynx-project.eu/res/doc001>
  a <http://lynx-project.eu/def/lkg/LynxDocument> ;
  rdf:value "This is the first Lynx document, a piece of identified text." .
```

Figure 14 Simple example of Lynx Document (Turtle)

Metadata is a collection of pairs property-list of values. This is better illustrated with the example below.

```
{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
  "@id": "doc002",
  "@type": "http://lynx-project.eu/def/lkg/LynxDocument",
  "text": "This is the second Lynx document.",
  "metadata": {
    "title": ["Second Document"],
    "subject": ["testing", "documents"]
  }
}
```

Figure 15 Example of Lynx Document with metadata

Which is rendered as RDF Turtle in the next listing.

```
@prefix lkg: <http://lkg.lynx-project.eu/def/lkg/> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://lkg.lynx-project.eu/res/doc002>
  a <http://lynx-project.eu/def/lkg/LynxDocument> ;
  lkg:metadata [
    dc:subject "testing", "documents";
    dc:title "Second Document"
  ] ;
  rdf:value "This is the second Lynx document." .
```

Figure 16 Example of Lynx Document with metadata (Turtle)

The language tag can be defined with the @language JSON-LD element, as an additional context element. This will make strings (RDF literals) to have the language tag set to Spanish.

```
{
  "@context": ["http://lynx-project.eu/doc/jsonld/lynxdocument.json", {"@language": "es"}],
  "@id": "doc003",
  "@type": "http://lynx-project.eu/def/lkg/LynxDocument",
  "text": "Un documento en español."
}
```

Figure 17 Example of Lynx Document with language tag (JSON-LD)

4.2.4 Lynx Documents with structuring information

Parts and structuring information can be included as shown in the next example. Parts are defined by the offset (begin and final character of the excerpt). They can be nested because they have a parent property and they can be possibly identified. Fragment identifiers can be built as described in the NIF specification²⁷. The example below shows an example of nested fragments, as Art. 2.1

```
{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
```

²⁷ <http://persistence.uni-leipzig.org/nlp2rdf/>

```

"@id": "doc004",
"@type": "http://lynx-project.eu/doc/lkg/LynxDocument",
"text": "Art.1 This is the fourth Lynx document. Art.2 This is the fourth Lynx document. Art 2.1.
Empty.",
"metadata": {
  "title": ["A document with parts."]
},
"parts": [
  {
    "offset_ini": 0,
    "offset_end": 39,
    "title": "Art.1"
  },
  {
    "@id": "http://lkg.lynx-project.eu/res/doc004/#offset_41_94",
    "offset_ini": 41,
    "offset_end": 94,
    "title": "Art.2"
  },
  {
    "offset_ini": 80,
    "offset_end": 94,
    "title": "Art.2.1",
    "parent": {
      "@id": "http://lkg.lynx-project.eu/res/doc004/#offset_41_94"
    }
  }
]
}
    
```

Figure 18 Example of Lynx Document with structure (JSON-LD)

In the following example, the Turtle RDF version is shown.

```

@prefix eli: <http://data.europa.eu/eli/ontology#> .
@prefix nif: <http://persistence.unileipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix lkg: <http://lkg.lynx-project.eu/def/lkg/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

<http://lkg.lynx-project.eu/res/doc004>
  a <http://lynx-project.eu/doc/lkg/LynxDocument> ;
  eli:has_part [
    nif:beginIndex 0 ;
    nif:endIndex 39 ;
    dc:title "Art.1"
  ], <http://lkg.lynx-project.eu/res/doc004/#offset_41_94>, [
    lkg:parent <http://lkg.lynx-project.eu/res/doc004/#offset_41_94> ;
    nif:beginIndex 80 ;
    nif:endIndex 94 ;
    dc:title "Art.2.1"
  ] ;
  lkg:metadata [ dc:title "A document with parts." ] ;
  rdf:value "Art.1 This is the fourth Lynx document. Art.2 This is the fourth Lynx document. Art 2.1. E
mpty."^^.

