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

D2.8 Final Report of Data Management Activities

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), Patricia Martín-Chozas (UPM), Ilan Kernerman (KD), Pompeu Casanovas (UAB)
CONTRIBUTORS	Jorge González-Conejero (UAB), Elena Montiel-Ponsoda (UPM)
REVIEWERS	TILDE, SWC
VERSION | STATUS	V1 | Draft
NATURE	EC Open Research Data Pilot
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.4651389
DATE	31/04/2021 (M40)

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

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	Initial contribution	11/03/2021	Víctor Rodríguez-Doncel integrating the contributions of every author.
02	Table with updated info on the publication of legislation.	30/03/2021	Socorro Bernarndos and Pompeu Casanovas
03	Corrections	31/03/2021	Artem Revenko, Ilan Kernerman, Socorro Bernardos, Christian Sageder

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 reports on all the data management activities of the Lynx project. The contents of this document summarize and integrate the Data Management Plan (D2.1, D2.4), the deliverables on the acquired vocabularies and corpora (D2.5, D2.6) and the catalogue of relevant and regulatory datasets (D2.7).

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

Table 1 List of recommended metadata fields and their representation (as of March 2021) 24

Table 2 List of some NIF-related properties and their values..... 24

Table 3 Prefixes used in this document..... 25

Table 4. URI patterns for different resources..... 26

Table 5. Harvested data sources 17

Table 6. Table of relevant legal datasets. 6

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. It refers to the FAIR Data Principles developed by the FORCE 11 community, that recommend data should be shared according to these four concepts.
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 deliverable reports on all the data management activities of the Lynx project, which are a work done in the context of Work Package 2. The contents of this document comprise the Data Management Plan (D2.1, D2.4), the deliverables on the acquired vocabularies and corpora (D2.5, D2.6) and the catalogue of relevant and regulatory datasets (D2.7).

The Data Management Plan (DMP) is a living document, and as such it is presented in full extent here, even if large parts of it have already been reported. This document contains (in Section 2) the final version of the DMP, which had as previous versions the deliverables «D2.1 Initial Data Management Plan» (May 2018) and in «D2.4 Data Management Plan». (May 2019). This section is also 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. **Section 2** follows the template proposed by the EC, and Lynx adopts policies compliant with the official FAIR guidelines [1] (findable, accessible, interoperable and re-usable).

Data models were extensively presented in D2.4, but relevant progress have been made in two aspects: a change in the URI minting policy and an entirely renewed and extended validation schemes. **Section 3** describes this progress. **Section 4** presents the catalogue of datasets, describing the final tables and the incremental contribution over D2.5 and D2.6, including an updated review of the most relevant legal datasets for Lynx (originally as D2.7, updates reported here as section 4.4 and Annex B).

2 DATA MANAGEMENT PLAN

This Section constitutes the Data Management Plan at the end of the project, with final statistics. It follows the template proposed by the EC.

The EC promotes the access to and reuse of research data generated by Horizon 2020 projects through the Open Research Data Pilot. This project commits to the rules[1] 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 maintained an up-to-date Data Management Plan (this version is the final one).
- Lynx has deposited the research data in a **research data repository** –Zenodo. Lynx has a community in Zenodo, and CKAN provides a stable **catalogue of datasets** with their description (own and others). Final figures are shown in Figure 1.
- Lynx makes sure third parties can freely access, mine, exploit, reproduce and disseminate any piece of research data – where applicable and not in conflict with any IPR considerations.
- Lynx has made clear what tools are necessary to use the raw data to validate research results – standard formats have been used for data at every moment.

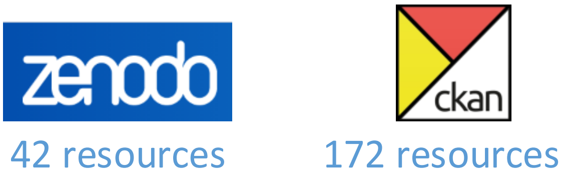


Figure 1. Resources published in Zenodo and in the data portal

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

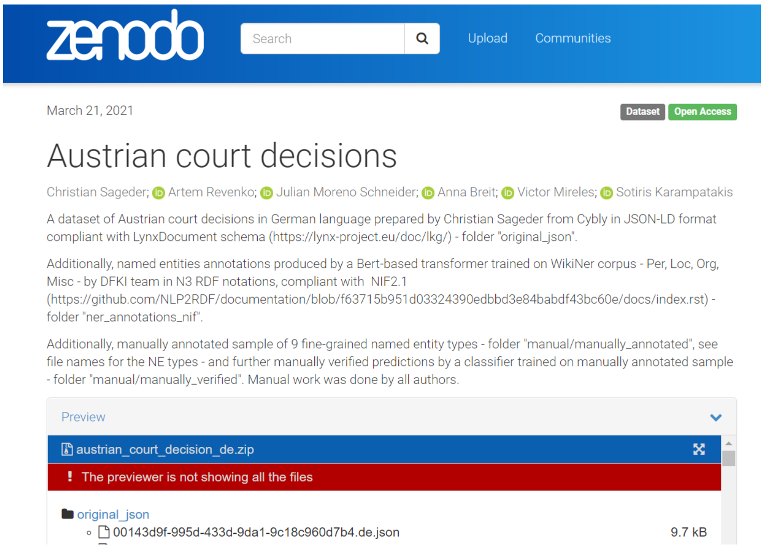
1. Data summary
a) What is the purpose of the data collection / generation and its relation to the objectives of the project?
The main objective of Lynx is "to create an ecosystem of smart cloud services to better manage compliance, based on a legal knowledge graph (LKG) which integrates and links heterogeneous compliance data sources including legislation, case law, standards and other aspects". In order to deliver these smart services, data has been collected and integrated into a Legal Knowledge Graph.
b) What types and formats of data will the project generate / collect?
The very nature of this project made the number of formats too high as to be foreseen in advance. However, the project has been be keen on gathering data in RDF format or producing RDF data itself. RDF has been the format of choice for the meta model, using standard vocabularies and ontologies as data models. Special care has been given to not forcing data users to learn RDF, but allowing them to use its JSON-LD version in a friendly manner. More details on the data models are given in Section 3.

[1] https://www.openaire.eu/what-is-the-open-research-data-pilot

c) Will you re-use any existing data and how?
The core part of the LKG was created by reusing existing datasets, either copying them into the consortium servers (only if strictly needed) or using them directly from the sources.
d) What is the origin of the data?
Lynx has been greedy in gathering and linking as much compliance-related data as possible from any possible source exceeding the initial geographical constraints to other jurisdictions, with the purpose of making experiments.
e) What is the expected size of the data?
The strong reliance of Lynx in external open data sources has minimized the amount of data that Lynx had to physically store.
f) To whom might the data be useful ('data utility')?
Data is useful for SMEs and EU citizens alike through different portals.

2.2 FAIR DATA

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. See an example of dataset in Figure 2. The description of the most relevant datasets for compliance have been published in a Lynx Data Portal, using the open source data portal CKAN software[2]. Metadata has been provided for every relevant dataset, and data has been selectively provided whenever it could be republished without license restrictions and relevance for the project was high.



March 21, 2021

Austrian court decisions

Christian Sageder; Artem Revenko; Julian Moreno Schneider; Anna Breit; Victor Mireles; Sotiris Karampatakis

A dataset of Austrian court decisions in German language prepared by Christian Sageder from Cybly in JSON-LD format compliant with LynxDocument schema (https://lynx-project.eu/doc/lkg/) - folder "original_json".

Additionally, named entities annotations produced by a Bert-based transformer trained on WikiNer corpus - Per, Loc, Org, Misc - by DFKI team in N3 RDF notations, compliant with NIF2.1 (https://github.com/NLP2RDF/documentation/blob/f63715b951d03324390edbbd3e84babdf43bc60e/docs/index.rst) - folder "ner_annotations_nif".

Additionally, manually annotated sample of 9 fine-grained named entity types - folder "manual/manually_annotated", see file names for the NE types - and further manually verified predictions by a classifier trained on manually annotated sample - folder "manual/manually_verified". Manual work was done by all authors.

austrian_court_decision_de.zip

! The previewer is not showing all the files

original_json
- 00143d9f-995d-433d-9da1-9c18c960d7b4.de.json 9.7 kB

Figure 2. Example of dataset in Zenodo.

[2] https://ckan.org/

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 D2.4. Data assets have been identified and harmonized.

b) What naming conventions do you follow?

A specific URI minting policy has been defined 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 has been "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 were created from existing private resources (e.g. dictionaries by KDictionaries), whose publication would irremediable damage their business model. These datasets have not been released as open data.

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 has assured Open Access to all peer-reviewed scientific publications relating to its results. Lynx has used 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 has acted as a central information node, with pointers to every relevant resource.

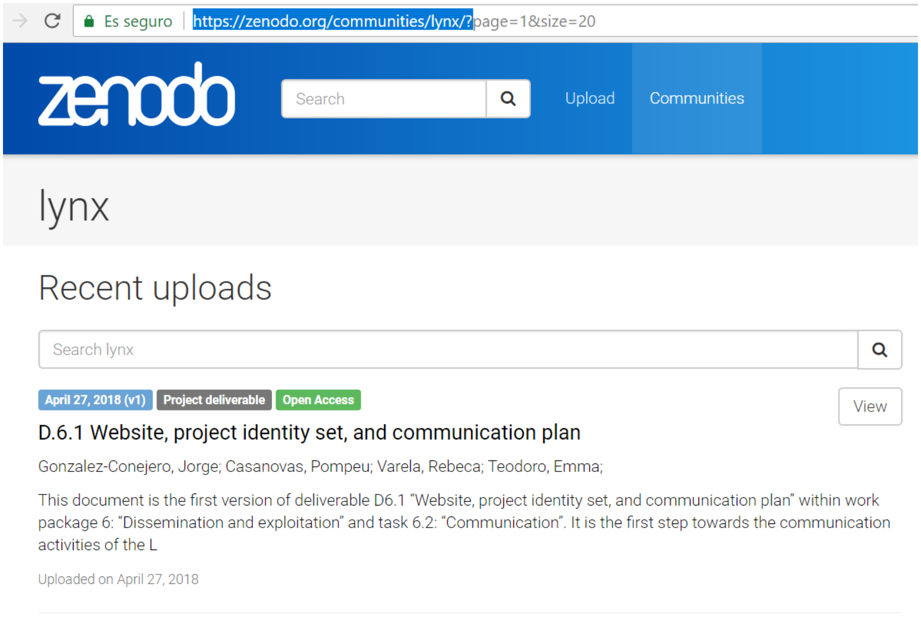
The following categories of outputs require Open Access to be provided free of charge by Lynx partners, to related datasets, in order to fulfil the H2020 requirements of making it possible for third parties to access, mine, exploit, reproduce and disseminate the results contained therein:

- *Public deliverables* are available both at Zenodo and the Lynx website at `http://lynx-project.eu/publications/deliverables`. See Figure 3 and Figure 4.

