Deliverable D4.5

eLENS Miner

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No.821918

Project acronym:	enviroLENS
Project title:	Copernicus for environmental law enforcement support
Project number:	821918
Instrument:	Horizon 2020
Call identifier:	H2020-SPACE-2018
Topic	DT-SPACE-01-EO-2018-2020
Type of action	Innovation action

Start date of project:	01-12-2018
Duration:	30 months

Deliverable number	**D4.5**
Deliverable title	**eLENS Miner**
Deliverable due date	**31-10-2020**
Lead beneficiary	**JSI**
Work package	**4**
Deliverable type	**Demonstator**
Submission date: **Revision:**	**Version 1.0**

Dissemination Level		
PU	Public	x
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Title:
eLENS Miner
Author(s)/Organisation(s):
Erik Novak (JSI)
Contributor(s):
Alexander Schultmeyer (DLA Piper), Alexandra Ibragimova (IUCN), Klemen Kenda (JSI)

Short Description:
The deliverable provides an overview of the eLENS Miner System, its architecture and components. The components serve as wrappers of the methodologies developed and presented in deliverables 4.3 Semantic Toolbox and 4.4 eLENS Knowledge Extraction Components. In addition, the system was manually evaluated by the legal project partners.
Keywords:
Text Mining, enviroLENS, Legal Documents, Word Embedding, Ontology, Wikification, Enrichment, Web Service

History:

Version	Author(s)	Status	Comment	Date
0.1	Erik Novak	Draft	Initial draft with TOC	20.10.2020
0.2	Alexander Schultmeyer	Draft	Evaluation report by DLA Piper	22.10.2020
0.3	Alexandra Ibragimova	Draft	Evaluation report by IUCN	26.10.2020
0.4	Klemen Kenda	Draft	Added section 4	09.11.2020
0.5	Erik Novak	Draft	Added rest of the content	09.11.2020

Review:

Version	Reviewer	Comment	Date
1.0	Mario Dohr	Final review, Ready for Submission	13.11.2020

Table of Contents

1 Introduction 7
2 eLENS Miner System 8
2.1 Pre-processing and Annotation Component 9
2.2 Search Engine and Query Expansion 9
2.3 Document Comparison Component 10
2.4 Text Embedding Component 10
3 Evaluation 11
3.1 Evaluation Analysis 11
4 Global Digital Twins Initiative 13
5 Conclusion 15
Appendix A: API Documentation 16
Retrieves the requested documents 16
Retrieves the requested document 16
Retrieve the most similar documents 17
Search for relevant documents 18
Create the text vector representation 20

List of Figures

Figure 1. The eLENS Miner Architecture. It connects different components and enables the documents to be enriched and searched through. 8

Figure 2. The annotation pipeline. It is able to annotate the legal documents with syntactic and semantic annotations, Wikipedia concepts and environmental and geospatial metadata. 9

Figure 3: Earth Observation feed created using Event Registry data. 13

List of Tables

No table of figures entries found.

Abbreviations

API	Application Programming Interface
EFTA	European free trade association
EU	European Union
FAO	Food and Agriculture Organization of the United Nations
HTML	Hypertext Markup Language
HTTPS	HyperText Transfer Protocol Secure
IUCN	International Union for Conservation of Nature
JSON	JavaScript Object Notation
NUTS	Nomenclature of territorial units for Statistics
UNEP	United Nations Environment Programme
URL	Uniform Resource Locator

1 Introduction

This document is the fifth and final document of the enviroLENS WP4 and is dedicated to providing a short overview of the eLENS Miner System components, its integration into the eLENS Portal and initial evaluation by our partners DLA Piper and IUCN.

The purpose of this deliverable is to provide a general overview of what has been developed within WP4, i.e. the final components of the eLENS Miner System, as well as its initial evaluation within the eLENS Portal. The evaluation has been performed by our internal partners who have given a positive feedback – but some improvements are still required.

The document is structured as follows. Section 1 describes the eLENS Miner System and its components. Since the components were presented in detail in previous WP4 deliverables, we provide only a brief description of their role in the whole system. Section 3 presents the evaluation of the systems performance done by our project partners, namely DLA Piper and IUCN. Afterwards, we present the Global Digital Twin Initiative in section 4. Here, we intend to use and extend the components of the eLENS Miner System on the news domain. We conclude the deliverable in section 5 and provide an Appendix containing the eLENS Miner System API documentation.

2 eLENS Miner System

In this section we describe the eLENS Miner System architecture, its components and how they are connected in system. The eLENS Miner System is the main result of WP4 and is intended to enrich legal documents with semantic and geo-spatial information, as well as enable searching through the enriched documents using both text and location metadata in the search query. The first prototype of the eLENS Miner System was presented in deliverable 4.4 – eLENS Knowledge Extraction Components. Since then, the system was stabilized, made it available through a secure protocol (HTTPS) and is running on https://envirolens.ijs.si/. Figure 1 shows the general schema of the architecture.

