



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

D2.3 Intermediate report on Lynx acquired corpora

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	Ēriks Ajausks (TILDE), Víctor Mireles-Chaves (SWC), Christian Sageder (openlaws), Andis Lagzdīņš (TILDE), Elena Montiel-Ponsoda (UPM)
CONTRIBUTORS	Roberts Rozis (TILDE), Rinalds Vīksna (TILDE), Matīss Rikters (TILDE), María Navas-Loro (UPM), Patricia Martín-Chozas (UPM), Socorro Bernardos Galindo (UPM)
REVIEWERS	UPM, SWC
VERSION STATUS	V1.0 Final
NATURE	Report
DISSEMINATION LEVEL	Public
DOCUMENT DOI	10.5281/zenodo.2655048
DATE	30/04/2019



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

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Draft of TOC	31/03/2019	Ēriks Ajausks (TILDE), Roberts Rozis (TILDE), Andis Lagzdiņš (TILDE), Artūrs Vasiļevskis
0.2	Draft of the document body	16/04/2019	Ēriks Ajausks (TILDE), Roberts Rozis (TILDE), Andis Lagzdiņš (TILDE), Rinalds Vīksna (TILDE), Matīss Rikters (TILDE)
0.3	Proofreading	17/04/2019	
0.4	Draft of the document body	19/04/2019	Ēriks Ajausks (TILDE), Elena Montiel-Ponsoda (UPM), Víctor Mireles-Chaves (SWC), Christian Sageder (OLS), Andis Lagzdiņš (TILDE), Patricia Martín-Chozas (UPM), María Navas-Loro (UPM)
0.5	Draft of the document body	26/04/2019	Ēriks Ajausks (TILDE), Elena Montiel-Ponsoda (UPM), Víctor Mireles-Chaves (SWC), Christian Sageder (OLS), Andis Lagzdiņš (TILDE)
1.0	Proofreading and final version of the document	30/04/2019	Elena Montiel-Ponsoda (UPM), Socorro Bernardos Galindo (UPM), María Navas Loro (UPM)

DISCLAIMER

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

ACRONYMS LIST

CKAN – Web based management system for the storage and distribution of data

CSV – Comma separated values (file format)

DOC, DOCX – Filenames extension for document files

HTML – Hypertext Markup Language

IPR – Intellectual Property Rights

JPG/JPEG – Joint Photographic (Experts) Group (compression for digital images)

JSON – Java Script Object Notation

MT – Machine Translation

NMT – Neural Machine Translation

OCR – Optical Character Recognition

PDF – Portable Document Format

RDF – Resource Description Framework

SKOS – Simple Knowledge Organization System

SMT – Statistical machine translator

TBX – TermBase eXchange

TMX – Translation Memory eXchange (file format)

TSV – Tab-separated values

TXT – A filename extension for text files

XLS, XLSX – Filenames extension for Microsoft Excel sheet format

XML - Extended Markup Language

XSL – Extensible Stylesheet Language

XSLT –Extensible Stylesheet Language Transformations (for transforming XML)

TABLE OF CONTENTS

TABLE OF CONTENTS	3
LIST OF FIGURES.....	5
EXECUTIVE SUMMARY.....	6
INTRODUCTON.....	7
1 INDEXED LEGISLATION CORPORA	8
1.1 CONTRACT CORPORA.....	8
1.1.1 Extracted terminology	9
1.2 STANDARDS, RECOMMENDED PRACTICES, REGULATIONS, AND RESOLUTIONS	10
1.2.1 Extracted Terminology.....	11
1.3 LABOR LAW CORPORA	11
1.3.1 Extracted Terminology.....	20
1.4 GENERAL LEGAL CORPORA	21
1.5 CORPORA INDEXING METHOD.....	22
1.6 NIF ANNOTATION.....	25
1.6.1. Methodology/Approach/Conversion Process	26
2 CREATED TRANSLATION CORPORA.....	27
2.1 CORPORA CREATION WORKFLOW	27
2.1.1 General parallel corpora creation workflow	27
2.1.2 Corpora creation from web crawled data	27
2.2 LIST OF TRANSLATION CORPORA.....	29
2.3 LIST OF CORPORA FOR MACHINE TRANSLATION TRAINING	32
3 CONCLUSIONS AND FUTURE WORK	35
ANNEX 1. MAIN EU LABOR LAW LEGISLATION CORPORA.....	36
REFERENCES.....	37

Table 1. Number of documents by document type in business case 2.....	10
Table 2. Summary of size of each component of the corpus for business case 3.....	20
Table 3. Number of extracted candidate concepts	20
Table 4. Information about the structure of a document.....	24
Table 5. Annotation properties defined until now in the Lynx project.....	26
Table 6. Resources used to build corpora	31
Table 7. Segment count in LYNX_Geothermal	31
Table 8. Corpora used for EN-NL NMT systems	33
Table 9. Corpora used for EN-ES NMT systems.....	33
Table 10. Corpora used for EN-DE NMT systems	34

LIST OF FIGURES

Figure 1. Example of the extracted candidate concept “storage tank”	11
Figure 2. Compilation of Labor law related documents in the Spanish Official Gazette website.....	12
Figure 3. Spanish Center for Judicial Documentation website.....	13
Figure 4. Website of the Spanish Register for Collective Agreements.....	14
Figure 5. Spreadsheet with sectoral collective agreements in Spain.....	14
Figure 6. Example of Austrian collective agreements per sector	16
Figure 7. Wikiversity website with links to Italian Labor law	16
Figure 8. Official website for accessing UK legislation	17
Figure 9. Official website for accessing UK case law	18
Figure 10. Irish legislation accessible from the Irish Statute Book.....	19
Figure 11. Irish case law accessible from Workplace Relations website.....	19
Figure 12. Example of the extracted candidate concept “Holiday Scheme”	21
Figure 13. Example of NIF annotated document in the Lynx project.....	25
Figure 14. General language resource processing workflow used for Lynx corpora creation.....	27
Figure 15. Web-crawled data processing workflow	28
Figure 16. Parallel corpus processing workflow	28

EXECUTIVE SUMMARY

This deliverable summarizes the intermediate work on acquired corpora (as part of WP2) within the context of the Lynx project. The aim of this task is to provide a description of the corpora collection methods, and the resulting collected corpora by Lynx partners around the different use cases. There are three business cases for which corpora are being collected. The first case is related to Compliance Assurance Services for Contracts, the second is related to Compliance Assurance Services in Oil & Gas and Energy, and the third Business Case is about Compliance Assurance Services in Labor Law. This document serves as reference material for the corpora collected to cover the needs of the three business cases, and for the first steps in the method followed to index that corpora. Furthermore, the document describes the corpora preparation workflow to be used in the training of Neural MT engines for specific languages and domains. Finally, this document reports on the term extraction process performed so far on the compiled corpora and briefly outlines its further use in the Lynx MT systems.

INTRODUCTON

This report summarizes the work done in Task 2.4 “Indexing of corpora” and Task 2.5 “Translation corpora creation” in WP2. T2.4 includes interim work done by openlaws, UPM, SWC, and DFKI and is scoped to provide a compilation of data to cover the needs of the different Lynx business cases. In the respective sections within section 1, partners provide information about the corpora gathered. The report summarizes the work done on data compilation and preparation. The detailed description and corpora interlinkage with services will be included in deliverable D3.4. The second part of the report focuses on 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 provides the workflow of web-crawling, automated extraction, and other data collection methods applied for the Lynx business cases. A significant part of the report consists of the description of language resource processing (i.e. aligning, and re-formatting), which is a crucial part of translation corpora creation for developing MT engines.

1 INDEXED LEGISLATION CORPORA

In this section we describe the process of acquisition and indexing of textual resources to support the three business cases defined in the project, that is, Business Case 1 “Compliance Assurance Services for Contracts”, Business Case 2 “Compliance Assurance Services in Oil & Gas and Energy”, and Business Case 3 “Compliance Assurance Services in Labor Law”.

By *indexing* of textual resources we mean the curation, storage and organization of the acquired textual resources. Importantly, this requires that documents be described using the provenance and bibliographic descriptors agreed upon, and that they be made available to the services developed in WP3. These descriptors are listed in the data-models of the Lynx project¹, and include subsets of cataloguing standards such as DCAT, and legal-specific standards such as ELI. We note that this definition of indexing implies (and goes beyond) the one used in the term "document indexer" (e.g. Elasticsearch, Solr), in that documents are organized in a way that they can be efficiently accessed.

In the following subsections we describe the process of corpora acquisition performed by several partners to cover the needs specified by the business cases, and the type and number of documents compiled. The format of the original data is also specified, and, whenever performed, the result of extracting candidate terms from documents. After that, we also describe the general legislation corpora acquired directly through the openlaws.com platform for different legislations. Finally, we refer to the method followed to index the compiled resources.

1.1 CONTRACT CORPORA

The contract corpora collected in Lynx are intended to fulfil the needs of Business Case 1 “Compliance Assurance Services for Contracts”. In the current step, only a certain type of contracts is chosen (rental contracts). The pilot showcase will analyze rental contracts and extract the necessary information from them in order to be compliant with the current legislation in Austria.

There are two main sets of corpora:

- Rental contracts: mainly in German, but there are also some in English. Other languages are currently out of scope. These contracts are private documents of a real client of the Openlaws partner. For this reason, they have to be handled differently than other documents. The contracts cannot be stored within Lynx. For training the system, a set of anonymized documents will be used, but also this training set cannot be made public. The training set includes approx. 100 different documents of different types of rental contracts:
 - Flat, House
 - Offices
 - Garage, Parking slot
 - Shopping center
- Legal corpora: based on the National and Federal law and the Jurisdiction of Austria. The following is an overview of the Austrian legislation that is taken into account for rental contracts.
 - Main corpora:
 - ABGB §§ 1090 - 1107; 1109-1121

¹ <http://lynx-project.eu/data2/data-models>

- Mietrechtsgesetz
- Wohnungseigentumsgesetz 2002
- Heizkostenabrechnungsgesetz
- Additional corpora
 - Maklergesetz
 - Verordnung des Bundesministers für wirtschaftliche Angelegenheiten über Standes- und Ausübungsregeln für Immobilienmakler
 - Konsumentenschutzgesetz §§ 30a-31
 - Wohnungsgemeinnützigkeitsgesetz
 - Bauträgervertragsgesetz
 - Bundesgesetz über die Festsetzung des Richtwertes für die mietrechtliche Normwohnung
 - Wohnhaus-Wiederaufbaugesetz
 - Kleingartengesetz
 - Landpachtgesetz
 - Sportstättenchutzgesetz
 - Zivilprozessordnung §§ 560-576
 - Energieausweis-Vorlage-Gesetz 2012
 - Gewerbeordnung 1994
 - Wohnrechtsänderungsgesetz
- Case law corpora:
 - All decisions linked to the above legislation.

The legislation / case law is available in openlaws.com internal document format and is converted into the Lynx defined format by a tool (see Section 1.5).

The format for rental contracts is PDF, either already from an electronic version of the contract or a scan and OCR of the contract. Scanning and OCR are out of the scope of this project. These documents are also passed from openlaws.com to Lynx for analysis.

1.1.1 Extracted terminology

Terminology extraction is still in progress, as the corpus of rental contracts is currently being prepared.

The workflow will be similar as the one followed for the other two pilots: lists of candidate terms are to be generated with the Tilde Terminology Extraction service. Such a service relies on the TaaS (Terminology as a Service) API, which, in its turn, makes use of state-of-the-art terminology extraction techniques.