- Some *Conference and Workshop presentations* have been published at Slideshare under the account `https://www.slideshare.net/LynxProject`.

- Some *Conference and Workshop papers and articles for specialist magazines* have also been 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.



lynx

Recent uploads

April 27, 2018 (v1) Project deliverable Open Access

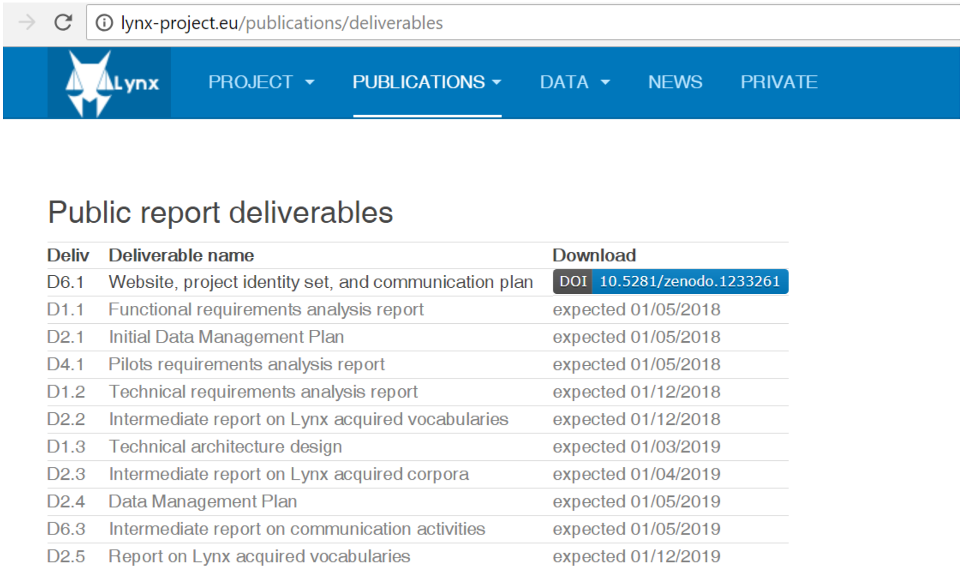
D.6.1 Website, project identity set, and communication plan

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 3. Lynx public deliverable at Zenodo.



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 4. 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.

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

Relevant datasets whose license is liberal are available as downloadable files. Some SPARQL endpoints have been set in place for those datasets in RDF form (see webpage for the pointers). 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 in the months after the project. Other software to be developed in Lynx has been 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 has used 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. Cloud computing is used in Austria.

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?

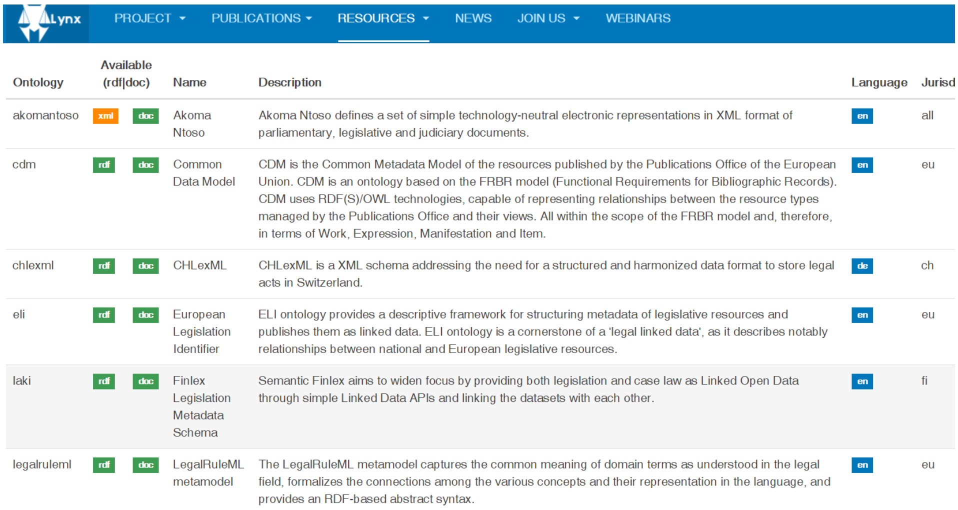
	There is no need for a Data Access Committee[3].
j) Are there well described conditions for access (i.e. a machine readable license)?	
	Description of data assets include a link to well-known licenses, for which machine readable versions exist. Either Creative Commons Attribution International 4.0 (CC-BY) or Creative Commons Attribution Share-Alike International 4.0 (CC-BY-SA) have been 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 have been defined throughout the entire project. `http://lynx-project.eu/data2/data-models` (see also Figure 5).
c) Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?	
	Standard vocabularies have been used inasmuch as possible, like the ECLI ontology, the Ontolex model and other vocabularies similarly spread. These choices grant inter-disciplinary collaboration. For example, Ontolex[4] is standard in the language resources and technologies communities, whereas the ELI ontology[5] (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?	

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

[4] http://lemon-model.net/

[5] http://publications.europa.eu/mdr/eli/

Figure 5 illustrates a visualization for **14** existing data models.



Ontology	Available (rdf|doc)	Name	Description	Language	Jurisd
akomantoso	xml doc	Akoma Ntoso	Akoma Ntoso defines a set of simple technology-neutral electronic representations in XML format of parliamentary, legislative and judiciary documents.	en	all
cdm	rdf doc	Common Data Model	CDM is the Common Metadata Model of the resources published by the Publications Office of the European Union. CDM is an ontology based on the FRBR model (Functional Requirements for Bibliographic Records). CDM uses RDF(S)/OWL technologies, capable of representing relationships between the resource types managed by the Publications Office and their views. All within the scope of the FRBR model and, therefore, in terms of Work, Expression, Manifestation and Item.	en	eu
chlexml	rdf doc	CHLexML	CHLexML is a XML schema addressing the need for a structured and harmonized data format to store legal acts in Switzerland.	de	ch
eli	rdf doc	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
laki	rdf doc	Finlex Legislation Metadata Schema	Semantic Finlex aims to widen focus by providing both legislation and case law as Linked Open Data through simple Linked Data APIs and linking the datasets with each other.	en	fi
legalruleml	rdf doc	LegalRuleML metamodel	The LegalRuleML metamodel captures the common meaning of domain terms as understood in the legal field, formalizes the connections among the various concepts and their representation in the language, and provides an RDF-based abstract syntax.	en	eu

Figure 5. 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 have been used.

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

Lynx aimed at building a LKG towards compliance. After the project end, 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 include a field informing on the last modification date.

e) Are data quality assurance processes described?

Only formal aspects of data quality have been assured.

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 have been (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 has been 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 (Cybly, Cuatrecasas, DNV GL). UPM has been responsible ('curator') for the Zenodo account and had to approve (*curate*) every upload.

d) Are the resources for long term preservation discussed?

Public deliverables and research data have been uploaded to Zenodo, which grants the long-term preservation. A specific community has been created in Zenodo[6].

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. Data has also lived in servers of Semantic Web Company (PoolParty) and Cybly (Vienna and Salzburg respectively). No personal data has been kept.

b) What provisions are in place for data security?

The relevant data that is open, has been 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*

[6] https://zenodo.org/communities/lynx/

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, no pilot has gathered personal data.

Ethical and societal aspects

The processing of personal data had a great ethical interest for –but no personal data was necessary in the end. The processing of personal data was a possibility in the framework of Pilot 1. However, GA Article 34 "Ethics and research integrity" is binding and would have been respected. Ethical and privacy related concerns were fully addressed in Section 3.2 of Deliverable 7.2 "*IPR and Data Protection management documents*".

The societal impact of this project has been 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 were requested to issue a statement on the ethical and societal impact of the Lynx project.

Finally, the Lynx websites have complied 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.

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.

The UAB partner has proceeded to review the degree of implementation of the risk management strategy, and 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.

Assessment

An Ethical and Societal impact assessment has been carried out. This assessment has been conducted following the methodology developed by the H2020 e-SIDES project.[7] In particular,

[7] https://e-sides.eu/e-sides-project

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. This assessment was made in D2.4
b) Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?
No personal data was gathered.

3 DATA MODELS

3.1 DATA MODEL FOR THE LYNX DOCUMENT

Data models were extensively described in D2.4 Data Management Plan, and modifications have been kept up to date online at `https://lynx-project.eu/doc/lkg/`. Because the information about the data models is so relevant, it is also reproduced here for convenience of the reader, with only minor updates, regarding the representation of jurisdiction and the document URI form.

The main element in the data models serves to describe the Lynx Document. Because the Lynx Document is in JSON-LD (RDF), the best possible form for the data model is an ontology. The ontology was also reported in detail in D2.4. The latest summary is offered here, with part of the text also being offered in that URL.

This ontology has been made with the purpose of supporting the representation of Lynx Documents, e.g., any document related to compliance worth to be in a Legal Knowledge Graph. This specification serves as a data model for documents in the Lynx project.

3.1.1 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 a Lynx Document.
- **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 temporal entities, identified terms in thesauri, etc. Other types of enrichments, not included in Lynx Annotations are summaries and translations.

Because most AI algorithms dealing with documents focus on text -manipulation of images, videos or tables are 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 6: 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.).

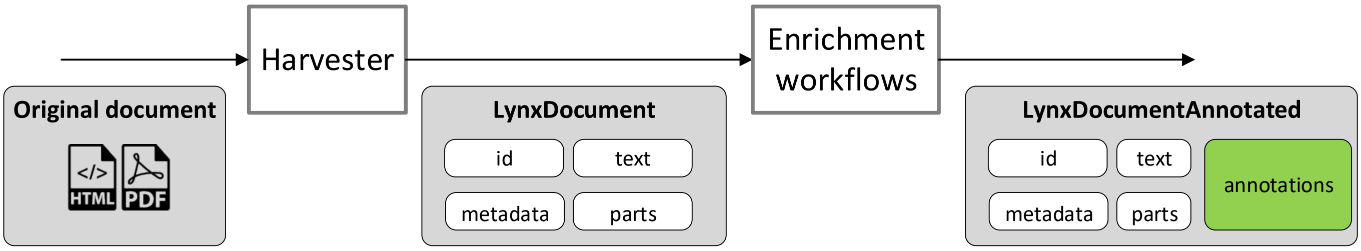


Figure 6 Original documents and Lynx Documents

The elements in a complete Lynx Document, with annotations, are depicted in Figure 7. 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').