<http://lkg.lynx-project.eu/res/doc004/#offset_41_94>
  nif:beginIndex 41 ;
  nif:endIndex 94 ;
  dc:title "Art.2" .
    
```

Figure 19 Simple example of Lynx Document (Turtle)

Two classes suffice for representing Lynx Documents without annotations as UML objects (See Figure 20).

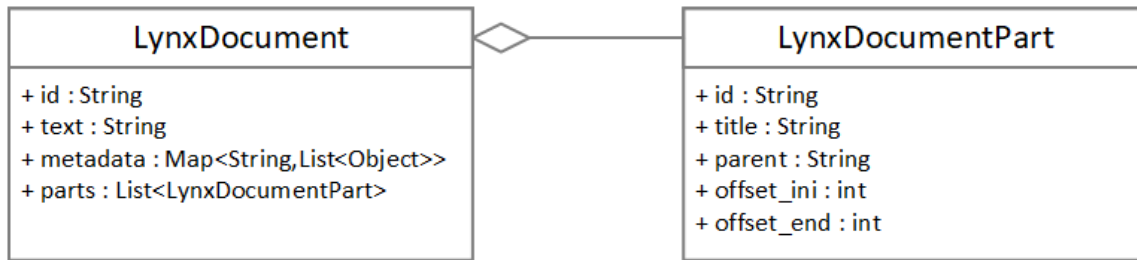


Figure 20 UML class diagram representation of Lynx document and Lynx document part.

4.2.5 Lynx document with annotations

Annotations are represented using NIF. The next example shows a Lynx Document with one annotation, highlighting the existence of a reference to London, which is a Location.

```

{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
  "@id": "doc005",
  "@type": "http://lynx-project.eu/doc/lkg/LynxDocument",
  "text": "I was born in London long time ago.",
  "metadata": {
    "title": [
      "An annotated document"
    ]
  }
},
"annotations": {
  "annotation": [
    {
      "@id": "http://lynx-project.eu/res/id000#offset_29_35",
      "@type": [
        "nif:String",
        "nif:RFC5147String"
      ],
      "anchorOf": "London",
      "offset_ini": "14",
      "offset_end": "20",
      "referenceContext": "http://lkg.lynx-project.eu/res/doc005",
      "taClassRef": "http://dbpedia.org/ontology/Location",
      "taIdentRef": "http://dbpedia.org/resource/London"
    }
  ]
}
}

```

Figure 21 Annotated Lynx Document (JSON LD).

The equivalent RDF Turtle excerpt follows, with the prefixes as above.

```

<http://lkg.lynx-project.eu/res/doc005>
  a <http://lynx-project.eu/doc/lkg/LynxDocument> ;
  lkg:metadata [ dc:title "An annotated document" ] ;
  lkg:annotations [ lkg:annotation <http://lynx-project.eu/res/id000#offset_29_35> ] ;
  rdf:value "I was born in London long time ago." .

<http://lynx-project.eu/res/id000#offset_29_35>
  a nif:String, nif:RFC5147String ;
  nif:anchorOf "London" ;
  nif:beginIndex 14 ;
  nif:endIndex 20 ;
  nif:referenceContext <http://lkg.lynx-project.eu/res/doc005> ;
  
```



```
itsrdf:taClassRef <http://dbpedia.org/ontology/Location> ;
itsrdf:taIdentRef <http://dbpedia.org/resource/London> .
```

Figure 22 Annotated Lynx Document (Turtle).

The use of `nif:annotationUnit` is optional, but useful for avoiding colliding annotations. The last line should be replaced then by the following excerpt. See more details on NIF on Table 6.