The TaaS platform was born from an European FP7 project led by Tilde. In addition to the term extraction services, the TaaS platform also provides automatic acquisition of translation equivalents and facilities for cleaning, sharing and reusing terminology. The TaaS API Specification can be found online².

Based on this service, Lynx extracted terminologies contain entries modeled as `skos:concept`, composed by the term itself, represented with the property `skos:prefLabel`; the source from which the term has been extracted is represented as an `rdfs:comment`; the confidence of each entry,

² https://term.tilde.com/Content/api_spec.pdf

represented with `itsrdf:taConfidence`; and examples of usage context for each term, represented as `lynxlang:hasExample`.

More detailed information and actual examples of the extracted entries are presented in the Extracted Terminology sections (1.2.1 and 1.3.1) of the remaining pilots.

1.2 STANDARDS, RECOMMENDED PRACTICES, REGULATIONS, AND RESOLUTIONS

In Business Case 2, Compliance Assurance Services in Oil & Gas and Energy, lead by DNV-GL, documents uploaded by the user will be analyzed, and after that documents from the LKG, will be recommended to the user. The uploaded document will be a proposal for a project in the field of geothermal energy, and will include, for example, mentions of the location where the project will take place, the technologies and processes being employed, or the names of organizations involved. After this uploaded document is analyzed, matching documents will be recommended to the user, for example, regulations applicable to this jurisdiction, or best practices of the particular processes involved.

To comply with the use case above, documents have to be collected that contain regulatory and compliance information that might be relevant to the uploaded document. A selection of document sources and types was made guided by DNV-GL needs.

The assembled corpus consists of four types of documents: (a) Standards (b) Recommended Practices, (c) Regulations, and (d) Government resolutions. Documents of types (a) and (b) are produced by DNV-GL experts, based on several decades of experience, and have been pre-categorized, being only the “Energy” and “Oil and Gas” categories relevant for this project. Documents of type (c) are published by the EU Publications Office (as part of EurLex) and local official legislation publishers (in this case, we consider the Dutch Mining Act). Documents of type (d) are published by the relevant local authorities, in this case, the Dutch Ministry of Economic Affairs and Climate Policy.

	Energy	Oil and Gas	Total
Standards	13	17	30
Recommended Practices	11	76	87
Regulations			3
Resolutions			409

Table 1. Number of documents by document type in business case 2

Standards and Recommended Practices are accessible from the DNV-GL website³, all in English language, and they are produced by the different departments of DNV-GL. In terms of Regulations, Directives 2009/28/EC and 2007/2/EC have been considered, as well as the *Dutch mining act*⁴. Resolutions are harvested from the Dutch Ministry of Economic Affairs and Climate Policy⁵. The listed documents all come as PDF files, in a variety of layouts and types of content (text, pictures, mathematical formulae, diagrams, and tables), which makes parsing text a daunting task.

³ <https://rules.dnvgl.com/servicedocuments/dnvgl/#!/home>

⁴ <https://zoek.officielebekendmakingen.nl/stb-2002-542.html>

⁵ <https://www.sodm.nl>

1.2.1 Extracted Terminology

Terminology extraction from collected corpora was performed. The result is two SKOS monolingual taxonomies – one for the English corpus and another for the Dutch corpus. The taxonomy created for the English corpus consists of 5,275 candidate concepts, while the taxonomy for the Dutch corpus consists of 3,059 candidate concepts. Each candidate concept consists of term, score, and a list of exemplifying contexts (mentions in the text, 25 words before the term, and 25 words after the term).

```
<https://term.tilde.com/lynx/termExtraction/a6fec492-d76b-4287-b037-06a7d009bfe5/#6>
a skos:Concept;
skos:prefLabel "storage tank"@en;
itsrdf:taConfidence 0.31;
lynxlang:hasExample "oil pumps and gas compressors and closing of isolation valves between ship 's
cargo and/or vapour system and the vapour recovery system . Also valves on liquid and vapour lines '
connection to any <hi>storage tank</hi> (s) for liquefied gas , if applicable, shall close. 1) If
separators have drain to slop/ cargo tanks . In order to protect against high pressure gas blow"@en;
rdfs:comment "Source: DNVGL-CG-0042.txt";
lynxlang:hasExample "3.3.8 Verify that the liquid level indicator is reading at the proper level. MSC
.1/ Circ .1318, par . 4.2.2 3.3.9 Verify that the manually operated <hi>storage tank</hi> main
service valve is secured in the open position. MSC .1/ Circ .1318, par . 4.2.3 By (see Sec .1 [6])
Remark All To be"@en;
rdfs:comment "Source: DNVGL-CG-0058.txt";
lynxlang:hasExample "in each is above 90% of the nominal charge. Cylinders containing less than 90%
of the nominal charge shall be refilled. The liquid level of low pressure <hi>storage tanks</hi>
shall be checked to verify that the required amount of carbon dioxide for protection against the
largest hazard is available. 3.3 CO2 fireextinguishing systems Regulation 3.3.18"@en;
rdfs:comment "Source: DNVGL-CG-0058.txt";
lynxlang:hasExample "proper range. MSC .1/ Circ .1432, par . 5.3 All Monthly Crew 3.6.2 Verify that
the proper quantity of foam concentrate is provided in the foam system <hi>storage tank</hi> . MSC
.1/ Circ .1432, par . 6.2 All Quarterly Crew 3.6.3 Visually inspect all accessible components for
proper condition . MSC .1/ Circ .1432, par . 7.4.1"@en;
rdfs:comment "Source: DNVGL-CG-0058.txt".
```

Figure 1. Example of the extracted candidate concept “storage tank”

1.3 LABOR LAW CORPORA

This section describes the corpora acquired in relation to Labor Law. The Labor Law corpora collected in Lynx are intended to fulfil the needs of Business Case 3, Compliance Assurance Services in Labor Law. The resulting pilot will showcase the access to interlinked relevant legal information in the labor law sector across multiple orders, jurisdictions, and languages. CUATRECASAS, the Spanish law firm participating in this project, leads the development of this pilot.

According to CUATRECASAS, the main types of documents in the labor law domain are (i) legislation at EU level and Member State level; (ii) case law (judgements related to labor law in the different jurisdictions; (iii) collective bargaining agreements (official documents, agreed upon between unions and business, that determine the conditions of work for a specific sector and function at the same level as ordinary laws) and (iv) employment contracts (standard binding contracts between workers and companies), and customs (other legal documents with special features).

To accomplish the purpose of accessing labor law information across jurisdictions, we created the following collections, as explained below:

- EU labor law legislation and case law
- Spanish labor law legislation, case law and collective bargaining agreements
- Austrian labor law legislation, case law and collective bargaining agreements
- Italian labor law legislation
- UK labor law legislation, case law and collective bargaining agreements
- Irish labor law and case law

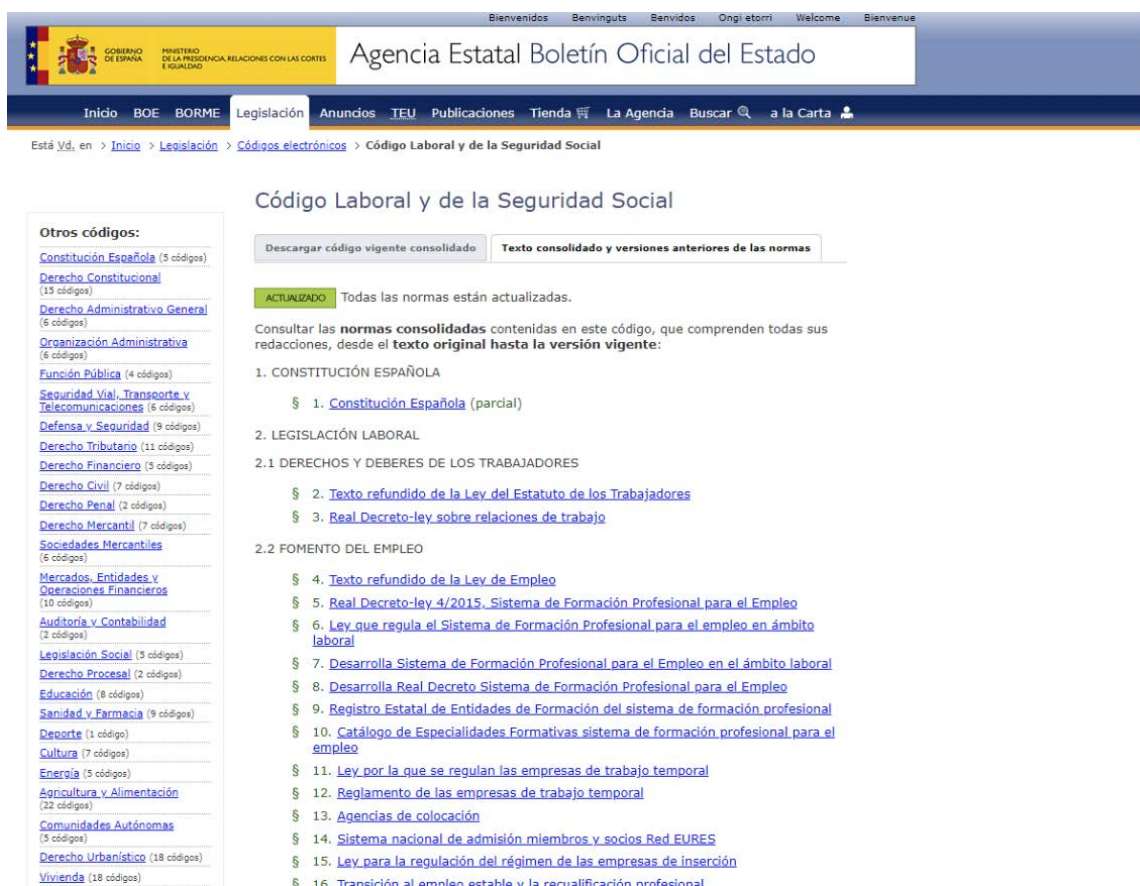
The choice of these jurisdiction has been made considering the interest of CUATRECASAS and the interest of having.

EU labor law legislation and case law

EU legislation, as described in Section 1.5, is taken directly from openlaws.com. Within openlaws.com, a set of regulations has been identified as relevant for labor law and is publicly available⁶. The list of main EU labor law legislation and case law is available in Annex 1. In addition, more than 100 decisions of the Court of Justice have been identified, which are also available in the collection of openlaws.com.

Spanish labor law legislation, case law and collective bargaining agreements

In Spain, legislation is published in the Official Gazette website or BOE⁷. Documents in the BOE are not available for bulk download, but within this website, there is a section⁸ devoted to labor law (see Figure 2), which has been used to guide the downloading of the documents (available in XML and PDF formats).



The screenshot shows the website 'Agencia Estatal Boletín Oficial del Estado'. The main content is titled 'Código Laboral y de la Seguridad Social'. It features a navigation menu with options like 'Inicio', 'BOE', 'BORME', 'Legislación', 'Anuncios', 'TEU', 'Publicaciones', 'Tienda', 'La Agencia', 'Buscar', and 'a la Carta'. Below the navigation, there is a breadcrumb trail: 'Está Vd. en > Inicio > Legislación > Códigos electrónicos > Código Laboral y de la Seguridad Social'. The main content area includes a section for 'Otros códigos:' with links to various legal codes such as 'Constitución Española', 'Derecho Constitucional', 'Derecho Administrativo General', etc. The main section is titled 'Código Laboral y de la Seguridad Social' and contains a list of sections: 1. CONSTITUCIÓN ESPAÑOLA, 2. LEGISLACIÓN LABORAL, and 2.2 FOMENTO DEL EMPLEO. Each section has a corresponding link to the consolidated text.