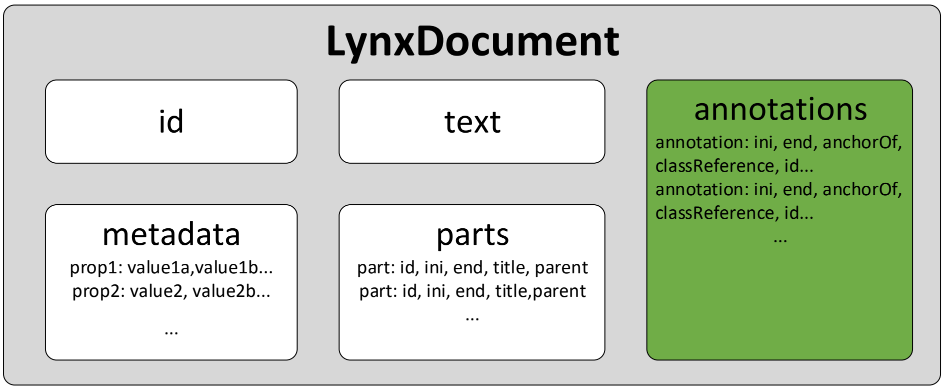

LynxDocument
id
text
annotations
annotation: ini, end, anchorOf, classReference, id...
annotation: ini, end, anchorOf, classReference, id...
...
metadata
prop1: value1a,value1b...
prop2: value2, value2b...
...
parts
part: id, ini, end, title, parent
part: id, ini, end, title,parent
...

Figure 7 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.

3.1.2 Lynx Documents with metadata

The simplest possible Lynx Document as a JSON file is shown in Figure 8.

```
{
  "@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 8 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.

To save space, the text is not repeated in the fragments. Alternative transformations of this JSON structure are possible and recommended for every specific implementation need (e.g. CYBLY 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 (please note that URIs are not fully conformant with what is described in Section 3.3, but valid for educational purposes)

```
@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 9 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 10 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 11 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 12 Example of Lynx Document with language tag (JSON-LD)

3.1.3 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 specification[8]. The example below shows an example of nested fragments, as Art. 2.1

```
{
  "@context": "http://lynx-project.eu/doc/jsonld/lynxdocument.json",
  "@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 13 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>
```

[8] http://persistence.uni-leipzig.org/nlp2rdf/

```
  nif:beginIndex 41 ;
  nif:endIndex 94 ;
  dc:title "Art.2" .
```

Figure 14 Simple example of Lynx Document (Turtle)

Two classes suffice for representing Lynx Documents without annotations as UML objects (See Figure 15).

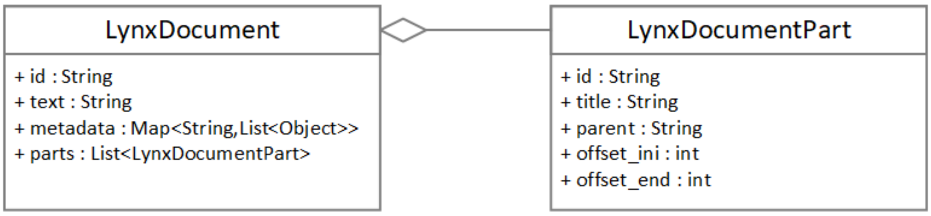


Figure 15 UML class diagram representation of Lynx document and Lynx document part.

3.1.4 Lynx document with annotations

Annotations are represented using NIF and live in class instances well differentiated from the Lynx Documents. A Lynx Annotation aggregates zero or more Lynx Annotation Units, as described in Figure 16, allowing thus that the same passage is annotated with different tags.

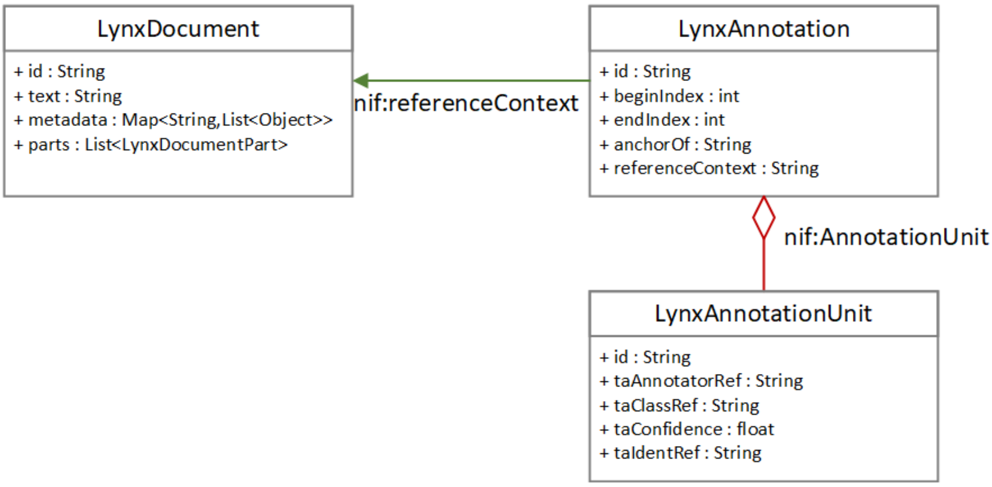


Figure 16. Lynx documents, annotations and annotation units.

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 17 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 18 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 3.

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

3.1.5 List of recommended metadata fields and their representation

Group	JSON Property	Usage	RDF property
general	language	Language of the document. Preferrably non-capitalized, no dialect variations.	dct:language
type_document	Sub-type of document (constitution, law, etc.). Defined for every legislation. For example, in Spain, the controlled vocabulary is here. Values in Appendix below. See also more strings in the controlled vocabulary here	eli:type_document	0-1
jurisdiction	ISO 3166-1 or ISO 3166-2 alpha 2 codes (with two letters), in capitalized form, such as AT (for Austria), or ES (Spain). Regions are therefore also possible: ES-MA (Madrid). Some LynxDocument processors may drop the region-specific letters. Entries in the ATU list may be accepted in the future for compatibility with ELI but not accepted in this version.	eli:jurisdiction	0-1
title	Title of the document. Language-tagged strings (one per language).	dct:title	0-1 per language
hasAuthority	Authority issuing the document. Single string with no language tag. See many elements in the NAL	lkg:hasAuthority	0-1
alternative	Alternative names of the document	dct:alternative	0-*

version	Consolidated, draft or bulletin. The value is string from a controlled vocabulary, for example 'con' (e.g. for Spain).	eli:version	0-1
subject	Subjects or keywords of the document. Array of language-tagged strings.	dtc:subject	0-*
summary	Summary of the text of the document. Language-tagged strings (one per language).	lkg:summary	0-1 per language
identifiers.	id_local	Local identifier (e.g. BOE-A-2019-1234).This metadata element is mandatory.	eli:id_local
accessGroup	Determines to which access groups a document belongs. E.g. "CocaCola". Array of non-language tagged strings.	lkg:accessGroup	0-*
dates	first_date_entry_in_force	Date when enters into force. In the form yyyy-mm-dd.	eli:first_date_entry_in_force
date_no_longer_in_force	Date when repealed / expired. In the form yyyy-mm-dd.	eli:date_no_longer_in_force	0-1
version_date	Date of publication of the document. In the form yyyy-mm-dd.	eli:version_date	0-1
Provenance	creator	Creators of the documents in Lynx (person or software)	dct:creator
created	Date when created in Lynx (internal). XSD datetime.	dct:created	0-1
rightsHolder	Who is the rightsholder (e.g. http://example.com/John)	dct:rightsHolder	0-1
source	Original URL if the document was extracted from the web	dct:source	0-1
hasEli	Official identifier (ELI, ECLI or equivalent)	lkg:hasEli	0-1
hasPDF	Link to the PDF version. Single URI.	lkg:hasPDF	0-1
hasDbpedia	Link to the equivalent dbpedia version. Single URI.	lkg:hasDbpedia	0-1
hasWikipedia	Link to the equivalent wikipedia version. Single URI.	lkg:hasWikipedia	0-1
sameAs	Equivalent document. Array of URIs	owl:sameAs	0-*
seeAlso	Related documents. Array of URIs	rdfs:seeAlso	0-*

Table 1 List of recommended metadata fields and their representation (as of March 2021)

The next table lists the recommended metadata fields and their representation.

Element	Meaning	Values / example
itsrdf:taClassRef	Class of the annotated context	dbo:Person, dbo:Location, dbo:Organization, dbo:TemporalExpression
itsrdf:taIdentRef	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 2 List of some NIF-related properties and their values

Table 3 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 3 Prefixes used in this document

3.2 VALIDATION

Validation is the operation that determines whether a document represents a valid Lynx Document or not –according to certain rules. Validation has been currently implemented within the DocumentManager, and it is executed upon ingestion of documents by the creation workflow[9].

Most validation operations are done automatically by checking against RDF Shapes[10]. The following RDF Shapes have been defined:

1) Validation rules related to NIF features[11].

These are some of the rules related to NIF, described in plain text:

- Instances of `nif:String` should have a valid URI.
- A `nif:String` that is not a `nif:Context` should have a `nif:referenceContext` property with one nif:Context as object
- A `nif:OffsetBasedString` must have exactly one `nif:beginIndex` and one `nif:endIndex`
- Instances of `nif:OffsetBasedString` should have a URI pattern conforming to the pattern '`#offset_{beginIndex}_{endIndex*}`'
- `nif:beginIndex` and `nif:endIndex` should match the number of word positions in `nif:isString`

An example of SHACL shape is presented here for convenience:

```
:OffsetBasedStringShape
    a sh:NodeShape ;
    sh:targetClass nif:OffsetBasedString ;
    sh:message "R008. Instances of nif:OffsetBasedString should have a URI pattern conforming to the
pattern '#offset_{beginIndex}_{endIndex*}'" ;
    sh:pattern "#offset\\_\\d+\\_\\d*\\b" ;
    sh:flags "i" .
```

Validation rules directly related to the Lynx Document[12]:

These are some of the rules related to the Lynx Document:

- Language of the target must exist and be unique

[9] The method `/validate/validate` in DCM determines whether a Lynx Document is valid against a series of rules or not.

[10] Shapes Constraint Language (SHACL). W3C Recommendation 20 July 2017. https://www.w3.org/TR/shacl/

[11] https://gitlab.com/superlynx/common/lynx_mvn/-/blob/master/src/main/resources/shapes/nif-shapes.ttl

[12] https://gitlab.com/superlynx/common/lynx_mvn/-/blob/master/src/main/resources/shapes/lkg-shapes.ttl

- A valid Lynx Document should only use the `nif:isString` property to define the text of the context. nif:`contextStringRef` is not supported.
- There must be one and exactly one metadata element in every Lynx Document
- Every document may have 0 or 1 parent, but no more.
- A Lynx Document must be also a NIF document (the type being at least `lkg:LynxDocument` and `nif:Context`)
- Language metadata element must be unique and a valid ISO code
- Title must have a unique language tag.
- Summary must have a unique language tag.
- There can be zero or one version date
- There can be zero or one creation date.
- There must exist exactly one id_local element.
- There can be at most one jurisdiction
- Jurisdiction must be a valid ISO code (2 letters), or regional ISO code.
- There can be at most one link to dbpedia
- There can be at most one PDF connected to the document
- There can be at most one ELI connected to the document
- Lynx Documents of type “legislation” must have exactly one “Jurisdiction” element.