```
nif:annotationUnit [
  itsrdf:taIdentRef <http://vocabulary.semantic-web.at/CBeurovoc/C8553> .
] .
```

4.2.6 List of recommended metadata fields and their representation

Group	Property	Usage	RDF property
basic elements	id	Lynx identifier of the document	dct:identifier
	text	Text of the document	rdf:value
	parts	Parts of the document	eli:has_part
general	type	Type of document (legislation, case law, etc.)	dct:type
	rank	Sub-type of document (constitution, law, etc.)	eli:type_document
	language	Language of the document	dct:language
	jurisdiction	Jurisdiction using ISO	eli:jurisdiction
	wasDerivedFrom	Original URL if the document was extracted from the web	prov-o:wasDerivedFrom
	title	Title of the document	dct:title
	hasAuthority	Authority issuing the document	lkg:hasAuthority
	nick	Alternative names of the document	foaf:nick
	version	Consolidated, draft or bulletin	eli:version
	subject	Subjects or keywords of the document	dct:subject
identifiers	id_local	Local identifier (e.g. BOE-A-2019-1234)	eli:id_local
	identifier	Official identifier (e.g. ELI etc.)	dct:identifier
dates	first_date_entry_in_force	Date when enters into force	eli:first_date_entry_in_force
	date_no_longer_in_force	Date when repealed / expired	eli:date_no_longer_in_force
	version_date	Date of publication of the document	eli:version_date
mappings	hasEli	Official identifier (ELI, ECLI or equivalent)	lkg:hasEli
	hasPDF	Link to the PDF version	lkg:hasPDF
	hasDbpedia	Link to the equivalent dbpedia version	lkg:hasDbpedia
	hasWikipedia	Link to the equivalent wikipedia version	lkg:hasWikipedia
	sameAs	Equivalent document	owl:sameAs
	seeAlso	Related documents	rdfs:seeAlso
Internal	creator	Creators of the documents in Lynx (person or software)	dct:creator
	created	Date when created in Lynx (internal)	dct:created

Table 4 List of recommended metadata fields and their representation

Table 4 lists the recommended metadata fields and their representation and.

Element	Meaning	Values / example
itsrdf:taClassRef	Class of the annotated context	dbo:Person, dbo:Location, dbo:Organization, dbo:TemporalExpression
itsrdf:taldentRef	URL from external resource, such as DBPedia, Wikidata, Geonames, etc.	http://dbpedia.org/resource/London
itsrdf:taConfidence	Confidence	[0..1]
nif:summary	Summary	text

Table 5 List of some NIF-related properties and their values

Table 6 lists the prefixes used in this section.

Vocabulary	Prefix	URL
LKG Ontology	lkg	http://lkg.lynx-project.eu/def/
Dublin Core	dct	http://purl.org/dc/terms/
RDF	rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
European Legislation Ontology	eli	http://data.europa.eu/eli/ontology#
W3C Provenance Ontology	prov-o	https://www.w3.org/TR/prov-o/
Friend of a Friend Ontology	foaf	http://xmlns.com/foaf/spec/
NLP Interchange Format	nif	http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#
ITS 2.0 / RDF Ontology	itsrdf	http://www.w3.org/2005/11/its/rdf#

Table 6 Prefixes used in this document

5 URI MINTING POLICY

5.1 BACKGROUND

This section highlights the importance of choosing a good URI naming strategy.

URIs (or IRIs to be more precise, as per RFC 3987 on Internationalized Resource Identifiers) are the natural identifiers for resources in Lynx. An IRI is a sequence of Unicode characters (Unicode/ISO 10646) that can be used to mint identifiers that use a wider set of characters than the one defined for the URIs in RFC3986. Choosing good IRIs are key at least for the following reasons:

- Make humans easier to understand what is the resource in question. URIs with information on the identified resource and its nature (e.g. class) are easier for humans to remember and understand. URIs play the role of documenting ontologies and RDF resources in natural language. This is a misuse of URIs, and hardens the operation of resources in multilingual environments [32], but it is a common practice.
- Make easier the execution of automated tasks, such as resource mapping [34], information extraction [35] or natural language generation.

The W3C Consortium does not provide a normative recommendation on how to mint URIs. However, it was Tim Berners-Lee himself who as early as 1998 wrote in his article *Cool URIs don't change*²⁸ a list of good practices. Berners-Lee introduced the concept of *URI design*, which has proven to be a challenge for the Semantic Web community.

A second reference is the W3C Note *Common HTTP Implementation problems*²⁹, issued in the context of the Technical Architecture Group. This note elaborates the ideas of Berners-Lee's article, specifying some rules for choosing URIs: (i) Use short URIs as much as possible, (ii) Choose a case policy, (iii) Avoid URIs in mixed case, and (iv) As a case policy choose either "all lowercase" or "first letter uppercase". More recently, the *Best Practices for Publishing Linked Data*³⁰ specification issued by a W3C Working Group only recommended: 'A URI structure will not contain anything that could change' and that URIs shall be constructed 'to ensure ease of use during development'.

However, no more precise rules are given by W3C. Some recommend using hyphens, other claim a camel case policy for the local names suffices.

5.2 ALTERNATIVE URI MINTING STRATEGIES

Given that technically there is no clear recommendation on how to choose sets of URIs, two alternatives can be considered: either they are meaningful conveying information on the resource and its structure or they are meaningless because URIs should not be semantically interpreted. For example, given a certain sentence (judgment), one might consider including in the URI either:

- the title of the judgment
- the unique reference number for the judgment
- the internal record number in the Lynx databases

This section describes the pros and cons of these alternatives.

²⁸ <https://www.w3.org/Provider/Style/URI>

²⁹ <https://www.w3.org/TR/chips/>

³⁰ <https://www.w3.org/TR/ld-bp/#HTTP-URIS>

5.2.1 Structured, non-opaque URIs

Once the semantic web has grown mature and widely accepted, public institutions have also issued guides on URI minting, all of them leaning towards structured, non-opaque URIs. Most notably, the UK Cabinet Office published the recommendation “Designing URIs for the public sector”, the government in Netherlands issued a similar document³¹ and the Spanish one issued the Norma Técnica de Interoperabilidad contains a chapter for that “*Definición de un esquema de URI*”³². Finally, the European Commission published in 2014 the document *Towards a common approach for the management of persistent HTTP URIs by EU Institutions* to be used in the EU portals. These documents specify the path structure for URIs, establishing a clear separation of different types of data (a bus line is not a police office) and defining naming conventions.

These conventions emphasize the need of stability and scalability and specifically address the problem of managing large amounts of data on the Web.

Spanish case. For example, the Spanish norm defines the following URI pattern:

```
http://{base}/{carácter}/{sector}/{dominio}[.{ext}][#{concepto}]
```

If this strategy was applied to Lynx, *Base* would be `lynx-project.eu`; *character* would be either **def** (for ontologies and vocabularies), **kos** (for dictionary data, thesauri, taxonomies and other knowledge organization data), **cat** (for catalogue information) or **res** (for resources, such as a document); *sector* would be one word describing the domain sector (economy, justice-legislation, etc.). For Lynx this might be (standards/legislation/caselaw/doctrine/others); *dominio* would be the specific data type (e.g. Judgment) and *concept* would be the id of the resource (ext being the extension). An example of URI using the Spanish recommendation would be:

```
http://lynx-project.eu/res/caselaw/judgment/C23987
```

UK case. The UK recommendation is a well detailed document, which proposes the following URI pattern for documents: `http://{domain}/doc/{concept}/{reference}`. This would mean, applied to Lynx, having this URI for the same:

```
http://lynx-project.eu/doc/judgment/C23987
```

Holland case. The Dutch administration has adopted the URI pattern: `http://{domain}/{type}/{concept}/{reference}`. The example for Lynx would read:

```
http://lynx-project.eu/id/judgment/C23987
```

5.2.2 Opaque URIs

In the *Architecture of the World Wide Web*³³, which is a W3C Recommendation, we read “*Agents making use of URIs should not attempt to infer properties of the referenced resource*”. This recommendation is directly opposed to the strategies mentioned in the section before, and leads to enabling opaque URIs or at least with less semantics in it. For example, Tim Berners-Lee (1998) recommended not to put too much semantics in the URI, and not to bind URIs to some classification or topic (as one change the point of view).

³¹ http://www.pilod.nl/wiki/Bestand:D1-2013-09-19_Towards_a_NL_URI_Strategy.pdf

³² https://datos.gob.es/sites/default/files/20160726_guia_de_aplicacion_de_la_nti_reutilizacion_recursos_de_informacion_1.pdf

³³ <https://www.w3.org/TR/webarch/>

Opaque URIs can be generated automatically, are easier to manage and do not convey character encoding problems –in a project intrinsically multilingual such as Lynx, there should not be a cultural bias against languages with accents and other local characters (such as the Spanish Ñ).