Figure 2. Compilation of Labor law related documents in the Spanish Official Gazette website

Case law is published in Spain by CENDOJ⁹ (*Centro de Documentación Judicial del Consejo General del Poder Judicial*), the National Center for Judicial Documentation of Spain. Documents cannot be downloaded in bulk, and harvesting is not possible due to legal restrictions. Here the search was restricted

⁶ <https://openlaws.com/public-folder-categories/4c44ce46-6ebf-4ef0-aeef-29f0e1427b48>

⁷ <http://www.boe.es>

⁸ <https://www.boe.es/legislacion/codigos/codigo.php?id=93&modo=1¬a=0&tab=2>

⁹ <http://www.poderjudicial.es/>

to case law issued by the Spanish Supreme Court, using keywords such as “laboral” (labor) and “empleo” (employment) (see Figure 3).

Due to the legal restrictions, a small set of documents was downloaded for research purposes (e.g. characterization of documents, design of data models) but not published.

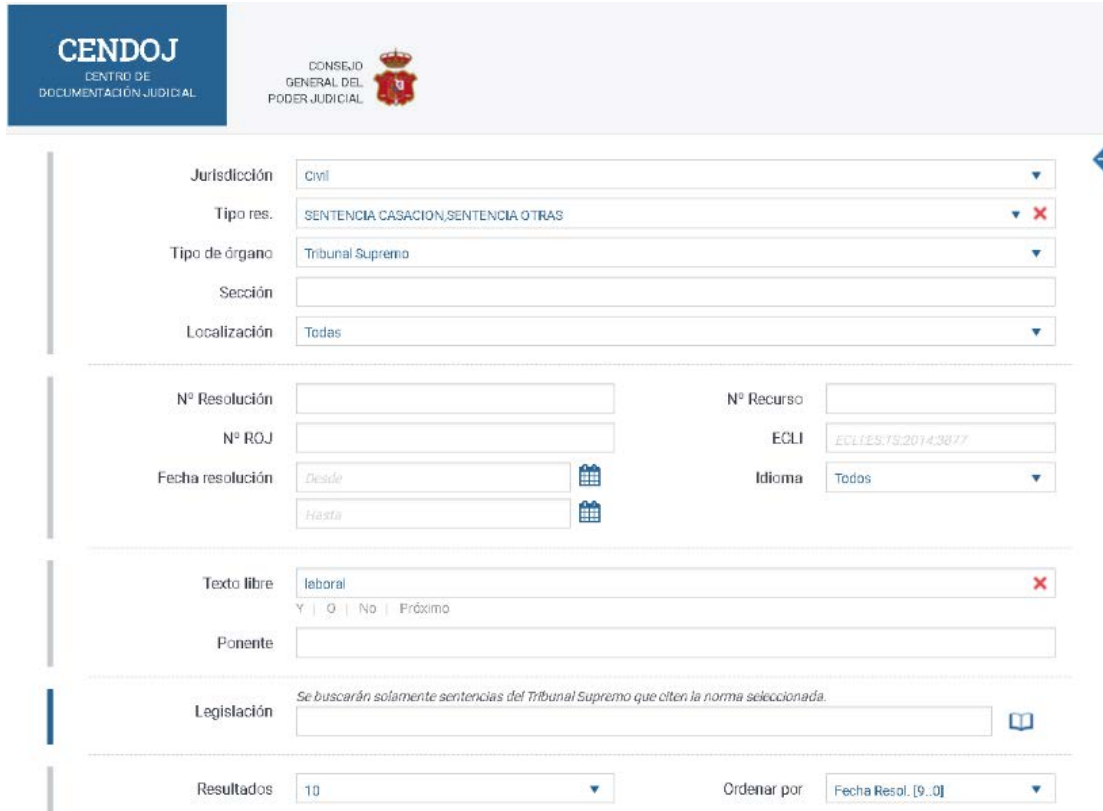


Figure 3. Spanish Center for Judicial Documentation website

In Spain, collective bargaining agreements are published by different authorities at national, regional (Autonomous Communities) and province level. We downloaded in bulk all sectoral agreements available for all provinces in Spain from the REGCON website¹⁰, the Register for Collective Agreements from the Ministry of Labor, Migration, and Social Affairs (see Figure 4).

After filtering agreements by sectors, we obtained structured information with the name of the agreement, province or autonomous community in which it applies, issuing date, period of validity, and URL (see Figure 5). Before downloading them, we manually removed some that were unreadable, as well as those written in languages other than Spanish. This has mainly to do with the fact that right now we do not have the required NLP tools at our disposal to process documents in the other official languages in Spain (Catalan, Basque, Galician), but we hope to include them at a later stage in the project. The number of collective agreements obtained with this procedure was 351. In addition to this, CUATRECASAS provided a set of 20 collective bargaining agreements from companies. Finally, 6888 collective agreements at Spanish national level were obtained from BOE but not used for the terminology extraction.

¹⁰ <https://expinterweb.empleo.gob.es/regcon/pub/consultaPublicaEstatual>

Castellano | Catalán | Euskera | Gallego | Valenciano

Inicio | Consulta de Trámites

Consulta de Trámites

* Puede seleccionar más de un elemento de la lista manteniendo pulsada la tecla Control o Mayúsculas.

Código del Acuerdo/Localizador:

Denominación:

Tipo de Trámite*
 PROMOCIÓN DE HEGOCIACIÓN
 DENUNCIA
 NUEVO ACUERDO
 TEXTO NUEVO

Estado de Vigencia*
 NO DENUNCIADO
 DENUNCIADO
 TEXTO DEROGADO POR OTRO POSTERIOR
 ANULADO POR RESOLUCIÓN JUDICIAL

Fecha de Inscripción / Publicación Desde: Fecha de Inscripción / Publicación Hasta:

Naturaleza*
 CONVENIO COLECTIVO
 ADHESIÓN A CONVENIO COLECTIVO
 LAUDO ARBITRAL
 ACUERDOS DE MEDIACIÓN
 ACUERDOS DE FIN DE HUELGA
 EXTENSIÓN DE CONVENIO
 ACUERDO MARCO
 ACUERDO COLECTIVO PARA EMPRESAS SUJETAS

Ámbito Funcional*
 CONVENIOS O ACUERDOS FRANJA
 UNO O VARIOS CENTROS DE LA EMPRESA
 EMPRESA O TODOS LOS CENTROS DE UNA EMPRESA
 GRUPO DE EMPRESAS / EMPRESAS VINCULADAS

Autoridad Laboral: *
 Álava
 Albacete
 Alicante/Albacic
 Almería

CNAE*
 0 - TODAS
 01 - Agricultura, ganadería, caza y servicios relacionados con las mismas
 011 - Cultivos no perennes
 0111 - Cultivo de cereales (excepto arroz), leguminosas y semillas oleaginosas

Figure 4. Website of the Spanish Register for Collective Agreements

	B	C	D	E	F	G	H	I	J	K
	Denominación	Tipo de Trámite	Autoridad Laboral	Inscripción / Vigencia	Desdénfencia	Has	URL	Boletín		
2	SIDEROMETALURGICAS (INDUSTRIAS)	CONVENIO COLECTIVO	Navarra	26/02/2019	01/01/2018	31/12/2021	http://www.navarra.es/home_es/Actualidad/BON/Bole			
3	SERVICIOS DE AYUDA A DOMICILIO	CONVENIO COLECTIVO	Zaragoza	25/02/2019	01/01/2018	31/12/2022	http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD="			
4	DERIVADOS CEMENTO	CONVENIO COLECTIVO	Rioja (La)	20/02/2019	01/01/2017	31/12/2020	https://www.larioja.org/bor/es/ultimo-boletin?tipo=2&l			
5	INDUSTRIA, SERVICIOS E INSTALACIONES DEL METAL DE LA COMUNI	CONVENIO COLECTIVO	Madrid	14/02/2019	01/01/2018	31/12/2020	http://www.bocm.es/boletin/CM_Ordern_BOCM/2019/C			
6	HOSTELERIA	CONVENIO COLECTIVO	Santa Cruz de Te	13/02/2019	01/07/2018	30/06/2022	http://www.bopsantacruztenerife.org/2019/02/019/			
7	Pastelería, bollería, galletas, repostería, elaboración de productos de	CONVENIO COLECTIVO	Huesca	13/02/2019	01/01/2018	31/12/2021	http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD="			
8	TRANSPORTES POR CARRETERA, GRUPOS DE TRACCION MECANICA Y	CONVENIO COLECTIVO	Vizcaya	13/02/2019	01/01/2017	31/12/2020	http://www.bizkaia.eus/lehendakaritza/Bao_bob/2019/			
9	CONSERVAS Y SALAZONES DE PESCADO	CONVENIO COLECTIVO	Vizcaya	13/02/2019	01/01/2018	31/12/2020	http://www.bizkaia.eus/lehendakaritza/Bao_bob/2019/			
10	FABRICANTES DE YESOS, ESCAYOLAS, CALES Y SUS PREFABRICADOS.	CONVENIO COLECTIVO	Estatal	13/02/2019	01/01/2018	31/12/2021	https://www.boe.es/boe/dias/2019/02/13/pdf/s/BOE-A-			
11	HOSTELERIA	CONVENIO COLECTIVO	Zaragoza	12/02/2019	01/01/2018	31/12/2020	http://www.boa.aragon.es/cgi-bin/EBOA/BRSCGI?CMD="			
12	COMERCIO EN GENERAL DEL PRINCIPADO DE ASTURIAS	CONVENIO COLECTIVO	Asturias	11/02/2019	01/01/2018	31/12/2020	https://sede.asturias.es/bopa/2019/02/11/2019-00727.j			
13	COMERCIO ALMACENISTAS DE COLONIALES	CONVENIO COLECTIVO	Cantabria	11/02/2019	01/01/2018	31/12/2021	https://boc.cantabria.es/boces/verAnuncioAction.do?id			
14	COMERCIO DETALLISTAS DE ALIMENTACION	CONVENIO COLECTIVO	Cantabria	11/02/2019	01/01/2018	31/12/2022	https://boc.cantabria.es/boces/verAnuncioAction.do?id			
15	AUTO-TAXI DE LA COMUNIDAD AUTONOMA DE ANDALUCIA	CONVENIO COLECTIVO	Andalucia	08/02/2019	01/01/2018	31/12/2020	https://juntadeandalucia.es/boja/2019/27/BOJA19-027-			
16	DERIVADOS CEMENTO	CONVENIO COLECTIVO	Segovia	06/02/2019	01/01/2017	31/12/2020	https://www.dipsegovia.es/documents/963029/efb696			
17	SUPERMERCADOS AUTOSERVICIOS Y DETALLISTAS DE ALIMENTACION	CONVENIO COLECTIVO	Alicante/Allicant	04/02/2019	01/01/2018	31/12/2019	http://www.dip.alicante.es/bop2/pdftotal/2019/02/04_			
18	INDUSTRIAS DE ALIMENTOS COMPUESTOS PARA ANIMALES	CONVENIO COLECTIVO	Estatal	04/02/2019	01/01/2018	31/12/2019	https://www.boe.es/boe/dias/2019/02/04/pdf/s/BOE-A-			
19	Industrias Vinícolas y Alcohólicas de Ciudad Real	CONVENIO COLECTIVO	Ciudad Real	04/02/2019	01/01/2018	31/12/2019	http://bop.sede.dipucr.es/bop/2019/02/04			
20	INDUSTRIAS DE TINTORERIAS, LAVANDERIA Y PLANCHADO DE ROPAS	CONVENIO COLECTIVO	Sevilla	02/02/2019	01/01/2017	31/12/2020	http://www.dipusevilla.es/system/modules/es.dipusevi			
21	COMERCIO DE VIDRIO Y CERAMICA	CONVENIO COLECTIVO	Valenciá/Valenci	31/01/2019	01/01/2018	31/12/2018	https://bop.dival.es/bop/drvtsapi.dll?MVal=DI_VerEdict			
22	ALMACENISTAS DE ALIMENTACION AL POR MAYOR	CONVENIO COLECTIVO	Castellón/Castell	29/01/2019	01/01/2017	31/12/2018	https://bop.dipcas.es/PortalBOP/obtenerPdfAnuncio.do			

Figure 5. Spreadsheet with sectoral collective agreements in Spain

Austrian labor law legislation, case law, and collective bargaining agreements

Austrian legislation related to labor law has been also taken from openlaws.com. In total, more than 200 Regulations have been identified as relevant for labor law.