3.3 URI MINTING STRATEGY

A very extensive report was presented in D2.4, “Section 5: URI minting policy”, to explain the design decisions taken around URI minting. Few changes have been made since then, except the very important identifier of Lynx Documents, whose final form is the following.

```
https://apis.lynx-project.eu/document-platforms/{implementationId}/collections/{collectionId}/documents/{docId}
```

Where `implementationId` is one of these strings: “`ldp`” or “`upm-elastic`”, `collectionId` is the name of the collection and `docId` is the document identifier. Lynx Documents were not the only document with URIs. For other URIs, see the next table.

Other URI patterns are shown in Table 4.

Type of resource	**URI pattern**
Ontology	`http:// lkg.lynx-project.eu/def/{onto_id}`
Example	*`http://lkg.lynx-project.eu/def/core`*
Ontology element	`http://lkg.lynx-project.eu/def/{onto_id}/{element}`
Example	*`http://lkg.lynx-project.eu/def/core/Document`*
KOS (thesauri, terminologies)	`http://lkg.lynx-project.eu/kos/{kos_id}/{id}`
Example	*`http://lkg.lynx-project.eu/kos/contracts_terms/24232`*

Table 4. URI patterns for different resources

4 DATA

Data collected during this project falls into four categories, that follow below. The list also includes the deliverable in which they have been reported.

- Domain Independent Vocabularies (D2.5)
- Domain Dependent Vocabularies (D2.5)
- Translation Corpora (D2.6)
- Legislation and documents to populate the LKG (2.4, D2.7).

4.1 DOMAIN INDEPENDENT VOCABULARIES

4.1.1 Introduction to Domain Independent Vocabularies

Domain-independent vocabularies in Lynx provide a common catalogue of word meanings which allow to traverse semantically annotated documents from different domains. They also serve as support for NLP tasks such as word sense disambiguation (WSD), question answering (QA), and cross-lingual search, and help to retrieve synonyms and translations, among other lexical data.

In Lynx, these vocabularies come in the form of multilingual lexicographic resources by the consortium partner K Dictionaries (KD), and they comprise, specifically, lexical data for Dutch, English, German and Spanish.

How is this data integrated into the LKG? For these lexical resources to be easily integrated into the Legal Knowledge Graph (LKG), we rely on the semantic representation of the data as Linked Data (LD). To this end, the data needs to be first converted to the Resource Description Framework (RDF) format and semantically annotated through the use of ontologies. This serves to ensure both syntactic as well as semantic interoperability of the resulting datasets, accessible at the end of this process via a REST API.

4.1.2 The transformation process: from XML to RDF

Figure 19 illustrates the entire conversion process.

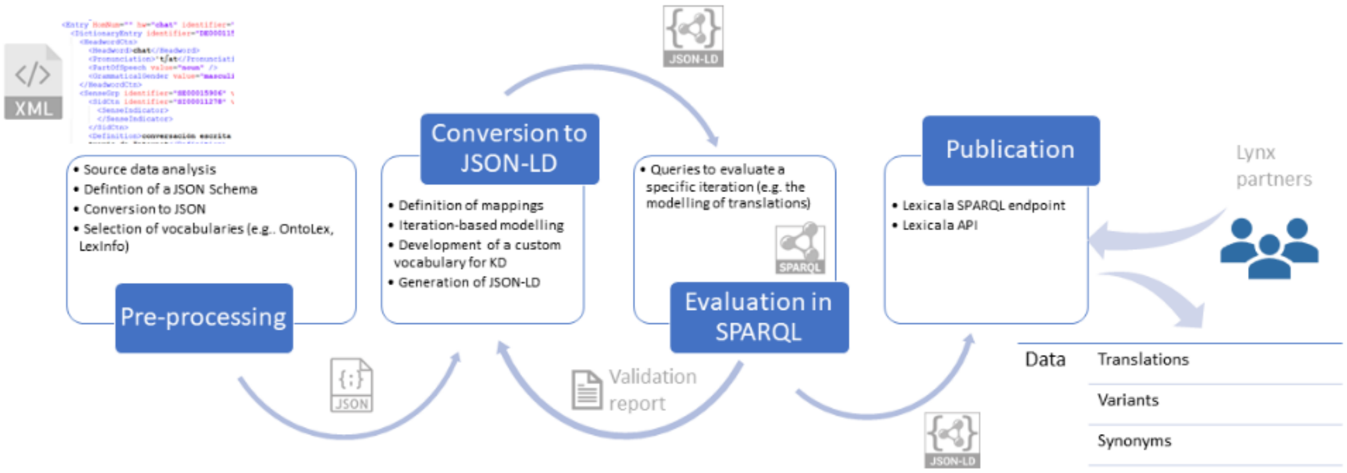


Figure 19. Conversion process of DIV from XML to RDF

4.1.3 Source data analysis

The source data is based on KD’s multilingual Global series, implementing the model on each individual entry and relying on pre-existing relations in KD resources to further link information across languages. The data is originally in XML, which is then converted into JSON format. Each XML element of an entry is mapped to a corresponding JSON type (including strings, arrays and objects) and each entry is represented by a single JSON object containing the entry components within.

A standard entry in the data consists of the headword, part of speech, grammatical information (such as sub-categorization, number or gender) and other semantic and syntactic information. Entries can be either monosemous (i.e. consisting of a single sense) or polysemous (i.e. divided into multiple senses) and each sense contains a definition and/or some indicator of the specific meaning. Such information can include register, subject field or sentiment, as well as synonyms and antonyms, example phrases, multiword expressions and translation equivalents in other languages.

These components are represented in a structured JSON format and governed by a schema to arrange the information in place and check that the relations between the different nested objects are correct, making sure that no ill-fitted relationship occurs in the data, thus enabling a stable and reliable basis for RDF implementation and conversion.

4.1.4 Model definition and conversion to JSON-LD

The semantic representation of the data as LD is based on the OntoLex-lemon model and its *lexicog* and *vartrans* modules. The OntoLex-lemon model is the de facto standard to represent lexical resources as LD. It consists of an RDF model that, with a set of core classes (such as ontolex:LexicalEntry or ontolex:LexicalSense) and several modules, allows for the representation of a wide range of lexical descriptions including morphosyntactic properties, translations, and pragmatic information. This model is commonly used with other ontologies such as LexInfo, which serves as a linguistic category registry. For example, OntoLex-lemon enables describing in RDF a dictionary entry as a lexical entry, its module lexicog adds further description to encode how this entry is arranged in the specific resource it comes from (e.g. if it is a headword or not), the *vartrans* module makes it possible to represent translations between this lexical entry and others, and LexInfo offers elements to specify its morphosyntactic features (part of speech, gender, number, etc.). A custom ontology for KD was also designed for those cases in which LexInfo did not provide a linguistic category equivalent to the original description in KD's DTD.

Once the model was defined, the process of converting KD data into LD was carried out following an incremental approach, starting with the very basics of a single entry (headword, part of speech, senses, definitions) and proceeding with more complicated elements, including synonyms, compounds, examples of usage, translations, etc. Each iteration focused on the modelling of a specific type of linguistic information and implementing the modelling decisions in the pipeline to generate JSON-LD.

The details on the URI naming strategy adopted for the conversion, as well as the linking advantages it offers across dictionaries, is presented in detail in Lonke and Bosque-Gil (2020) [36].

4.1.5 Evaluation

After generating the corresponding JSON-LD entries for each iteration, the resulting data was validated iteratively through SPARQL queries. A series of queries were defined, specifically targeted at evaluating how the representation of the linguistic information covered in that iteration was implemented in the pipeline and whether the model was followed accordingly. This incremental process has not only assured constant validation and error handling, but also allowed for an adaptation period, during which the process of writing queries for validation has shed more light on the model and methods of improvement. Taking into account the data requirements of the Lynx services and our initial experiments with the RDF data, we have been able to improve the queries and iteratively change the model so that the results optimally represent the users' needs.

The details on the first modelling efforts with KD lexicographic data in OntoLex are elaborated on Bosque-Gil et al. (2016) [37], while the application of the lexicog module and the iterative evaluation are presented in detail in Bosque-Gil et al (2019) [38].

4.1.6 Publication and exploitation

The resulting dataset provides a wide array of lexicographic components represented as RDF with OntoLex-Lemon, LexInfo and KD custom ontology, including the headword, part of speech, inflections, grammatical information, examples of usage, multiword expressions, synonyms and antonyms, and translation equivalents accessible via the SPARQL endpoint to the Lynx partners. The entries are linked across KD's multilingual Global series through translation relations (across source and target languages), and monolingual sense relations (antonomy, synonymy). This leads to the emergence of a graph of lexical data, illustrated here in the visual rendering of the English entry bow (noun) and its homographs:

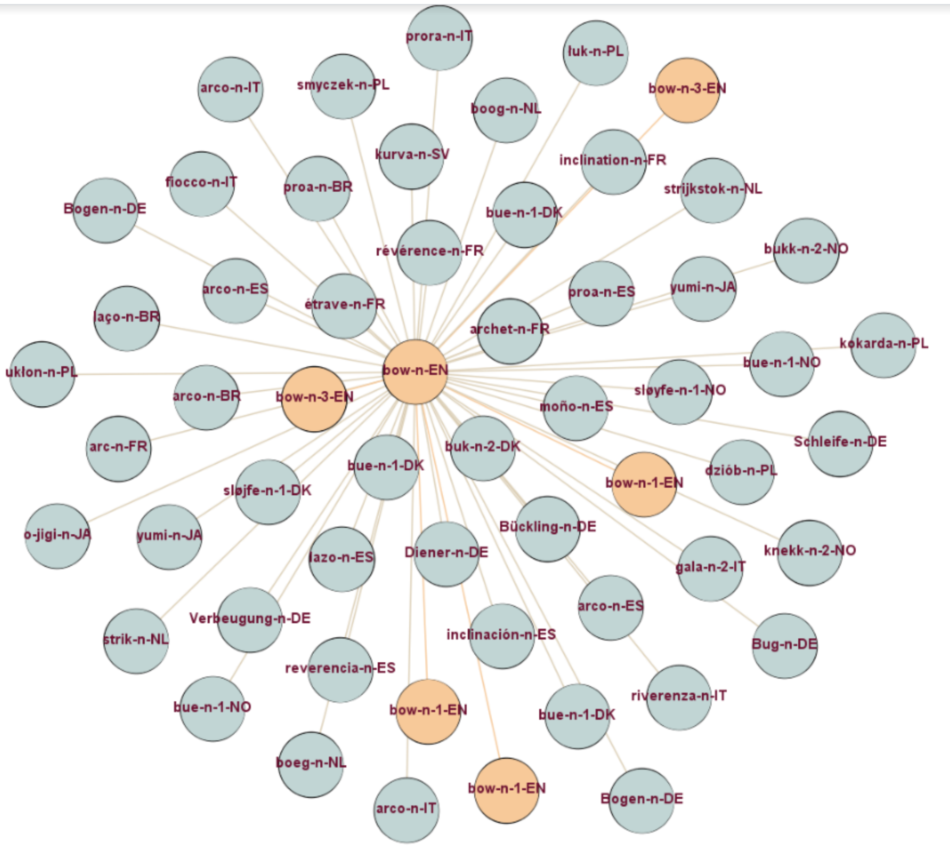


Figure 20. Graph of lexical data

Some example queries over the graph of lexical data are provided and documented in Lonke and Bosque-Gil (2020) [36].