An examples of opaque URI would be one chosen from the Spanish National Library (BNE) to identify the writer Miguel de Cervantes:

`http://datos.bne.es/persona/XX1718747`

From this URI, it can be inferred that it refers to a person, but no clue is given on which person. On the contrary, the dbpedia policy for cervantes hides the type of entity, but makes clear who is the referred writer:

`http://es.dbpedia.org/resource/Miguel_de_Cervantes`

5.3 LYNX URI MINTING STRATEGY

Considering the advantages and disadvantages examined in the previous section, Lynx has chosen the URI patterns as described in Table 7.

Type of resource	URI pattern
Ontology <i>Example</i>	<code>http://lkg.lynx-project.eu/def/{onto_id}</code> <code>http://lkg.Lynx-project.eu/def/core</code>
Ontology element <i>Example</i>	<code>http://lkg.lynx-project.eu/def/{onto_id}/{element}</code> <code>http://lkg.Lynx-project.eu/def/core/Document</code>
KOS (thesauri, terminologies) <i>Example</i>	<code>http://lkg.lynx-project.eu/kos/{kos_id}/{id}</code> <code>http://lkg.Lynx-project.eu/kos/contracts_terms/24232</code>
Resource <i>Example</i>	<code>http://lkg.lynx-project.eu/res/{id}</code> <code>http://lkg.Lynx-project.eu/res/23983</code>

Table 7. URI patterns for different resources

Advantages of this choice are:

- Problems derived from character encoding are solved
- Automatic generation of ids is possible, avoiding auto-increment derived problems
- Freedom of choice of ids for the different implementors
- No collision between resources sharing a name
- Relatively short URIs
- Easy scalability (no types of resources are predefined)
- Lynx URIs do not compete with official ones such as ELI or ECLI.

6 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.

6.1 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 claimed³⁴ 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 23.

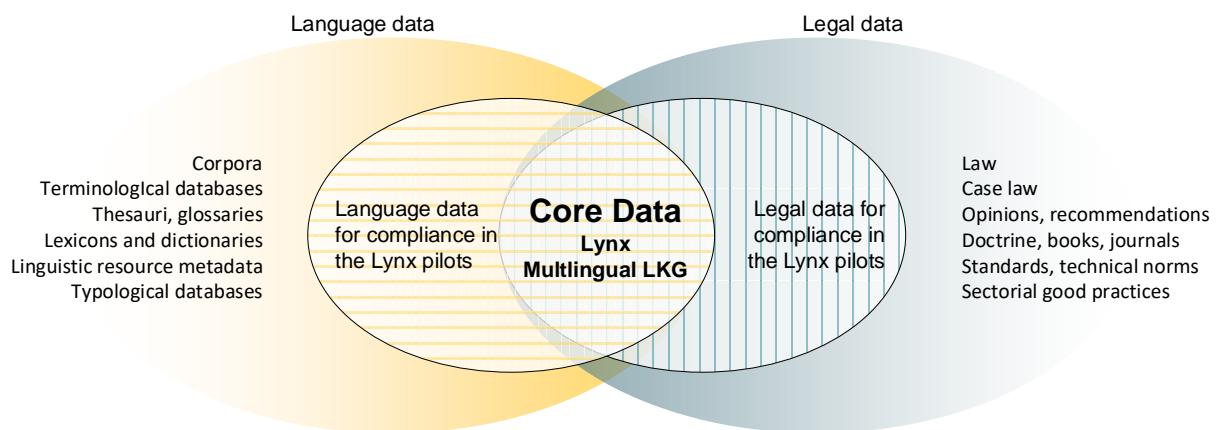


Figure 23. Scope of the multilingual Legal Knowledge Graph

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 inseparable 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 25.

³⁴Welcome to LexisNexis Legal & Professional". Lexisnexis.com. 2014-03-19.

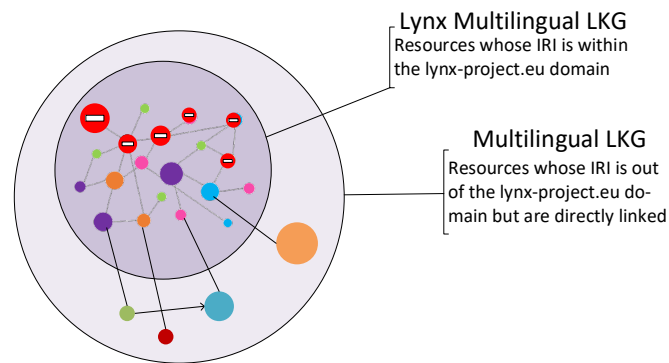


Figure 24 Lynx LKG and LKGs