The main documents are listed the following:

- AAV - Allgemeine Arbeitnehmerschutzverordnung
- ABVO - Arbeitsbescheinigungsverordnung
- AEntG 2009 - Arbeitnehmer-Entsendegesetz
- AltTZG 1996 - Altersteilzeitgesetz
- AIVG - Arbeitslosenversicherungsgesetz 1977
- AM-VO - Arbeitsmittelverordnung
- AngG - Angestelltengesetz
- APfIG - Ausbildungspflichtgesetz
- APG - Allgemeines Pensionsgesetz
- Arbeiter-Abfertigungsgesetz

- Arbeitszeitgesetz - Arbeitszeitverkürzung - Art. 1
- ArbIG - Arbeitsinspektionsgesetz 1993
- ArbNErfG - Gesetz über Arbeitnehmererfindungen
- ArbPlatSchG - Arbeitsplatzschutzgesetz
- ArbSchG - Arbeitsschutzgesetz
- ArbVG - Arbeitsverfassungsgesetz
- ArbZG - Arbeitszeitgesetz
- ArbZRG - Arbeitszeitrechtsgesetz
- ARG - Arbeitsruhegesetz
- ARG-VO - Arbeitsruhegesetz-Verordnung
- AÜG - Arbeitnehmerüberlassungsgesetz
- AuslBG - Ausländerbeschäftigungsgesetz
- AVO Verkehr 2017 - ArbeitnehmerInnenschutzverordnung Verkehr 2017
- AVRAG - Arbeitsvertragsrechts-Anpassungsgesetz
- AZG - Arbeitszeitgesetz
- AZHG - Auslandszulagen- und -hilfeleistungsgesetz
- AZV - AIVG-Auszahlungsverordnung
- EFZG - Entgeltfortzahlungsgesetz
- MiLoG - Mindestlohngesetz
- MuSchG - Mutterschutzgesetz
- NastV - Nadelstichverordnung
- NSchG - Nachtschwerarbeitsgesetz

Austrian collective bargaining agreements have been collected from the website of the Austrian Chamber of Commerce (Wirtschaftskammern Österreichs), which organizes them into seven different sectors:

- Commerce and Craftworks
- Commerce
- Industry
- Information and Consulting
- Tourism and Entertainment
- Transport and Traffic
- Bank and Insurance

From each of these sectors, we have selected the most representative topics that contain the desired information. Thus, we have discarded outdated documents and other types of data that are not required

at this phase of the project: salary tables, summaries, and additional information (Figure 6). As a result, a corpus of 50 PDF files in German language has been created.

Alle Kollektivverträge

Branche *

- > **Beilage zum Kollektivvertrag für das Dachdeckergewerbe, Lohnordnung gültig ab 1.5.2018**
inkl. der aktuellen Lohn tafeln
- > **Informationen zum Kollektivvertrags-Abschluss für ArbeiterInnen im Metallgewerbe 2019**
Die Verhandlungsergebnisse im Überblick
- > **Informationen zum Kollektivvertrags-Abschluss für Angestellte im Metallgewerbe 2019**
Die aktuellen Gehaltstafeln
- > **Kollektivvertrag für ArbeiterInnen im Dachdeckergewerbe - Stand 1.5.2016**
inkl. der Lohn tafeln
- > **Kollektivvertrag für Angestellte im Handwerk und Gewerbe, in der Dienstleistung in Information und Consulting gültig ab 1.1.2019**
inkl. der aktuellen Gehaltstafeln
- > **Gehaltstabellen für Angestellte im Handwerk und Gewerbe, in der Dienstleistung, in Information und Consulting gültig ab 1.1.2019**
Gehälter nach Verwendungsgruppen und Lehrlingsentschädigungen
- > **Zusatzkollektivvertrag zum Kollektivvertrag für Angestellte des Metallgewerbes vom 27.11.2017, in Kraft seit 1.1.2018**
inkl. der aktuellen Gehaltstafeln gültig ab 1.1.2019

Figure 6. Example of Austrian collective agreements per sector

Italian labor law legislation

Italian legislation is officially published at <http://normattiva.it> but it is not available for bulk download nor easily scappable by a web spider. However, Wikiversity provides links to different sources from which labor law can be collected (see Figure 7). The retrieved collection comprises 32 PDF documents.

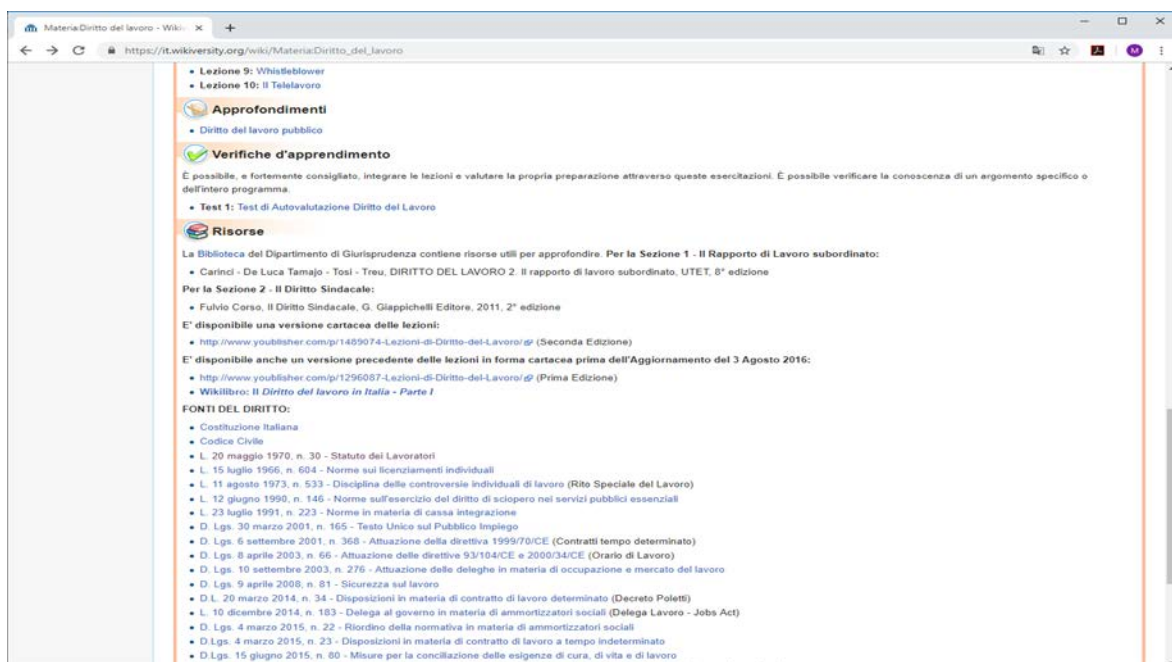
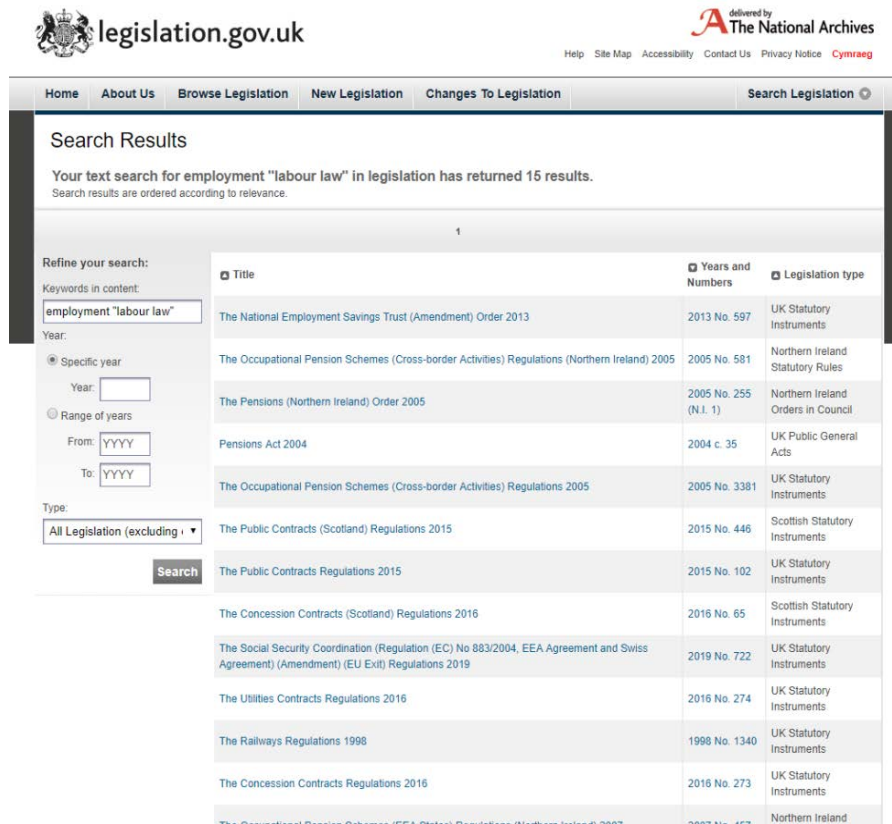


Figure 7. Wikiversity website with links to Italian Labor law

UK labor law legislation, case law, and collective bargaining agreements

In the UK, legislation can be accessed via an official website with well structured content (Metalex RDF is used).¹¹ We used keywords such as labor law and employment to search for Labor law related documents (see Figure 8). We retrieved 21 documents.



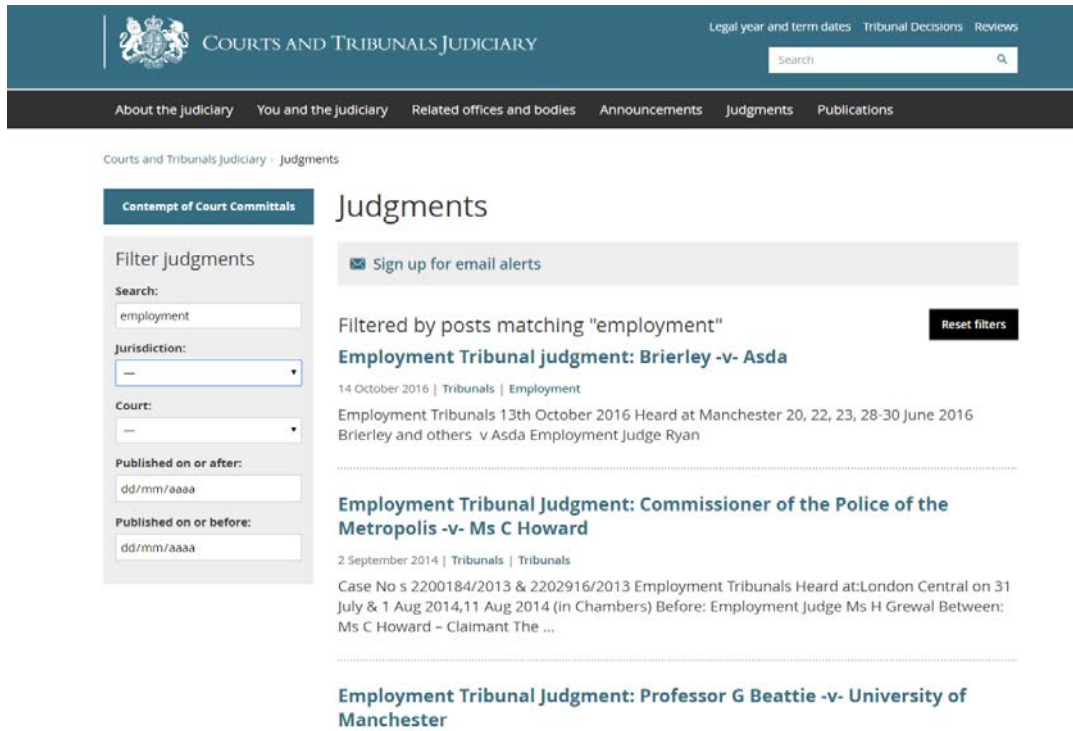
The screenshot shows the search results page on legislation.gov.uk. The search query was "employment 'labour law'". The results are sorted by relevance and show 15 results. The table below represents the visible results from the screenshot.