KD's multilingual data is being exploited by several Lynx services (`https://lynx-project.eu/doc/api/).` It is being used by the Search services to perform query expansion through a series of SPARQL queries which, given a lemma, return the lemmas of the synonyms (non-inflected forms) and the inflections (if available). It is also being used in the creation of ad-hoc terminologies to provide synonyms and translations of relevant terms, and also in the disambiguation process of data retrieved from the Linguistic Linked Open Data cloud.

A table with existing domain independent vocabularies is presented as Annex A in this document.

4.2 DOMAIN DEPENDENT VOCABULARIES

Domain dependent vocabularies have been necessary in Lynx. Some of them were existing resources, some of them had to be created ad-hoc. This section provides only a very brief summary of what was already presented in Deliverable D2.5

4.2.1 Existing DDV

High quality resources exist ready to be used. The main resource is the EuroTermBank portal (`www.eurotermbank.com`), a network of stakeholders for publishing and hosting EU-related open terminology data. Another platform is Tilde Terminology platform which provides services and tools (e.g. term annotation, automatic extraction) as well as a platform for managing private term collections. Lynx has benefitted from EuroTermBank by getting public and open resources, as well as from Tilde Terminology by using its services (automatic term extraction, translation lookup).

EuroTermBank was an initial EU project whose purpose was to gather information related to EU-linked terminology. As the EuroTermBank project focus was on harmonization and consolidation of terminology work in EU member states, transferring experience from other EU terminology networks and accumulating competencies and efforts of the accessed countries. EuroTermBank currently provides terminology in **34** languages and wide domain coverage, and it is perfectly suitable to addressing the harvesting process of DDV in Lynx. Conversely, Lynx will also contribute to this project, since all existing DDV now have the option of being converted into RDF format, as explained in the next section.

4.2.2 Transformation of existing resources

To acquire new domain vocabularies, the consortium looked at different data sources to harvest terminology resources that deal with legal vocabulary. Various paths were followed to identify and explore relevant resources for Lynx, including:

- General web search
- Lookup of resources described in papers from specialized literature
- Search in data portals specialized in language resources

The full workflow of processing different data formats is presented in Figure 21.

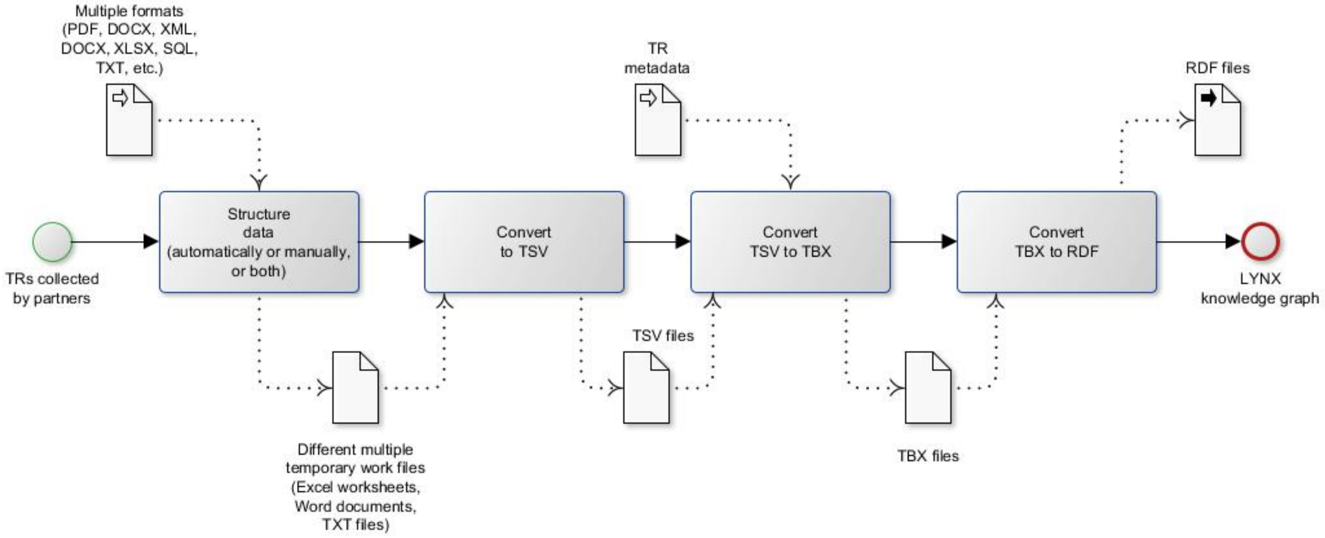

Multiple formats (PDF, DOCX, XML, DOCX, XLSX, SQL, TXT, etc.)
TR metadata
RDF files
TRs collected by partners
Structure data (automatically or manually, or both)
Convert to TSV
Convert TSV to TBX
Convert TBX to RDF
LYNX knowledge graph
TSV files
Different multiple temporary work files (Excel worksheets, Word documents, TXT files)
TBX files

Figure 21. Workflow of processing different data formats

4.2.3 Semi-automatic creation of rich SKOS Concept Schemes

After the analysis of the existing domain dependent vocabularies, we realised that the most specific assets were not openly available, and those published in open formats are, mostly, too general for Lynx pilots. Therefore, we worked on the generation of domain dependent vocabularies per pilot.

From the corpora collected per each of the pilots, already reported in previous deliverables, we extracted the main terms representing each legal sub domain: labour law, contracts and industrial standards. We also extracted the contexts in which those terms appear, that are good usage examples and also act as sense indicators to be used in a later disambiguation stage.

Term extraction algorithms, especially statistical algorithms, tend to extract constructions of words that are not correct from a terminological perspective. For this reason, we performed a semi-automatic refinement step to remove inconsistencies.

The above-mentioned disambiguation stage is needed since, in order to generate multilingual resources, we need to retrieve translations and additional information from more general resources such as Wikidata[13] and IATE[14], that may contain several senses for the same word. Thanks to the disambiguation algorithm provided by Semantic Web Company, we enriched our terms with translations, synonyms, and definitions in the four languages of the project: Dutch, English, German and Spanish. We also retrieved hierarchical relations provided by thesauri such as EuroVoc[15] and the Unesco Thesaurus[16].

Finally, the generated domain dependent resources were represented following the SKOS vocabulary. As a result, we therefore generated three concept schemes, one per pilot, that are linked with other vocabularies in similar RDF formats, taking part of the Linguistic Linked Open Data cloud[17].

4.3 TRANSLATION CORPORA

Translation corpora was presented in D2.6 and no relevant work has been made since its description. This deliverable presented the work done in Task 2.4 “Indexing of corpora” and Task 2.5 “Translation corpora creation” in WP2, which followed up D2.3 ‘Intermediate report on Lynx acquired corpora’.

This deliverable D2.6 also reported on the translation corpora creation, with specific focus on the identification of language resources in public online sources and data repositories in the legal domain, including parallel data, monolingual data, glossaries, etc. It provided the workflow of web-crawling, automated extraction, and other data collection methods which were in the end applied for the Lynx business cases.

Various resources (documents and websites) useful for the NMT systems were provided by Lynx partners and later processed to create our geothermal energy related corpora. They are listed in D2.6 and no changes have been made. Finally, a complete list of corpora used for the machine translation training was used, being all the details in Tables 5-8 in D2.6.

4.4 LEGISLATION CORPORA

Existing legislation corpora were described in «D2.7 Catalogue of relevant legal and regulatory datasets».

[13] https://www.wikidata.org/

[14] https://iate.europa.eu/

[15] http://publications.europa.eu/resource/dataset/eurovoc

[16] http://vocabularies.unesco.org/

[17] http://linguistic-lod.org/llod-cloud

4.4.1 Description

The catalogue in Annex B is the result of an exhaustive search for available legal and regulatory resources on the web from different jurisdictions across the European Union, first presented in D2.7 and updated here, mostly with the result of the survey of Filz et al. (2021) [45].

This effort is the result of the analysis made on the legal information sources provided by the 27 countries that conform the European Union (as of March 2021), plus the European Union itself and the United Kingdom. The analysis considers several criteria of interest for the Lynx project, namely, type of documents available (law and/or case law), level of implementation of the ELI (European Resource Identifier) standard, formats in which the legislation is published, language of the data sources, or accessibility options (access interface, possibility of a bulk download). Details on the methodology of this analysis have already been described, in D2.7 –this document only provides an update. The table columns of the table in Annex B are described again here, as they have been changed since its first version in D2.7.

In the following we explain the rationale behind the features of the datasets analysed in the table:

Area. The state-of-the-art done in the context of the Lynx project and reported in this deliverable focuses on the legislation at EU and Member-State levels. Some of the portals mentioned in the table include links to sub-national level resources, especially for countries with federal or autonomous structure, but that is not reflected in the table.

Web site. URIs to the resources publishing legal information.

Content. The focus of this work has been on law and case law resources, the most relevant sources of legal information for the Lynx project. We also account for the level of implementation of ELI or ECLI, with Id referring to the situation where documents are given an ELI (or ECLI) identifier and Id/M indicating that metadata for the documents are also provided.

License Type. Most of the datasets do not explicitly adhere to a particular type of license. In those cases, the table reflects if the access is free (F) or restricted (R).

Format. The documents can be downloaded in different formats (PDF, XML, HTML, etc.). In many cases, there is an HTML version, but it cannot be directly downloaded. This is reflected by the expression (online).

Language. The two-letter code in ISO 636-1 is used to capture the available languages for each dataset

Access Interface. The table reflects if an API exists and/or a bulk download of documents is allowed and/or an SPARQL endpoint is available. When this is the case and the access it to be found on sites different from the one in the first column, the links are included.

Each of the resources included in the table are also available through the Lynx Data Portal[18].

[18] http://data.lynx-project.eu/dataset

4.4.2 Harvested datasets

Corpora on contracts were already described in D2.6, Section 2.1. Standards, recommended practices recommendations and resolutions in Section 2.2, labor law corpora in Section 2.3, legislation in Section 2.4. The list of harvested documents, duely updated after its first version in 2.6 is shown in Table 5.

The total number of harvesated documents follows, broken down by Pilot. Numbers are not precise as harvesters are periodically run.