We may define the Lynx Multilingual LKG as the set of entities and relations whose IRIs are within the <http://lynx-project.eu> top level domain. However, the resources in it are connected to other resources published by other entities, which constitute a wider LKG. Figure 24 represents this idea, together with the notion of private resources, which are only accessible to the authorized users (e.g. contracts only visible for the parties).

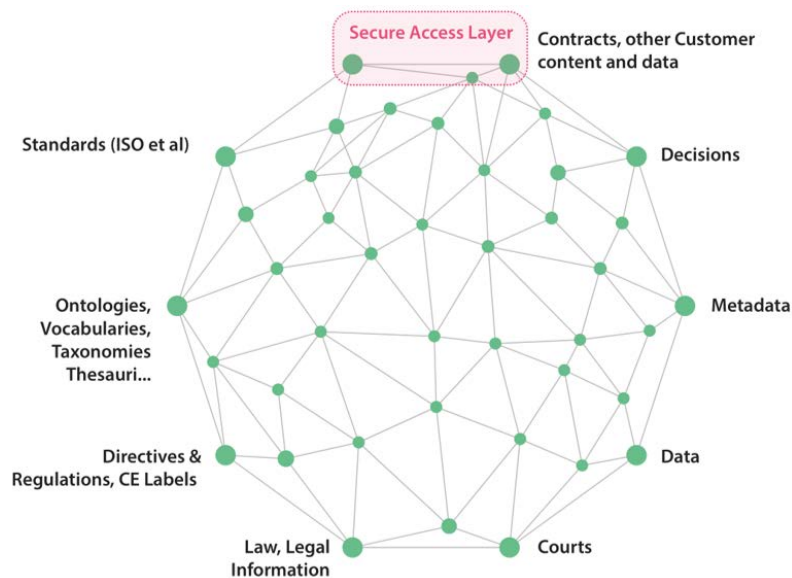


Figure 25. Types of information in the Legal Knowledge Graph

6.2 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's³⁵, 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

³⁵<https://www.google.es/intl/es/insidesearch/features/search/knowledge.html>

accessible online in different forms: as file dumps, through a SPARQL endpoints or dedicated APIs or simply published online as Linked Data [4].

6.2.1 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>).

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 Special³⁶, devoted to ease the collection of user consents and represent policies as RDF or the H2020 Mirel³⁷ (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.

6.2.2 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 26, resembles the one of the Linked Data Cloud, but limited to the language domain.

³⁶ <https://www.specialprivacy.eu>

³⁷ <http://www.mirelproject.eu>

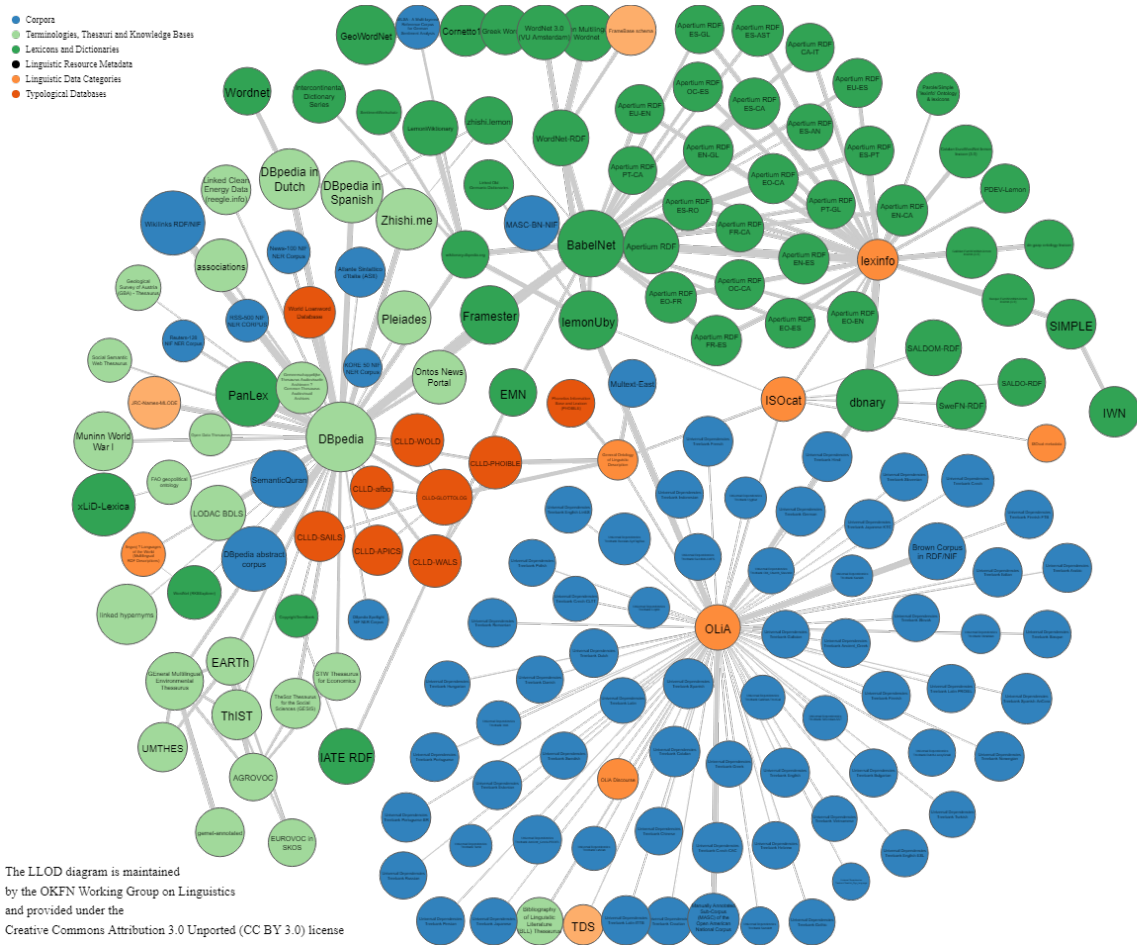


Figure 26. Linguistic Linked Open Data Cloud³⁸

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 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

³⁸ <http://linguistic-lod.org/llood-cloud>

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].

6.2.3 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
- [32] Montiel-Ponsoda, E., Vila Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E., & Gómez-Pérez, A. (2011). Style guidelines for naming and labeling ontologies in the multilingual web.
- [33] Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. v3.2, March 2017
- [34] Svab-Zamazal Ondrej and Svatek Vojtech. (2008). Analysing Ontological Structures through Name Pattern Tracking. In: EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management. Springer LNCS, pp. 213-228.
- [35] Müller, Hans-Michael, Eimear E. Kenny, and Paul W. Sternberg. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature PLoS Biol, 2004, 2, e309.

ANNEX I. JSON-LD CONTEXT FOR A LYNX DOCUMENT

This annex shows the content of the JSON-LD context for a Lynx Document as of M18. <http://lynx-project.eu/doc/jsonld/lynxdocument.json>.