Title	Years and Numbers	Legislation type
The National Employment Savings Trust (Amendment) Order 2013	2013 No. 597	UK Statutory Instruments
The Occupational Pension Schemes (Cross-border Activities) Regulations (Northern Ireland) 2005	2005 No. 581	Northern Ireland Statutory Rules
The Pensions (Northern Ireland) Order 2005	2005 No. 255 (N.I. 1)	Northern Ireland Orders in Council
Pensions Act 2004	2004 c. 35	UK Public General Acts
The Occupational Pension Schemes (Cross-border Activities) Regulations 2005	2005 No. 3381	UK Statutory Instruments
The Public Contracts (Scotland) Regulations 2015	2015 No. 446	Scottish Statutory Instruments
The Public Contracts Regulations 2015	2015 No. 102	UK Statutory Instruments
The Concession Contracts (Scotland) Regulations 2016	2016 No. 65	Scottish Statutory Instruments
The Social Security Coordination (Regulation (EC) No 883/2004, EEA Agreement and Swiss Agreement) (Amendment) (EU Exit) Regulations 2019	2019 No. 722	UK Statutory Instruments
The Utilities Contracts Regulations 2016	2016 No. 274	UK Statutory Instruments
The Railways Regulations 1998	1998 No. 1340	UK Statutory Instruments
The Concession Contracts Regulations 2016	2016 No. 273	UK Statutory Instruments
The Occupational Pension Schemes (Cross-border Activities) Regulations (Northern Ireland) 2005	2005 No. 581	Northern Ireland

Figure 8. Official website for accessing UK legislation

The UK also has an official website to access case law. Again, we used keywords such as employment and labor (see Figure 9) to search for relevant documents. The retrieved collection comprises 112 PDF files randomly selected (and manually revised) from the search result.

¹¹ <https://www.legislation.gov.uk/>



The screenshot shows the official website for Courts and Tribunals Judiciary. The header includes the Royal Coat of Arms and the text 'COURTS AND TRIBUNALS JUDICIARY'. A search bar is visible in the top right. The main navigation menu includes 'About the judiciary', 'You and the judiciary', 'Related offices and bodies', 'Announcements', 'Judgments', and 'Publications'. The page title is 'Judgments'. A filter sidebar on the left allows searching for 'employment' and filtering by jurisdiction and court. The main content area displays a list of judgments, including 'Employment Tribunal judgment: Brierley -v- Asda' (14 October 2016) and 'Employment Tribunal Judgment: Commissioner of the Police of the Metropolis -v- Ms C Howard' (2 September 2014).

Figure 9. Official website for accessing UK case law

As for collective agreements, we could download a small set of company agreements (13 documents) available from the Wage Indicator website.¹²

Irish labor law and case law

In Ireland, citizens can get information about employment at the Citizens information web portal¹³, which we used to help us identify relevant labor law and then collect it from the electronic Irish Statute Book¹⁴ (see Figure 10).

¹² <https://wageindicator.co.uk/advice/collective-agreements-database/compare-clauses>

¹³ <https://www.citizensinformation.ie/en/employment/>

¹⁴ <https://www.citizensinformation.ie/en/employment/>

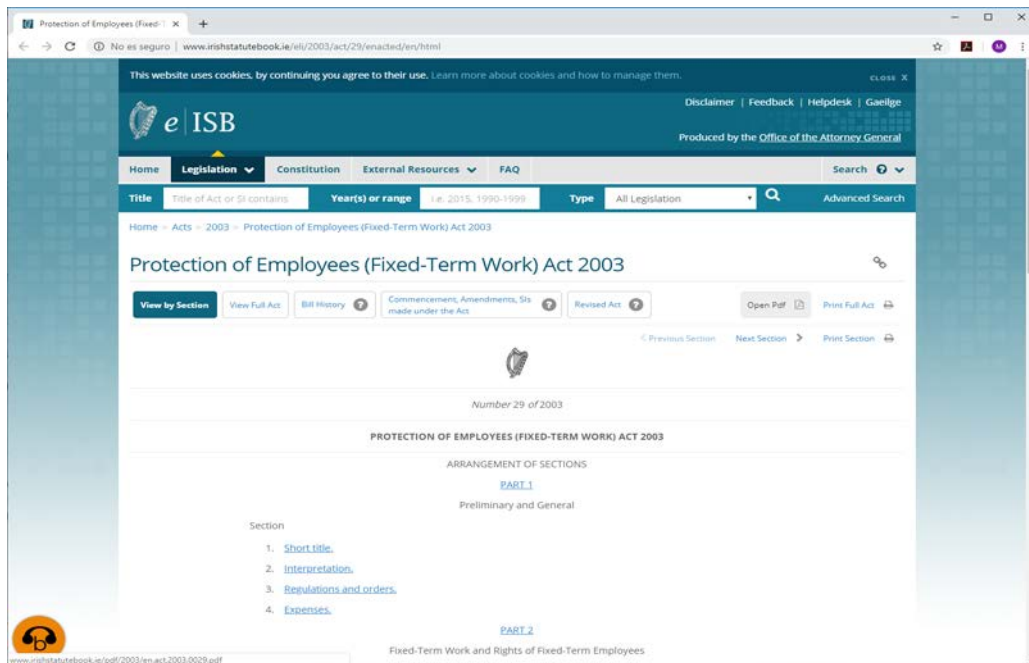


Figure 10. Irish legislation accessible from the Irish Statute Book

Irish cases regarding labor law can be accessed from the Workplace Relations website¹⁵, which we used to harvest documents differently dated (see Figure 11). The retrieved collection comprises 20 PDF files randomly selected (and manually revised) from the search result.

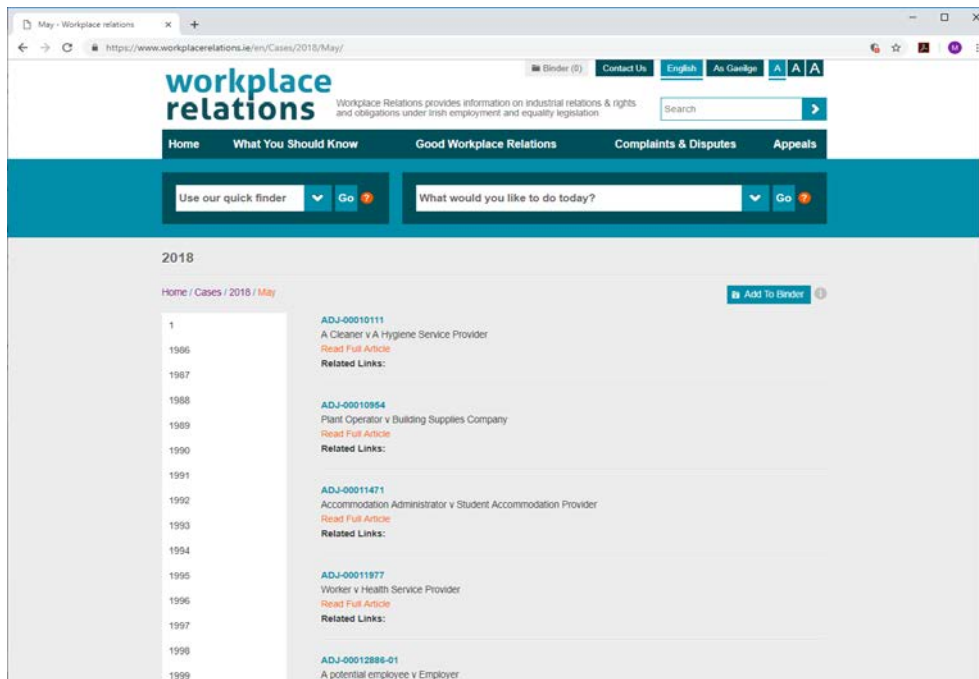


Figure 11. Irish case law accessible from Workplace Relations website

The documents were collected in PDF format and converted to plain text.

¹⁵ <https://www.workplacerelations.ie/en/Cases/>

	<i>Collective Agreement</i>	<i>Legislation</i>	<i>Judgments</i>	Total Size
<i>Austria</i>	50 PDF files	13 PDF files	-	20.8 MB
<i>Ireland</i>	-	4 PDF files	20 PDF files	6 MB
<i>Italy</i>	-	32 PDF files	-	7 MB
<i>Spain</i>	371 PDF files	50 PDF files	27 PDF files	359 MB
<i>United Kingdom</i>	13 PDF files	21 PDF files	112 PDF files	119.3 MB

Table 2. Summary of size of each component of the corpus for business case 3

1.3.1 Extracted Terminology

A first terminology extraction task was performed over the collected corpora, resulting in a list of SKOS concepts.

Table 3 shows the size of corpora and resulting collection of term candidates. As in the case of the DNV-GL corpora, each concept consists of term, score, and a list of example contexts (mentions in the text, 25 words before the terms, and 25 words after the term).