- Pilot on geothermal energy. Number of documents: ~300. Languages: nl, en.
- Pilot on contracts. Number of documents (as of March 2021): 721.986 Languages: de, nl, en, es. More specifically:
  - Austrian Laws: 8,845 in 89,532 versions (Lynx Documents)
  - Austrian Articles in the system: 584,425, in force: 244,133 (parts of a Lynx Document)
  - German Laws: 7,203 all in force (Lynx Documents)
  - German Articles: 129,592 all in force
  - EurLex Legislation: 85,102 in force 51,629 (Lynx Documents)
  - EurLex Articles: 544,295 in force 530,831
  - Dutch Law: 1 (PoC to check that data can be imported if necessary)
  - Dutch Articles: 1,747
  - Austrian Decisions: 523,431 (Lynx Documents)
  - EurLex Decisions: 16,717 (Lynx Documents)
  - A number of contracts that largely varies per case.
- Pilot on labor law. Number of documents: ~150,000 documents. Languages: es, de, en
  - Spanish legislation: ~120,000 documents as Lynx Documents
  - Spanish labor-law specific legislation: 10,000 articles as Lynx Documents
  - Spanish collective agreements: ~5,000 documents as Lynx Documents
  - Austrian legislation on labour law: ~1,000 documents as Lynx Documents
  - Various pieces of legislation of other jurisdictions for testing purposes (eu, ie, uk, nl, el, mx, au, at) in several languages (es, es-mx, en-au, en-ie, de, nl, el).

Jurisdiction	URI	Source Data Format	ELI implemented	Type of docs	Estimate of Volume	Domain	Sub-Domain	Harvester	Harvester is public?	Content availability online
ES	boe.es	PDF,HTML,XML	yes	Legislation	approx. 120,000	law	State and Autonomous Community legislation	UPM	yes	yes
ES	sede.asturias.es	XML, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES, CAT	caib.es	HTML, RDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES, CAT	https://dogc.gencat.cat	RDF, TURTLE, XML, HTML, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES	http://boa.aragon.es	XML, HTML, JSON, PDF	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
ES	bocyl.jcyl.es	XML	-	Collective Agreement		Labour Law	Collective Agreement	UPM	to be made public	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	10672 [19] [20]	law	federal	LawThek	no [21]	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1289	law	Burgenland	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	318	law	Kärnten	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1054	law	Niederösterreich	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	876	law	Oberösterreich	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1144	law	Salzburg	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	1098	law	Steiermark	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	573	law	Tirol	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	779	law	Voralberg	LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	level 1	Legislation	467	law	Wien	LawThek	no	yes

[19] Number of laws, not documents. Within RIS 1 article is 1 document
[20] Laws which are currently in force
[21] Only the converter from lawthek.eu to a Lynx-Document

AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	22619	VfgH		LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	116077	VwgH		LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	134473	Justiz		LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	132869	BVwG		LawThek	no	yes
AT	ris.bka.gv.at	PDF,HTML,XML	ECLI	jurisprudence	21554	LVwG		LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	Level 1	Legislation		Legal acts	Directives	LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	Level 1	Legislation		Legal acts	Regulations	LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	ECLI	jurisprudence		Judgment	Court of Justice	LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	ECLI	jurisprudence		Judgment	General Court	LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	-	Legislation		Consolidated texts	Directives	LawThek	no	yes
EU	eur-lex.europa.eu	PDF,HTML,XML	-	Legislation		Consolidated texts	Regulations	LawThek	no	yes
DE	https://www.gesetze-im-internet.de	PDF,HTML,XML, EPUB	No	Legislation	6497		Federal	LawThek	no	yes
DE	http://www.rechtsprechung-im-internet.de	PDF,HTML,XML		jurisprudence	51625			LawThek	no	yes

Table 5. Harvested data sources

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, dehttps://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] Rodriguez-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.

[36] Lonke, D., and Bosque-Gil, J. (2020) Applying the OntoLex-lemon lexicography module to K Dictionaries' multilingual data. Kernerman Dictionary News, 28. https://kln.lexicala.com/kln28/lonke-bosque-gil-ontolex-lemon-lexicog/

[37] Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., and Aguado-de-Cea, G. (2016). Modelling multilingual lexicographic resources for the Web of Data: The K Dictionaries case. In GLOBALEX 2016 Lexicographic Resources for Human Language Technology Workshop Programme (pp. 65-72). https://ufal.mff.cuni.cz/~kolarova/2016/docs/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf#page=71

[38] Bosque-Gil, J., Lonke, D., Gracia, J., & Kernerman, I. (2019). Validating the OntoLex-lemon lexicography module with K Dictionaries' multilingual data. In Electronic lexicography in the 21st century. Proceedings of the eLex 2019 conference (pp. 726-746). https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_41.pdf

[39] Bautista Salinero, J. (2020). Extracción y Normalización de Convenios Colectivos. Trabajo fin de Grado, Universidad Politécnica de Madrid. To appear in http://oa.upm.es

[40] Eurostat. (2009) Nace Rev. 2 Metadata. URL: https://ec.europa.eu/eurostat/ramon/nomenclatures/index.cfm?TargetUrl=LST_NOM_DTL&StrNom=NACE_REV2&StrLanguageCode=EN&IntPcKey=&StrLayoutCode=HIERARCHIC

[41] Pinnis, M. (2013). Context Independent Term Mapper for European Languages. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 562-570).

[42] Pinnis, M. (2018). Tilde's Parallel Corpus Filtering Methods for WMT 2018. In Proceedings of the Third Conference on Machine Translation (pp. 952–958).

[43] Shuyo, N. (2010). Language detection library for java. Retrieved Jul, 7, 2016.

[44] Zariņa, I., Ņikiforovs, P., Skadiņš, R. (2015). Word alignment based parallel corpora evaluation and cleaning using machine learning techniques. In Proceedings of the 18th Annual Conference of the European Association for Machine Translation.

[45] Filtz, E., Kirrane, S. & Polleres, A. (2021) The linked legal data landscape: linking legal data across different countries. Artif Intell Law. https://doi.org/10.1007/s10506-021-09282-8

Appendix 1. Relevant language resources documented in Lynx Data Portal

name	title	domain	publisher	jurisdiction	language	notes
abc-of-oil	ABC of oil	industry	Norwegian Petroleum Directorate	Norway	en, no	ABC of oil contains terminology and abbreviations used in oil industry.
agrovoc	Agrovoc				29 languages	Controlled general vocabulary
babelnet	BabelNet	General			270 languages	BabelNet is regarded as a multilingual encyclopaedic dictionary but also as a semantic network and a knowledge base that combines general data with lexical information that comes from WordNet.
biocides	Biocides	agriculture	Health and Safety Executive	European Union	en	This glossary contains terminology related to biocides and their regulation.
bpr	BPR - Bibliography of the Italian Parliament	Legal	Italian Parliament		it	The BPR - Bibliografia del Parlamento italiano e degli studi elettorali (Bibliography of the Italian Parliament and Electoral Studies) is a database of bibliographic references of books and articles in periodical journals addressing the history of the Italian Parliament and the history of elections.
calathe	Cadastral Vocabulary (CaLaThe)	Environmental Law	University of Denmark and University of Turkey		en	Monolingual thesaurus for the domain of cadastre and land administration that provides a controlled vocabulary. It contains 142 terms related to previous vocabularies, such as AGROVOC, GEMET and STW thesaurus for economics. Terms are arranged in graphical, tabular and alphabetical order.
cdisc-glossary	CDISC Glossary	Clinical Data	CDISC		en	CDISC develops data standards to streamline clinical research and enable connections to healthcare. This glossary contains definitions of terms and abbreviations. It is published as PDF and XLS formats, and it has also been converted into RDF.
conneticut-legal-glossary	Conneticut Legal Glossary	Legal	State of Conneticut Judicial Branch		en, es	Bilingual glossary from English into Spanish organised in alphabetical order that can be downloaded in PDF. It is published by the Conneticut Judicial Branch. Terms contained in this glossary cover general law area, including labour law and economic terminology.
copyrighttermbank	Copyright Termbank	Copyright	OEG	European Union	en, es, fr, pt	A multilingual termbank that contains copyright-related terms from WIPO definitions, IATE and Creative Commons licenses. This termbank is also connected to external resources such as DBpedia and Lexvo. Terms have been hierarchically organised, and they are useful to annotate licenses.
court-thesaurus	Court Thesaurus	legal	Wolters Kluwer		de	A monolingual thesaurus in German containing names of German and international courts.

dataprotectionglossary	Data Protection Lynx Glossary	Legal	OEG		en	Monolingual termlist from the Data Protection domain in RDF linked with BabelNet, DBpedia and EuroVoc.
dbpedia	DBpedia				>100 languages	Extensive general knowledge base providing information about approx. 4 million entities in more than one hundred languages.
ecb	European Central Bank Corpus	financial, legal	OPUS, the open parallel corpus	European Union	cs, da, de, el, en, es, et, fi, fr, hu, it, lt, lv, mt, nl, pl, pt, sk, sl	ECB Corpus is a multilingual corpus that contains financial vocabulary. It has been extracted from the website and documentation of the European Central Bank, and it is aligned among 19 European languages.
edp	European Data Portal	Several topics including: government, justice and legal fields	European Commission	European Union	bg, cs, da, de, el, en, es, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	The European Data Portal harvests the metadata of Public Sector Information available on public data portals across European countries. Not only linguistic information is presented here, but any kind of data related to diverse domains: environment, government, law, etc. Information regarding the provision of data and the benefits of re-using data is also included. Also, datasets can be filtered by country, language, format, domain, etc.
eige	EIGE Glossary	Gender Equality	EIGE		en	Glossary of gender-related terms in English published by the European Institute of Gender Equality. It contains over 400 terms in English that have been extracted from 92 resources. Each entry includes a link to its related source.
eucases	EUCases			eu, uk, bg	en, bg, de	This is the result of the EUCases Research Project which developed a European case law Linking Platform, transforming multilingual legal open data into linked open data. The EUCases Linking Platform links EU law with legislative acts and court decisions of six EU member states: Austria, Bulgaria, France, Germany, Italy and United Kingdom.
eugdpr-glossary	EUGDPR Glossary	Legal	EUGDPR	European Union	en	A Glossary of Terms and Definitions as used in relation to the GDPR in the EU.
eugo	EUGO Glossary	Business	EUGO	Spain	es	The glossary of EUGO is addressed to companies and entrepreneurs that need to comply with administrative or professional requirements to perform a remunerated economic activity in Spain.
eurlex	EUR-Lex	Law	Publications Office	European Union	bg, cs, da, de, el, en, es, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	EUR-Lex gives access to EU Law, the jurisprudence of the EU Court of Justice, other EU public documents and the electronic edition of the Official Journal of the EU, in 24 languages.
eurovoc	EuroVoc Thesaurus	Politics, international	European Union	n.a.	bg, cs, da, de, el, en, es, et, fi, fr,	EuroVoc is a multilingual, multidisciplinary thesaurus that covers the activities of the EU. It is available in 23