```
{
  "@context": {
    "@base": "http://lkg.lynx-project.eu/res/",
    "nif": "http://persistence.unileipzig.org/nlp2rdf/ontologies/nif-core#",
    "itsrdf": "http://www.w3.org/2005/11/its/rdf#",
    "rdf": "http://www.w3.org/1999/02/22-rdf-syntax-ns#",
    "rdfs": "http://www.w3.org/2000/01/rdf-schema#",
    "skos": "https://www.w3.org/TR/skos-reference/",
    "dct": "http://purl.org/dc/terms/",
    "lkg": "http://lkg.lynx-project.eu/def/lkg/",
    "eli": "http://data.europa.eu/eli/ontology#",
    "text": "rdf:value",
    "Concept": "skos:Concept",
    "ConceptScheme": "skos:ConceptScheme",
    "prefLabel": "skos:prefLabel",
    "altLabel": "skos:altLabel",
    "notation": "skos:notation",
    "definition": "skos:definition",
    "broader": "skos:broader",
    "narrower": "skos:narrower",
    "inScheme": "skos:inScheme",
    "hasTopConcept": "skos:hasTopConcept",
    "topConceptOf": "skos:topConceptOf",
    "parts": "eli:has_part",
    "offset_ini": {
      "@id": "nif:beginIndex",
      "@type": "xsd:integer"
    },
    "offset_end": {
      "@id": "nif:endIndex",
      "@type": "xsd:integer"
    },
    "parent": {
      "@id": "lkg:parent",
      "@type": "@id"
    },
    "annotation": {
      "@id": "lkg:annotation",
      "@container": "@set"
    },
    "annotations": {
      "@id": "lkg:annotations",
      "@container": "@set"
    },
    "metadata": {
      "@id": "lkg:metadata",
      "@container": "@set"
    },
    "subject": {
      "@id": "dct:subject",
      "@container": "@set"
    },
    "title": {
      "@id": "dct:title",
      "@container": "@set"
    },
    "first_date_entry_in_force": {
      "@id": "eli:first_date_entry_in_force",
      "@container": "@set"
    },
    "version_date": {
      "@id": "eli:version_date",
      "@container": "@set"
    },
    "version": {
      "@id": "eli:version",
      "@container": "@set"
    },
    "hasAuthority": {
      "@id": "lkg:hasAuthority",
      "@container": "@set"
    }
  }
}
```

```
},
"jurisdiction": {
  "@id": "eli:jurisdiction",
  "@container": "@set"
},
"language": {
  "@id": "dct:language",
  "@container": "@set"
},
"type_document": {
  "@id": "eli:type_document",
  "@container": "@set"
},
"links": {
  "@id": "rdfs:seeAlso",
  "@type": "@id",
  "@container": "@set"
},
"uri": {
  "@id": "dct:uri",
  "@type": "@id"
},
"taClassRef": {
  "@id": "itsrdf:taClassRef",
  "@type": "@id"
},
"taIdentRef": {
  "@id": "itsrdf:taIdentRef",
  "@type": "@id"
},
"referenceContext": {
  "@id": "nif:referenceContext",
  "@type": "@id"
},
"taConfidence": {
  "@id": "itsrdf:taConfidence",
  "@type": "xsd:decimal"
},
"anchorOf": {
  "@id": "nif:anchorOf"
}
}
}
```