	<i>Collective Agreement</i>	<i>Legislation</i>	<i>Judgments</i>	Total
<i>Austria</i>	1,592 concepts	2,504 concepts	-	4,096 concepts
<i>Ireland</i>	-	857 concepts	607 concepts	1,464 concepts
<i>Italy</i>	-	3,436 concepts	-	3,436 concepts
<i>Spain</i>	20,600 concepts	12,797 concepts	1,114 concepts	34,511 concepts
<i>United Kingdom</i>	1,399 concepts	7,473 concepts	8,453 concepts	17,325 concepts

Table 3. Number of extracted candidate concepts

As in the case of the DNV-GL corpora, each candidate concept consists of term, score, and a list of example contexts (mentions in the text, 25 words before the term, and 25 words after the term). In the example in Figure 12 there is a concept with an English term “Holiday Scheme”, it has confidence score, and three context examples coming from two different source documents.

```

<https://term.tilde.com/lynx/termExtraction/6f2b3046-be32-4f31-8739-6c125e1a0111/#793>
a skos:Concept;
    skos:prefLabel "Holiday Scheme"@en;
    itsrdf:taConfidence 0.14;
    lynxlang:hasExample "Colleagues on these contracts will only need to use their
holiday allowance to cover their core hours /days. Additional flexible hours
will not be scheduled on a bank holiday . <hi>Holiday Schemes</hi> Historically,
there are 5 holiday schemes . Colleagues will be on the holiday scheme that
applied at the time they joined. 1. Current Year Scheme Colleagues who"@en;
    rdfs:comment "Source: Our+Partnership+Agreement+with+Usdaw.pdf";
    lynxlang:hasExample "on these contracts will only need to use their holiday
allowance to cover their core hours /days. Additional exible hours will not be
scheduled on a bank holiday . <hi>Holiday Schemes</hi> Historically, there are 5
holiday schemes . Colleagues will be on the holiday scheme that applied at the
time they joined. 1. Current Year Scheme Colleagues who"@en;
    rdfs:comment "Source:
United+Kingdom+-+TESCO+Partnership+Agreement+with+USDAW+-+2016+-+2016+-+WageIndica
tor.co.uk.pdf";
    lynxlang:hasExample "applied at the time they joined. 1. Current Year Scheme
Colleagues who joined on or after 12th January 2003 will be on the Current Year
<hi>Holiday Scheme</hi> . Entitlement is calculated on the number of days worked
each week and how much service could be achieved in the current holiday year,
between 1st April"@en;
    rdfs:comment "Source:
United+Kingdom+-+TESCO+Partnership+Agreement+with+USDAW+-+2016+-+2016+-+WageIndica
tor.co.uk.pdf".
    
```

Figure 12. Example of the extracted candidate concept “Holiday Scheme”

1.4 GENERAL LEGAL CORPORA

General legal corpora from the EU, Austria and Germany have been made available by Openlaws directly from the openlaws.com platform. The Openlaws platform is a SaaS which helps to find legal information more easily, organizes it according to the user desires and shares it with others. The user can create a network of legislation, case law, legal literature and legal experts – both on a national and a European level. Table 4 summarizes the legislation and case law compiled for such jurisdictions.

	Legislation	Case law
Austria	Federal law National law	Verfassungsgerichtshof (VfGH) Constitutional Court Verwaltungsgerichtshof (VwGH) Administrative High Court Obersten Gerichtshofes (OGH) Supreme Court Oberlandesgerichte (OLG) Higher Regional Courts Landesgerichte (LG) District Courts Bezirksgerichte (BG) Regional Courts Obersten Patent- und Markensenats (OPMS) Bundesverwaltungsgericht (BVwG) Federal Administrative Court Landesverwaltungsgerichte (LVwG) National Administrative Court
EURLex	Sector 3: Legal acts R - Regulations	Sector 6: EU case law CJ: Court of Justice - Judgment

L - Directives	TJ: General Court - Judgment
Sector 1: Treaties	
C - Treaty of Nice 2001	
D - Treaty of Amsterdam 1997	
L - Treaty of Lisbon 2007	
M - Treaty on the European Union	
Sector 0: Consolidated texts	
R - Regulations	
L - Directives	
Germany	Federal law

Table 4. Legislation and case law from openlaws.com

The corpora in openlaws.com are already split into text fragments at the article level for legislation, even if the target source, e.g., EUR-Lex, does not provide this granularity level. This is necessary for further annotations and linking. In addition, these fragments are already linked to other text fragments and/or decisions at the national and European level. This information is either extracted by a human or during the load process from the original source to openlaws.com.

openlaws's website also has historic and future data, but this is not used within Lynx. For the time being, Lynx only works with the current version (as of Spring 2019).

1.5 CORPORA INDEXING METHOD

This section offers a summary of the methods followed to index the corpora compiled for the three business cases defined in Lynx. In order to process the compiled textual resources using the services that make up the Lynx Platform and to extract knowledge that will aid in the first steps of the compliance process for a new project, documents have to be put into an agreed standardized format. This allows for all services to have access to it and for all facilities for indexing and rendering to be shareable among use cases.

For business case 1 "**Compliance Assurance Services for Contracts**" the contracts themselves will not be indexed. The text itself is extracted from the pdf with Apache Tika. The text corpus of the contracts is annotated to train the different services, but will not be part of the lynx Legal Knowledge Graph. The main idea is to extract information out of the contracts using the Lynx services.

For business case 2 "**Compliance Assurance Services in Oil & Gas and Energy**", this has been achieved through a custom-made python script using the library PDFMiner.six¹⁶ and a set of hand-crafted regular expressions that help identify titles and subtitles. While this approach implies quite a bit of manual work, it is suitable for this dataset in which most documents come in the same format. Specifically, it is necessary to define the area of the documents in which the text is found (excluding headers and footers, for

¹⁶ <https://github.com/pdfminer/pdfminer.six>

example), as well as combinations of regular expressions and font types that determine that a given piece of text is a title or a subtitle. With this in hand, the whole document can be parsed and divided into sections. We note that this approach ignores all images and considers all text that does not fit the aforementioned regular expressions as equal. This means that text in figure legends, tables, and normal paragraphs is combined. However, since there are always separations between two pieces of text, all services developed in tasks 3.2 and 3.3 work as expected, except for Summarization, which might require adjustments for this behavior.

As for **Business Case 3 “Compliance Assurance Services in Labor Law”**, a service called Document Structure Extractor (StrEx) has been created. The first version of this service is described in D3.1. This service is able to handle multiple types of documents (legislation, case law...) and can extract the different sections in them. For specific sources, such as the Spanish official state gazette, customized extractors have been implemented to recognize sections (further information on this specific service can be found in deliverable 3.1). Regular expressions frequently appearing in certain types of documents have also been identified to improve the detection of their respective structure. Additionally, a generic algorithm has been designed to extract the basic structure of other kinds of documents, such as contracts or judgments, relying on common divisions such as sections or articles. All this information about the structure of a document is stored in the system as in the following table:

Lynx Document		
Information field	Description	Value in the example
id	The unique identifier of the document. It is used to update and retrieve it, and should not contain characters such as spaces or slashes.	P13ML591
text	The raw text in the document	An Act to impose a tax on the issue, the anniversary of the coming into force and the holding of a receiver licence under the Radiocommunications Act 1992 1 Short title\t\tThis Act may be cited as the Radiocommunications (Receiver Licence Tax) Act 1983.2 Commencement\t\tThis Act shall come into operation on the date fixed for the purposes of subsection 2(1) of the Radiocommunications Act 1983.
parts	The different parts (such as articles or paragraphs, depending of the text) the document is divided in.	part01, part02
metadata	Other information about the document, such as the language, the ELI or the jurisdiction.	{ "eli": ["http://localhost:8080/portal/res/eli/au/P13ML591"], "jurisdiction": ["au"], "language": ["en"], "title": ["Radiocommunications (Receiver Licence Tax) Act 1983"], "uri": [{ "uri": "http://localhost:8080/portal/res/eli/au/P13ML591" }] }
annotations	The annotations done by the services in Lynx	(empty yet)
LynxDocumentPart		
Information field	Description	Value in the example
id	A unique identifier of the part through the document.	part01

It should not contain characters such as spaces or slashes.

offset_ini	The number of the character from the whole text of the document where the part begins.	162
offset_end	The number of the character from the whole text of the document where the part ends.	260
title	The title of the part, if any.	1 Short title
parent	It is the part that the current part belongs to (eg, the parent of section 2.1 would be section 2). If it belongs nowhere (it is in the "root" of the document), it is null.	part0

Table 4. Information about the structure of a document

The current implementation of this structure is the one below:

```
public class LynxDocument {
    private String id;           //identifier. short slug with no strange characters (/, ' ', etc.)
    private String text;        //text of the document

    ///Structured parts
    private List<LynxDocumentPart> parts = new ArrayList();

    ///Metadata elements: version, author, date, etc.
    private Map<String, List<Object>> metadata = new HashMap();

    ///Annotations
    private List<LynxAnnotation> annotation = new ArrayList();

    @JsonIgnore
    private Model annotations;

    ...
}
```

where

```
public class LynxDocumentPart {
    public String id;           //id with no strange characters
    public Integer offset_ini;  //eg. 120. Start of part
    public Integer offset_end;  //eg. 200. End of part.
    public String title;        //title, e.g. "2.1 Introduction"
    public LynxDocument/ LynxDocumentPart parent; //parent or null if root
}
```

An example of an implemented document using this class would be the following:

```
{
    "id": "P13ML591",
    "text": "An Act to impose a tax on the issue, the anniversary of the coming into force and the holding of a receiver licence under the Radiocommunications Act 1992 1 Short title\t\tThis Act may be cited as the Radiocommunications (Receiver Licence Tax) Act 1983.2 Commencement\t\tThis Act shall come
```

```
into operation on the date fixed for the purposes of subsection 2(1) of the Radiocommunications Act
1983.",
  "parts": [{
    "id": "part01",
    "offset_ini": "162",
    "offset_end": "260"
    "title": "1 Short title",
    "parent": "part0"
  },
  "parts": [{
    "id": "part02",
    "offset_ini": "261",
    "offset_end": "406"
    "title": "2 Commencement",
    "parent": "part0"
  },
  ],
  "metadata": { "eli": ["http://localhost:8080/portal/res/eli/au/P13ML591"], "jurisdiction": ["au"], "language": ["en"], "title": ["Radiocommunications (Receiver Licence Tax) Act 1983"], "uri": [{"uri": "http://localhost:8080/portal/res/eli/au/P13ML591"}]},
  "annotation": []
}
```

This representation allows to run simple text queries on the whole text (since it is stored altogether) and to keep sections in the text and move up in the structure of the document (thanks to the “parent” field) without duplicating information.

The **General Legal Corpora** for Austria, Germany, EURLex, is provided by openlaws.com. To extract the content out of openlaws.com, a web service that will be described in Deliverable 3.4 has been developed that allows to receive the legislation / decision / collection of documents via the openlaws.com REST API, and to provide it in the Lynx defined format either in RDF or JSON. In the current stage the legislation contains sub parts on an article level, so each individual article can be referenced. For case law there are no sub parts, the decision is provided as one text block at all.

The tool will also be used to feed these documents directly into the Lynx document manager on a regular (e.g., daily) basis. Details of the tool will be described in D3.4.

1.6 NIF ANNOTATION

Document annotations in the Lynx project will be recorded in the Natural Language Processing Interchange Format (NIF). This format allows the annotation of linked data and the usage of external ontologies. The data (documents) of the project are going to be converted into NIF, annotated using the semantic annotation services, and stored in the Document Manager and Legal Knowledge Graph.

A basic example of a NIF annotated document is shown in Figure 13:

```
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#> .
@prefix lynxnif: <http://persistence.lynx-project.eu/ontologies/nif-core#> .
<http://lynx-project.eu/documents/#offset_0_26>
  a nif:RFC5147String , nif:String , nif:Context ;
  nif:beginIndex "0"^^xsd:nonNegativeInteger ;
  nif:endIndex "26"^^xsd:nonNegativeInteger ;
  nif:isString "Welcome to Berlin in 2018." ;
  lynxnif:summary "Welcome" ;
  lynxnif:summaryConfidence 0.68 .
```

Figure 13. Example of NIF annotated document in the Lynx project

1.6.1. Methodology/Approach/Conversion Process

For the Lynx project, there is different information that has to be annotated, such as semantic information (named entities, temporal expressions, or terms), translations, summaries, etc. We have agreed on using, as much as possible, existing ontologies to perform the annotation of semantic information such as named entities, temporal expressions, or translations.

For the cases in which we cannot find ontologies that already include suitable predicates for any annotations, we decided to define our own linked data (NIF) annotations (predicates). The annotation properties that have been defined until now in the project are shown in Table 5. The nif prefix stands for <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core>, its for <https://www.w3.org/2005/11/its/rdf-content/its-rdf.html> and lynx-nif for <http://lynx-project.eu/ontologies/nif>.