		relations, european communities, law			hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	languages of the European Union. It is intended to be used as a disambiguation tool by contextualising each term, offering univocal meanings. Also, the tool provides related terms and preferred and non preferred designations to guide the user.
eurovoc-agrovoc-semantic-alignment	EuroVoc-AgroVoc Semantic Alignment	Agricultural, General	Publications Office		en	Semantic Alignment of AgroVoc Thesaurus with EuroVoc Thesaurus.
eurovoc-eige	EuroVoc-EIGE Semantic Alignment	Gender Equality, Legal	Publications Office	European Union	en	Semantic Alignment of Gender Equality glossary with EuroVoc Thesaurus.
eurovoc-gemet-semantic-alignment	EuroVoc-GEMET Semantic Alignment	Environmental	Publications Office		Several languages including: en, de, es, it	Semantic Alignment of GEMET Thesaurus with EuroVoc Thesaurus.
eurovoc-inspire-alignment	EuroVoc-Inspire Alignment	General	Publications Office	European Union	en	Semantic Alignment of Inspire glossary with EuroVoc Thesaurus.
eurovoc-lcsh-semantic-alignment	EuroVoc-LCSH Semantic Alignment	General	Publications Office		en	Semantic Alignment of the Library of Congress Subject Headings with EuroVoc Thesaurus.
eurovoc-stw	EuroVoc-STW Semantic Alignment	Economics	Publications Office		en	Semantic Alignment of STW Thesaurus for economics with EuroVoc Thesaurus.
eurovoc-umthes-semantic-alignment	EuroVoc-Umthes Semantic Alignment	Environmental	Publications Office		de, en	Semantic Alignment of the Umthes (German Environmental Thesaurus) with EuroVoc Thesaurus.
eurovoc-unbis	EuroVoc-UNBIS Semantic Alignment	General	Publications Office		ar, cn, en, es, fr, ru	Semantic Alignment of UNBIS Thesaurus with EuroVoc Thesaurus.
eurovoc-unesco-semantic-alignment	EuroVoc-UNESCO Semantic Alignment	General	Publications Office	European Union	en, es, fr, ru	Semantic Alignment of the UNESCO Thesaurus with EuroVoc Thesaurus.
evroterm	Evroterm	legislation	General Secretariat of the Government of the Republic of Slovenia	European Union	en, sl	The Evroterm terminology collection contains terms from EU legal regulations and other documents related to the EU.
finnish-legislation	Finnish legislation			Finland		Finnish legislation as linked data. This dataset convers the following legal subdomains: legislation, case law (supreme court), case law (supreme administrative) and court. http://data.finlex.fi
gemet	GEMET (General Multilingual Environmental Thesaurus)	General	EIONET (European Environment Information and Observation Network)		Several languages including: en, de, es, it	GEMET is a compilation of the following resources. - “Umwelt Thesaurus” that has more than 2.000 descriptors out of 8.500 in German and English. - “Thesaurus Italiano per l'Ambiente (TIA)” with more than 4.000 descriptors in Italian, English, Dutch and German. - “Multilingual Environment Thesaurus (MET)” with more than 2.300 descriptors in Dutch, Danish, English, French, German, Italian, Norwegian and Spanish. - “EnVoc Thesaurus”, of UNEP Infoterra, 1997 edition, with about 2.000 descriptors in English, French

						and Spanish, with possibility of access to Arabic, Chinese and Russian. - "Thesaurus de Medio Ambiente" with more than 2.600 descriptors in Spanish, English, French, German. - "Lexique environnement - Planète" with more than 5.000 descriptors in French and English. - Descriptors of relevant documents of the EEA, namely "Europe's Environment, The Dobris Assessment", the "DPSIR Data Flow Scheme", as well as terminology of ETCs and Eionet, in English. - Descriptors of the "Thesaurus Eurovoc" in French, English, Dutch, German, Italian, and Spanish, with possibility of access to Danish, Greek, and Portuguese. The result are 6562 terms arranged in 3 super-groups, 30 groups plus 5 accessory, instrumental groups, hierarchically organised.
gllir	Glossary of labour law and industrial relations	employment	International Labour Organization	European Union	en	
gllt	German labour law thesaurus	Labor Law	Wolters Kluwer Deutschland GmbH	Germany	de	Labour law thesaurus covers all main areas of labor law, including the roles of employee and employer; legal aspects around labour contracts and dismissal; also co-determination and industrial action.
gowers-review-of-intellectual-property	Gowers Review of Intellectual Property	laws	Crown		en	Glossary about Intellectual Property can be found on pages 121 to 127.
iate	IATE	Law, information technology, agriculture, etc.	European Union	n.a.	bg, cs, da, de, el, en, es, et, fi, fr, ga, hr, hu, it, la, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	IATE is a terminological database developed by the European Union that contains around 8,5 million terms in the 24 official languages of the EU. It can also be downloaded both in RDF and TBX format. IATE uses EuroVoc Thesaurus to classify its entries by domain.
ilo-taxonomy	ILO Taxonomy	employment	International Labour Organization		en, fr, es	ILO Taxonomy contains terms related to the world of work in English, French, and Spanish.
imf	International Monetary Fund Terminology		IMF		de, en, es	This terminology list contains over 150,000 terms useful to translators working with IMF material. It provides English terms with their equivalents in a number of languages. This list includes words, phrases, and institutional titles commonly encountered in IMF documents in areas such as money and banking, public finance, balance of payments, and economic growth.
industrialstandardsglossary	Industrial Standards Lynx Glossary	Legal	OEG		en	Monolingual termlist from the Industrial Standards domain in RDF linked with BabelNet and DBpedia.
informea	InforMEA Glossary (UNESCO)	Environmental Law	UNESCO		en	The glossary contains terms, definitions and related information on Multilateral Environmental Agreements. Such terms are classified in 6 different domains: Water, Chemicals and Wastes, Biodiversity, Air and Climate,

						Land and Environmental Law and Governance. Terms are also organised hierarchically, grouped in small sets of terms that are are all dependent on a top term. This glossary also stablishes broader and narrower relations between terms. Moreover, it provides sources where the terms have been applied to check their right usage context.
inspire	INSPIRE Glossary (EU)	Spatial information	European Union		en, es	Glossary developed by the INSPIRE Knowledge Base of the European Union. The INSPIRE Glossary contains 195 general terms and definitions that specify the common terminology used in the INSPIRE Directive and in the INSPIRE Implementing Rules documents. This glossary contains definitions both in English and Spanish. It also provides information about the previous versions of each entry and it marks if terms are valid or not.
jrcacquis	JRC-Acquis	Legal	Ralf Steinberger - European Commission - Joint Research Centre (JRC)		bg, cs, da, de, el, en, es, et, fi, fr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	JRC-Acquis is a collection of legislative texts from the European Union generated between the years 1958 and 2006. They are available in xml format and in 22 languages of the EU.
jurivoc	Jurivoc	Law	Bundesgericht (Switzerland)	Switzerland	de, it, fr (and some terms in EU languages)	JURIVOC, is the juridical thesaurus of the Federal Supreme Court and the former Federal Insurance of Switzerland. It contains information in German, French and Italian. Terms are arranged as per the following relations: - LT: lead term SN: scope note - UF: used for - UFA: used for... and... - NT: narrower term - SA: see also - BT: broader term
labour-law-corpus	Labour Law Corpus	Labour Law	Lynx Project	National and European	en, es, de, it	Corpus of legal documents from different countries: Austria, Germany, Ireland, Italy, Spain and UK. And also from the European Union.
labourlawglossary	Labour Law Lynx Glossary	Legal	OEG		es	Monolingual glossary containing terms in Spanish from the labour law domain. Terms have been automatically extracted from legal corpora and manually reviewed by language and law professionals. The glossary was first generated in TBX and eventually converted into RDF to be linked with BabelNet, Eurovoc and DBpedia.
legislative-glossary	Legislative Glossary (Glossário Legislativo)	legislation	Câmara Municipal de Bento Gonçalves		pt	Monolingual glossary about legislation in Portuguese
libraryofcongress	Library of Congress	General: books, agreements, documents	Library of Congress	Global	en	The Library of Congress Linked Data Service enables both humans and machines to programmatically access authority data of the Library of Congress.
llcorpuses	Spanish Labour Law Corpus	Legal		Spanish	es	Labour law corpus composed by 20 agreements in Spanish provided by Cuatrecasas Lynx partners.

myndigheternas-foreskrifter	Government Regulations (Myndigheternas föreskrifter)	legislation	Sveriges regering	Sweden	sv	
nacerev2	NACE, Rev. 2	statistics	Eurostat	European Union	en	Statistical Classification of Economic Activities in the European Community, Rev. 2
nomothesia	NOMOTHESIA			Greece	el	Greek legislation modelled as per Metalex OWL ontology. It can be accessed through a SPARQL endpoint or downloaded as RDF file. The content of this resource can also be requested through a websearch application.
osh-thesaurus	OSH Thesaurus	employment	International Labour Organization	United Nations	en, fr, es	OSH Thesaurus contains over 15,000 multilingual terms and synonyms on occupational safety and health.
psi-glossary	PSI Glossary	inofrmation	European Commision		en	Contains terminology related to Public Sector Information
quality-glossary	Quality Glossary (Терминологичен речник на качеството)	statistics	Национален статистически институт.	European Union	bg, en	Bilingual glossary about quality in English and Bulgarian with definitions.
saij	SAIJ Legal Thesaurus (Argentine Juridical Information System)	Law	Government of Argentine		es	SAIJ Thesaurus organises legal knowledge through a list, modelled with SKOS, of controlled terms which represent concepts. It is used to ease users' access information related to the Argentinian legal system that can be found in a file or in a documentation centre. Terms are also hierarchically organised with broader and narrower relations.
stw	STW Thesaurus for Economics	Economics	ZBW		en	The thesaurus provides vocabulary on any economic subject: almost 6,000 standardized subject headings and about 20,000 additional entry terms to support individual keywords. Terms used in law, sociology, or politics can also be found. This thesaurus is provided by the Leibniz Information Centre and it has been published in RDFa format to boost the reuse of such resources in the Semantic Web. It is one of the first resources in the Linked Open Data Cloud since it is mapped with a great number of related resources.
tcnen	TERMCAT Collective Negotiation Glossary EN	Legal	TERMCAT	dbpedia	en	Monolingual termlist from the Collective Negotiation (Labour Law) domain in RDF linked with DBpedia and EuroVoc. (English)
tcnes	TERMCAT Collective Negotiation Glossary ES	Legal	TERMCAT		es	Monolingual termlist from the Collective Negotiation (Labour Law) domain in RDF linked with DBpedia and EuroVoc. (Spanish)