Annotation Properties	Description
nif:anchorOf nif:beginIndex nif:endIndex nif:referenceContext	These are the common NIF properties for every annotation.
its:taClassRef	This class determines the type of the entity: <ul style="list-style-type: none"> • Person: http://dbpedia.org/ontology/Person • Location: http://dbpedia.org/ontology/Location • Organization: http://dbpedia.org/ontology/Organization
its:taIdenRef	It determines an (several) external link(s) referring to the same entity. For example, a value could be " https://www.wikidata.org/wiki/Q90 " for "Paris".
lynx-nif:geolocationSubclass	Geographical region, City area, Neighborhood, Postal address
lynx-nif:summary	Refers to the text of the created summary
lynx-nif:summaryConfidence	Refers to the confidence of the generated summary. It is a rather subjective value.
its.taConfidence	Expect a float value - confidence of annotation
its:taIdentRef	Specifies which entity is found in the current nif:phrase

Table 5. Annotation properties defined until now in the Lynx project

2 CREATED TRANSLATION CORPORA

2.1 CORPORA CREATION WORKFLOW

2.1.1 General parallel corpora creation workflow

This section contains information about the creation workflow for parallel corpora. This workflow depends on resource type, file format, content of source data, and other factors. In general, the process starts with acquiring original data, processing data into two parallel Moses files (two plaintext files with aligned sentences in utf-8 encoding), converting it to final delivery format, and performing quality evaluation (see Figure 14). Data processing includes Cleaning, Aligning, Language identification, Converting, Anonymization, Filtering, Evaluation, and other steps. If problems with data are found at any of these steps, adaptation of some previous steps and reprocessing is required.

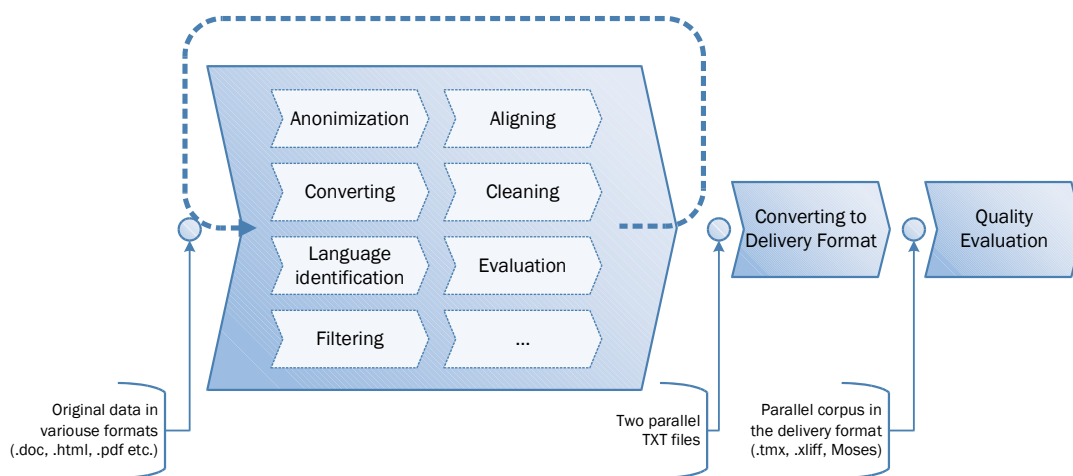


Figure 14. General language resource processing workflow used for Lynx corpora creation

Original data are in many different file formats (such as .doc, .docx, .pdf, .xls, .rtf, .odt, etc.), web content (.html files, lists of links, etc.), translation memory and other localization file formats (.tmx, .xliff, .ttx, etc.), and database dumps (.json, .xml, .sql, etc.). Delivery data is prepared in the agreed format as TMX (Translation Memory eXchange) format, which is standard in localization, or Moses format used by Moses SMT and other MT engines. Quality evaluation is done to ensure that the delivery data meets required quality requirements and does not contain translator or automatic processing errors.

2.1.2 Corpora creation from web crawled data

Web crawling that was used to compile corpus for all business cases. Depending on partners resources, this process was applied to further compile corpora. The first step in this regard was the identification of useful websites. Useful websites are the ones that contain parallel or at least comparable content in two languages. This resource identification was performed by the use case partners and yielded a list of URLs to be crawled and processed.

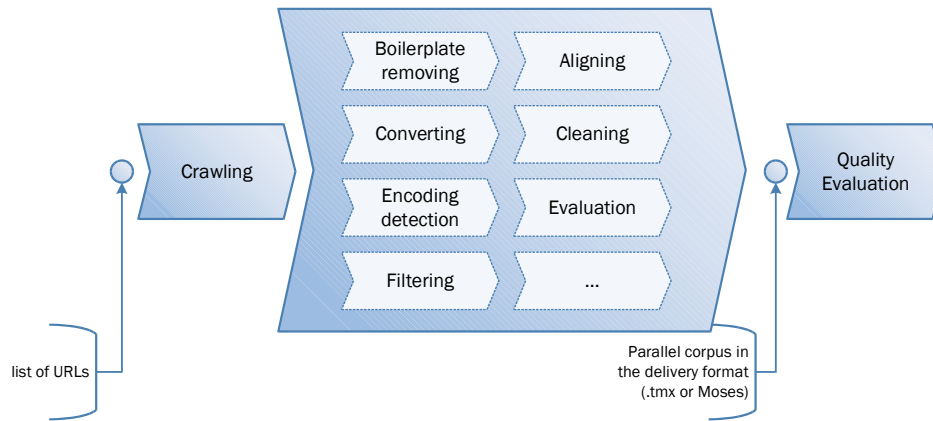


Figure 15. Web-crawled data processing workflow

In general, corpora creation start by crawling specified URLs and extracting more links from the same domain. Collected links are retrieved, contents are stored, and additional links may be extracted and collected in a loop, until no new links can be found. Web crawling yields .html, .pdf, .doc, and other file types, and for every file type, the appropriate tool are used for text extraction in correct encoding (usually utf-8), boilerplate (ads, headers, footers, etc.) removal, and metadata (title, keywords, author, publisher, etc.) extraction (see Figure15). Further processing is used for language identification, text segmentation, duplicate document removal, and anonymization (see first part of Figure 16).

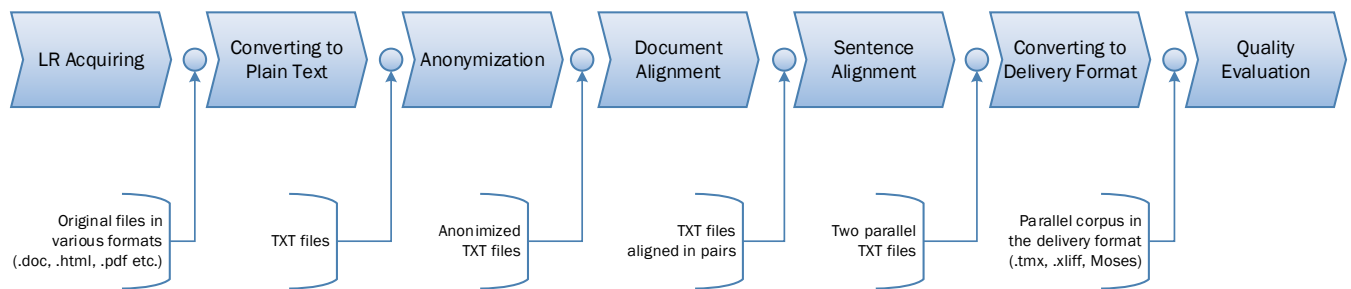


Figure 16. Parallel corpus processing workflow

Processed documents are then aligned at the document level using various heuristics, i.e., similar filenames (for .doc, .pdf, etc.), links pointing to the same document in another language (for .html), and content similarity (pictures, numbers, structure, etc.). This produces a list of paired files that should contain the same text in different languages, and each file pair is then further processed to find matching segments (sentences) in both languages. Sentence alignment is done using either Microsoft Bilingual sentence aligner or HunAlign and produces plaintext files containing matching sentences. Matching sentences may further be simply concatenated into a single Moses file or exported into the required output format. As a post-processing step, several filters may be applied with the purpose of favoring document pairs and aligned segments that are most useful (e.g., having good maximum/minimum length of segment, length ratio of segments, language filters, etc.) for training MT engines. The output file is then evaluated by taking random samples of segments and giving them to a human evaluator. Figure 16 summarizes this process.

2.2 LIST OF TRANSLATION CORPORA

Further in the workflow, various resources (documents and websites) were provided by Lynx partners and later processed to create our geothermal energy related corpora. Table 6 provides examples of processed documents.

ID	Category	Title	Link	Document type	Language	Tool used
1	Directive	DIRECTIVE 2009/28/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL	https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32009L0028&from=EN	PDF	DE EN ES NL	General workflow
2	Directive	Directive 2007/2/EC as regards interoperability of spatial data sets and services	https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=LEGISSUM:l28195&from=EN	HTML	DE EN ES NL	Web crawling
3	Act	Dutch Mining Act	https://zoek.officielebekendmakingen.nl/stb-2002-542.html	HTML	NL	Monolingual Web crawling
4	Website	NLOG	https://www.nlog.nl/en	HTML	EN NL	Web crawling
5	Website	NLOG	https://www.nlog.nl/geothermie	HTML	EN NL	Web crawling
6	Website	NLOG	https://www.nlog.nl/en/geothermal-energy	HTML	EN NL	Web crawling
7	Report	EBN Annual report (see Dutch translation below)	https://www.ebn.nl/wp-content/uploads/2017/06/Focus-on-Energy-2017.pdf	PDF	EN	General workflow
8	Report	EBN Jaarverslag	http://cdn.instantmagazine.com/upload/11408/web_-_2018_04_24_-_ebn_-_jaarverslag_2017.8bcde623ed6b.pdf	PDF	NL	General workflow
9	Website	DAGO - Dutch Association Geothermal Operators	https://www.dago.nu/nl/geothermie	HTML	NL	Web crawling
10	Website	DAGO - Dutch Association Geothermal Operators	https://www.dago.nu/en/geothermie	HTML	EN	Web crawling
11	Website	Thermogis - Geothermal mapping	https://www.thermogis.nl/	HTML	NL	Web crawling
12	Website	Thermogis - Geothermal mapping	https://www.thermogis.nl/en	HTML	EN	Web crawling

13	Article	Gebruik van warmte uit bodem in 5 jaar verdubbeld	https://www.cbs.nl/nl-nl/nieuws/2015/41/gebruik-van-warmte-uit-bodem-in-5-jaar-verdubbeld	HTML	NL	Web crawling
14	Article	Use geothermal heat doubled in the past 5 years	https://www.cbs.nl/en-gb/news/2015/41/use-geothermal-heat-doubled-in-the-past-5-years	HTML	EN	Web crawling
15	Website	Hoewerktaardwarmte.nl	Hoewerktaardwarmte.nl	HTML	NL	Monolingual Web crawling
16	Website	WHAT IS DEEP GEOTHERMAL ENERGY	https://vito.be/en/vito-insights/deep-geothermal/what-deep-geothermal-energy	HTML	EN	Web crawling
17	Website	WAT IS DIEPE GEOTHERMIE?	https://vito.be/nl/diepe-geothermie/wat-diepe-geothermie	HTML	NL	Web crawling
18	Website	DEEP GEOTHERMAL ENERGY IN FLANDERS	https://vito.be/en/vito-insights/deep-geothermal/deep-geothermal-energy-flanders	HTML	EN	Web crawling
19	Website	GEOTHERMIE IN VLAANDEREN	https://vito.be/nl/diepe-geothermie/geothermie-vlaanderen	HTML	NL	Web crawling
20	Website	VITO REVEALS GEOTHERMAL ENERGY POTENTIAL IN THE BORDER REGION	https://vito.be/en/vito-reveals-geothermal-energy-potential-border-region	HTML	EN	Web crawling
21	Website	GEOTHERMISCH POTENTIEEL IN GRENSREGIO	https://vito.be/nl/geothermisch-potentieel-grensregio	HTML	NL	Web crawling
22	Website	GEOTHERMAL ENERGY BOOSTS EMPLOYMENT	https://vito.be/en/geothermal-energy-boost-employment	HTML	EN	Web crawling
23	Website	GEOTHERMIE BOOST WERKGELEGENHEID	https://vito.be/nl/geothermie-boost-werkgelegenheid	HTML	NL	Web crawling
24	Website	DEEP GEOTHERMAL ENERGY IN THE KEMPEN: WHAT'S NEXT?	https://vito.be/en/news/deep-geothermal-energy-kempen-what%E2%80%99s-next	HTML	EN	Web crawling
25	Website	DIEPE GEOTHERMIE IN DE KEMPEN: WHAT'S NEXT?	https://vito.be/nl/nieuws/diepe-geothermie-de-kempen-what%E2%80%99s-next	HTML	NL	Web crawling
26	Website	DEVELOPMENT & TESTING OF INNOVATIVE	https://vito.be/en/news/development-testing-innovative-geophysical-methods-	HTML	EN	Web crawling