temcoord	Temcoord Glossaries	Several topics	DG TRAD	European Union	bg, cs, da, de, el, en, es, et, fi, fr, ga, hr, hu, it, lt, lv, mt, nl, pl, pt, ro, sk, sl, sv	Repository that contains around 300 hundred glossaries developed by European institutions and bodies in the 23 languages of the European Union. Most of them are available as PDF files, while others have been published in HTML format. The content of this repository is heterogeneous in terms of domain and information exposed in each glossary: some glossaries provide with definitions of the terms while others establish equivalences in different languages, etc.
termcat	Termcat Terminological Database	General	Termcat		ca, en, es, de, fr, it	TERMCAT’s mission is to ensure the development and integration of Catalan terminology into both specialist sectors and society in general.
thesoz	TheSoz - Thesaurus for Social Sciences	Social Sciences	Leibniz Institut für SozialWissenschaften		en, de, fr, ru	Thesaurus about social sciences organised according SKOS vocabulary containing terms in English, German, French and Russian.
uk-legislation	UK legislation					UK legislation as linked data
umwelt-thesaurus	Umwelt-Thesaurus	Environmental	Federal Environment Agency, Germany		de, en	German Thesaurus containing terms on environmental protection. It includes 13500 descriptors and 40000 linked German-language synonyms with 35000 English translations and around 11000 definitions.
unbis	UNBIS Thesaurus	General	UN Library		ar, cn, en, es, fr, ru	The UNBIS Thesaurus is a multilingual database of the controlled vocabulary used to describe UN documents, and terms are extracted from these documents.
unesco	UNESCO Thesaurus	Education, Science, Culture, Politics, Countries, Information	UNESCO		en, es, fr, ru	The UNESCO Thesaurus is a controlled and structured list of terms used in subject analysis and retrieval of documents and publications in several fields. The thesaurus is divided into seven greater domains and other smaller thesauri, where terms are hierarchically organised as per the following properties: SN - Scope Note MT - Microthesaurus UF - Used For BT - Broader Term NT - Narrower Term RT - Related Term The UNESCO Thesaurus can be downloaded as RDF file, accessed through a SPARQL endpoint and in the website: search by domain, by alphabetical order and by hierarchical order.

ANNEX A. EXISTING DOMAIN INDEPENDENT VOCABULARIES FROM KD

name	title	domain	publisher	description	language	format
api	Lexicala API	DIV	K Dictionaries	data transfer facility	40+ languags	JSON, JSON-LD
mdls-de	Global German	DIV	K Dictionaries	German monolingual dictionary core, with translation to English, Dutch (+ other languages)	de, en, nl + 7	XML, RDF
mdls-en	Global English	DIV	K Dictionaries	English monolingual dictionary core, with translation to Spanish (+ other languages)	en, es + 6	XML, RDF
mdls-es	Global Spanish	DIV	K Dictionaries	Spanish monolingual dictionary core, with translation to English, Dutch (+ other languages)	es, en, nl + 6	XML, RDF
mdls-it	Global Italian	DIV	K Dictionaries	Italian monolingual dictionary core, with translation to English (+ other languages)	it, en + 2	XML
mdls-nl	Global Dutch	DIV	K Dictionaries	Dutch monolingual dictionary core, with translation to German, English, Spanish	nl, de, en, es	XML
pw	Password Multilingual Dictionary	DIV	K Dictionaries	English multilingual dictionary (translation 40+ languages)	en, de, es, it, nl +40	XML
kmt-de	MultiGloss German	DIV	K Dictionaries	German-English semi-automated multilingual glossary	de, en, es, it, nl + 40	XML
kmt-es	MultiGloss Spanish	DIV	K Dictionaries	Spanish-English semi-automated multilingual glossary	es, en, de, it, nl + 40	XML
kmt-it	MultiGloss Italian	DIV	K Dictionaries	Italian-English semi-automated multilingual glossary	it, en, de, es, nl + 40	XML
kmt-nl	MultiGloss Dutch	DIV	K Dictionaries	Dutch-English semi-automated multilingual glossary	nl, en, de, es, nl + 40	XML
rhwcd	Random House Webster's College Dictionary	DIV	K Dictionaries	English monlingual glossary	en	XML
wfl	Word Form Lists	DIV	K Dictionaries	morphol;ogical lists of inflected forms and keywords	de, en, es, it, nl	XLSX

ANNEX B. TABLE OF RELEVANT LEGAL DATASETS

Table 6, presented in alphabetical order after the entry for the European Union, is updated as of March 2021.

Area	Web site	Content		License type	Format				Access interface	Language (ISO 639-1)
		Law	Case law		PDF	XML	HTML	Other		
European Union	https://eur-lex.europa.eu/	x (ELI)	x (ECLI)	F	x		x			EU languages
	https://data.europa.eu/euodp/es/data/dataset/official-journals-of-the-european-union-in-english	x (ELI)		F		Formex format	x	xsl	API	EN (EU languages)
	https://data.europa.eu/euodp/es/data/dataset/eu-case-law		x (ECLI)	F	x		x		API	EU languages
Austria	https://www.ris.bka.gv.at/	x (ELI Id)	x (ECLI Id)	F	x	x	x	rtf	API (https://data.bka.gv.at/ris/api/v2.5/)	DE EN (some laws)
Belgium	https://belgiumlex.be/	X (ELI Id)	X (ECLI Id)	F	x		x (online)			NL, FR, DE
Bulgaria	https://dv.parliament.bg/DVWeb/index.faces	x		F	x		x (online)			BG
	https://legalacts.justice.bg/		X (ECLI Id)	F	x		x (online)			BG

Croatia	http://nn.hr/	x (ELI Id)		F			x (online)			HR
	https://sudskapraksa.csp.vsrh.hr/home		x (ECLI Id)	F	x			txt		HR
Cyprus	http://www.cylaw.org/	x	x	F	x (law)		x (case law)			EL
Czech Republic	https://aplikace.mvcr.cz/sbirka-zakonu/	x		F	x					CS
	http://www.nsoud.cz		X (ECLI Id)	F	x		x	doc		CS
Denmark	https://www.retsinformation.dk/	x (ELI Id/M)		F	x		x (online)			DA
	https://domstol.dk/				x		X (online)			DA
Estonia	https://www.riigiteataja.ee/	x	x (ECLI Id)	F	x	x (law)	x (law, online)	txt (law)		ET EN (some laws)
Finland	http://www.finlex.fi/	X (ELI Id/M) (SFL)	x (ECLI (SFCL))	F	x (law)		x (online)			FI, sámi SV, EN (some texts)

France	https://www.legifrance.gouv.fr/	x (ELI Id/M)	x (ECLI Id)	F	x		x (online)	rtf	API (information at https://www.legifrance.gouv.fr/contenu/pied-de-page/open-data-et-api)	FR EN, DE, IT, ES (some texts)
Germany	https://www.bgbl.de/	x		F	x		x (online)			DE EN (some laws)
	https://www.bundesverfassungsgericht.de/		X (ECLI Id/M)	F	x		x			DE
Greece	http://www.et.gr/index.php/	x		F	x					EL
	http://www.adjustice.gr/		X (ECLI Id)	F			x (online)	txt		EL
	http://legislation.di.uoa.gr/	x (ELI)		F			x (online)	rdf, json	Endpoint	EL
Hungary	http://www.njt.hu/	x		F			x (online)			HU, EN (some laws)
	https://birosag.hu/birosagi-hatarozatok-gyujtemenye		x (ECLI Id)	F				rtf		HU
Ireland	http://www.irishstatutebook.ie/	x (ELI Id/M)		F	x		x (online)			EN, GA
	https://courts.ie/		x	F	x		x (online)			EN

Italy	https://www.normattiva.it/	X (ELI Id/M)		F		x	x			IT
	https://www.gazzettaufficiale.it/	X (ELI Id/M)		F	x					IT
	http://pst.giustizia.it/PST/		X (ECLI Id)	R						IT
	http://www.italgiure.giustizia.it/		X (ECLI Id)	R						IT
Latvia	https://likumi.lv/	x (OGPM)		F			x (online)			LV EN, RU (some laws)
	https://manas.tiesas.lv/eTiesas/		x (ECLI Id)	F	x					LV
Lithuania	https://www.e-tar.lt/portal/lt/index	x		F	x		x (online)	docx, odt		LT
	http://liteko.teismai.lt/viesasprendimupaieska/		x	F			x (online)			LT
Luxembourg	http://legilux.public.lu/	x (ELI Id/M) (JOLUX)		F	x	x	x	rdf	Endpoint, Bulk (http://legilux.public.lu/editorial/casemates)	FR
	https://justice.public.lu/fr/jurisprudence/jurisprudence-judoc.html		x (ECLI)	F	x					FR

Malta	https://justice.gov.mt/ (https://legislation.mt/)	x		F	x					EN
	https://justice.gov.mt/ (https://ecourts.gov.mt/onlineservices/Judgements)		x (ECLI Id)	F	x					EN or MT
Netherlands	https://www.officielebekendmakingen.nl/	x		F	x	x	x	odt, rdf	API (information at https://www.koopoverheid.nl/documenten/instructies/2021/02/09/handleiding-voor-het-uitvragen-van-de-collectie-officiele-publicaties)	NL, FR EN (some laws)
	https://uitspraken.rechtspraak.nl/		x (ECLI Id/M)	F		x (online)				NL
Poland	http://isip.sejm.gov.pl/	x		F	x (some)					PL
	http://orzeczenia.nsa.gov.pl/cbo/query		x	F				rtf		PL
Portugal	https://dre.pt/	x (ELI)		F	x		x (online)			PT EN (summaries)
	http://www.dgsi.pt/		X (ECLI Id) (OGPM)	F			x			PT
Romania	http://legislatie.just.ro/	x		F			x (online)			RO

	http://www.rolii.ro/		x (ECLI Id)	F	x		x (online)			RO
Slovakia	https://obcan.justice.sk/ (https://www.slov-lex.sk/domov)	x		F	x		x (online)			SK
	https://obcan.justice.sk/infosud/-/infosud/zoznam/sud		x (ECLI Id)	F	x	x	x (online)			SK
Slovenia	http://www.pisrs.si/Pis.web/	X (OGP M)	x (ECLI Id)	F	x		x (online)	docx		SL EN (some laws)
	http://www.sodnapraksa.si/		x	F			x (online)			SL
Spain	https://www.boe.es/	x (ELI Id/M)		F	x	x	x (online)	epub	API (information at https://www.boe.es/ datosabiertos/)	ES (CA, GL, EU)
	http://www.poderjudicial.es/search/indexAN.jsp		x (ECLI Id)	F	x					ES
Sweden	https://lagrummet.se/ (http://rkrattsbaser.gov.se// https://rattsinfosok.domstol.se/)	x	x	F	x (case law)		x (online)			SV
United Kingdom (not EU)	http://www.legislation.gov.uk/	x		F	x		x (online)			EN, CY
	https://www.bailii.org/		x	F	x		x	rtf		EN
	https://www.supremecourt.uk/decided-cases/		x (SC)	F	x					EN

Table 6. Table of relevant legal datasets.