		GEOPHYSICAL METHODS FOR EVALUATION GEOTHERMAL POTENTIAL	evaluation-geothermal-potential			
27	Website	ONTWIKKELEN & TESTEN VAN INNOVATIEVE GEOFYSISCHE METHODES VOOR EVALUATIE GEOTHERMISCH POTENTIEEL	https://vito.be/nl/nieuws/ontwikkelen-testen-van-innovatieve-geofysische-methodes-voor-evaluatie-geothermisch	HTML	NL	Web crawling
28	Website	VITO BRINGS GEOTHERMAL HEAT TO THE SURFACE	https://vito.be/en/news/vito-brings-geothermal-heat-surface	HTML	EN	Web crawling
29	Website	VITO BRENGT AARDWARMTE AAN DE OPPERVLAKTE	https://vito.be/nl/nieuws/vito-brengt-aardwarmte-aan-de-oppervlakte	HTML	NL	Web crawling
30	Website	BALMATT ENERGY PLANT	https://vito.be/en/vito-insights/deep-geothermal/balmatt-energy-plant	HTML	EN	Web crawling
31	Website	BALMATT-SITE	https://vito.be/nl/diepe-geothermie/balmatt-site	HTML	NL	Web crawling
32	Website	GEOTHERMAL ENERGY SOURCE	https://vito.be/en/news/geothermal-energy-source	HTML	EN	Web crawling
33	Website	AARDWARMTE BRON VAN ENERGIE	https://vito.be/nl/nieuws/aardwarmte-bron-van-energie	HTML	NL	Web crawling

Table 6. Resources used to build corpora

The given resources were examined and aggregated for convenient processing by domain. In this way, resources #4, #5, and #6 were processed together, as were #7- #8, #9-#10, #11-#12, #13-#14, and #16-#33, to produce bilingual corpora. The processed corpora (LYNX_Geothermal) was uploaded to <https://www.letsmt.eu> for further use.

	MONO	DE	EN	ES	NL
DE	1,939		261	824	854
EN	10,982	261		976	9,745
ES	2,100	824	976		300
NL	13,066	854	9,745	300	

Table 7. Segment count in LYNX_Geothermal

Uploaded corpora size statistics are given in Table 7. Most segments belong to the EN-NL language pair, as most resources were Dutch geothermal-energy-related websites. Some DE-EN and EN-ES segments were extracted from resources #1 and #2.

Resources #3 and #15 were found to be unilingual; therefore, only NL content was extracted and used to build the NL monolingual corpus. The monolingual corpus contains 2,167 segments and, similarly to the bilingual, resources it was uploaded to <https://www.letsmt.eu> for further use.

2.3 LIST OF CORPORA FOR MACHINE TRANSLATION TRAINING

Table 8 and Table 9 show which public or proprietary corpora were used to train our NMT systems. We used the corpora filtering workflow of Pinnis (2018) to remove most of the lower-quality data from these corpora before training the NMT systems. The following issues are addressed by various filters:

1. Source-source or target-target entries in parallel data. (Equal source/target entries are filtered out).
2. Sentence splitting issues. (Segments with more than 1000 symbols or more than 400 tokens are filtered out; the numerical thresholds can be adjusted for each individual training task.)
3. Data corruption through optical character recognition (OCR), e.g., when processing PDF documents. (Segments containing tokens with >50 symbols are filtered out.)
4. Redundancy issues. (Duplicate entries are filtered out).
5. Partial translation (also sentence splitting) issues. (Entries where the length ratio between the source and target segments is too small (e.g., <0.3) are filtered out.)
6. Foreign language data issues. (Entries containing letters from neither source nor target languages are filtered out)
7. Sentence misalignment issues. (Sentences failing a cross-lingual alignment test using c-eval (Zariņa et al., 2015) are filtered out.)
8. Incorrect language filtering using an automatic language detection tool (Shuyo, 2010).
9. Low content overlap filtering using the cross-lingual alignment tool MPAligner (Pinnis, 2013).
10. Digit mismatch filtering. (This showed to be effective in identifying parallel corpora sentence segmentation issues).

Table 8 lists the corpora used for the EN-NL NMT systems with a total size of 41 639 229 parallel sentences.

Corpus name

LYNX_Geothermal

DGT-TM

EUBookShop

DCEP

EESC

Europarl v7

TAUS - Legal

JRC-Acquis (v.3.0)

EMA

OPUS - EMEA

TAUS - Information Technology

RAPID

Tatoeba

OPUS - ECB

TAUS - Business

Global Voices Parallel Corpus

TAUS - Electronics
European Ombudsman
EUROSTAT Combined Nomenclature
OPUS - European Constitution
TAUS - Manufacturing
EUROSTAT PRODCOM
JRC-Names
Geo Names
Europe's Languages in the Digital Age
Regions
TAUS - Misc

Table 8. Corpora used for EN-NL NMT systems

Table 9 lists the corpora used for the EN-ES NMT systems with total size of 81 176 632 parallel sentences.

Corpus name
DGT-TM
MultiUN
DCEP
Europarl v7
TAUS - Legal
JRC-Acquis (v.3.0)
European Ombudsman
OPUS - European Constitution
United Nations Parallel Corpus
TAUS - Information Technology
EUBookShop
EESC
EMA
OPUS - EMEA
RAPID
Global Voices Parallel Corpus
Tatoeba
TAUS - Telecommunications
TAUS - Electronics
OPUS - ECB
TAUS - Automotive
TAUS - Misc
TAUS - Business
TAUS - Financial
TAUS - Manufacturing

Table 9. Corpora used for EN-ES NMT systems

Table 10 lists the corpora used for the EN-DE NMT systems with total size of 87 542 066 parallel sentences.

Corpus name

OPUS - Open Subtitles
TAUS - Information Technology
DGT-TM
DCEP
EESC
Europarl v7
RAPID
EMA
JRC-Acquis (v.3.0)
Tatoeba
TAUS - Electronics
MultiUN
OPUS - ECB
TAUS - Business
Libre Office
Global Voices Parallel Corpus
German-English Parallel Corpus de-news
TAUS - Manufacturing
TAUS - Misc
OPUS - European Constitution
JRC-Names
EUROSTAT PRODCOM
ECDC-TM
Geo Names
ParaCrawl parallel corpus
Common Crawl parallel corpus
News Commentary Corpus

Table 10. Corpora used for EN-DE NMT systems

3 CONCLUSIONS AND FUTURE WORK

This report summarizes the work done in Task 2.4 “Indexing of corpora” and Task 2.5 “Translation corpora creation” under WP2 of the Lynx project. Firstly, the corpora collection approach, the analysis of data format, and the results of the terminology extraction have been explained. Then, the corpora creation workflow for Lynx acquired corpora has been described, including the workflow of parallel corpora and corpora creation from the web crawled data. Finally, the list of translation corpora and the one for machine translation used for training have been presented.

Future work will be directed to multiple areas, including collection and processing of parallel corpora, and terminology preparation for MT. As for corpora creation, the future work will be embedded within the general parallel corpora creation workflow described in section 2.1 for Lynx languages.

As for terminology preparation for MT, it must be noted that a domain-specific bilingual terminology is a resource often used to control lexical quality of domain-specific MT systems. However, not all term collections that have been prepared by terminologists, translators, or automatic means can be directly used (or should be used) in MT. This is because of ambiguity of the terms in the term collections, which, in turn, is due to insufficient morphological, syntactic, or semantic information that describes each term in the term collections. Furthermore, MT systems are source-to-target systems that in typical scenarios (i.e., if we ignore multi-way NMT systems and other multi-task neural network-based systems) translate from one source language to one target language. Therefore, term collections for MT systems have to be prepared as bilingual collections that define term pairs. However, we know that a term entry in a term database may consist of a term in multiple languages and even multiple term variants, e.g., different variants for different registers (e.g., neutral, technical, slang, etc.), types (e.g., abbreviation, initialism, short form, full form, etc.), etc. This means that before integrating terms into an MT system, each term collection has to be transformed and pre-processed into a bilingual term collection that consists of term pairs. Therefore, one focus of future work could be the preparation of term collections for use in the Lynx MT systems.

Furthermore, professional and automatically generated dictionaries are potential sources of domain-specific knowledge that could be useful for MT systems in areas where term collections are not available. Further work will focus on investigating the feasibility of cleaning bilingual dictionaries and transforming them in useful resources for MT integration.

ANNEX 1. MAIN EU LABOR LAW LEGISLATION CORPORA

- Council Directive 1999/70/EC of 28 June 1999 concerning the framework agreement on fixed-term work concluded by ETUC, UNICE, and CEEP
- Council Directive 2001/86/EC of 8 October 2001 supplementing the Statute for a European company with regard to the involvement of employees
- Council Directive 2010/18/EU of 8 March 2010 implementing the revised Framework Agreement on parental leave concluded by BUSINESSEUROPE, UEAPME, CEEP, and ETUC and repealing Directive 96/34/EC (Text with EEA relevance)
- Council Directive 89/391/EEC of 12 June 1989 on the introduction of measures to encourage improvements in the safety and health of workers at work
- Council Directive 91/533/EEC of 14 October 1991 on an employer's obligation to inform employees of the conditions applicable to the contract or employment relationship
- Council Directive 92/104/EEC of 3 December 1992 on the minimum requirements for improving the safety and health protection of workers in surface and underground mineral-extracting industries (twelfth individual Directive within the meaning of Article 16 (1) of Directive 89/391/EEC)
- Council Directive 92/85/EEC of 19 October 1992 on the introduction of measures to encourage improvements in the safety and health at work of pregnant workers and workers who have recently given birth or are breastfeeding (tenth individual Directive within the meaning of Article 16 (1) of Directive 89/391/EEC)
- Council Directive 94/33/EC of 22 June 1994 on the protection of young people at work
- Directive 2003/41/EC of the European Parliament and of the Council of 3 June 2003 on the activities and supervision of institutions for occupational retirement provision
- Directive 2003/88/EC of the European Parliament and of the Council of 4 November 2003 concerning certain aspects of the organization of working time
- Directive 2008/94/EC of the European Parliament and of the Council of 22 October 2008 on the protection of employees in the event of the insolvency of their employer (Codified version) (Text with EEA relevance)
- Directive 96/71/EC of the European Parliament and of the Council of 16 December 1996 concerning the posting of workers in the framework of the provision of services.

REFERENCES

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).

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

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

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*.