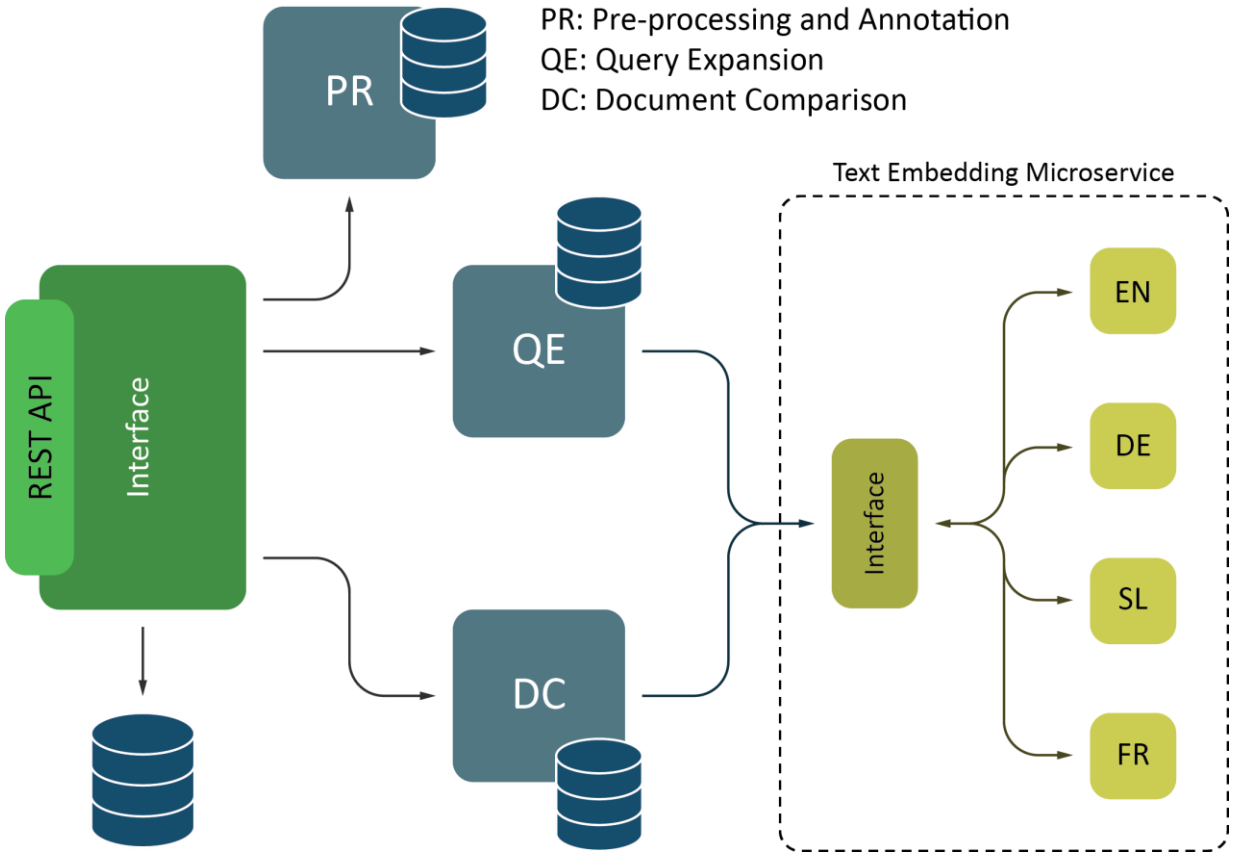

Figure 1. The eLENS Miner Architecture. It connects different components and enables the documents to be enriched and searched through.

The architecture connects three components: 1) the Pre-processing and Annotation component, 2) the query expansion and search engine, and 3) the document comparison component. The architecture was designed and developed to be general enough to be also used on documents outside the legal domain. The code to the eLENS Miner System is openly available on Github[1]. Although the components were described in

[1] https://github.com/JozefStefanInstitute/eLENS-miner-system

the previous deliverables, we include their brief descriptions and functionalities in the following sections to have a complete report. The API documentation of the system is available as an Appendix to this report.

2.1 Pre-processing and Annotation Component

The pre-processing and annotation component contain the pipeline that enriches a document with syntactic and semantic annotations, Wikipedia concepts[2], as well as environmental and geographical terms found in the InforMEA, NUTS[3], and Protected Planet[4] ontologies, respectively. Figure 2 shows the general architecture of the pipeline.

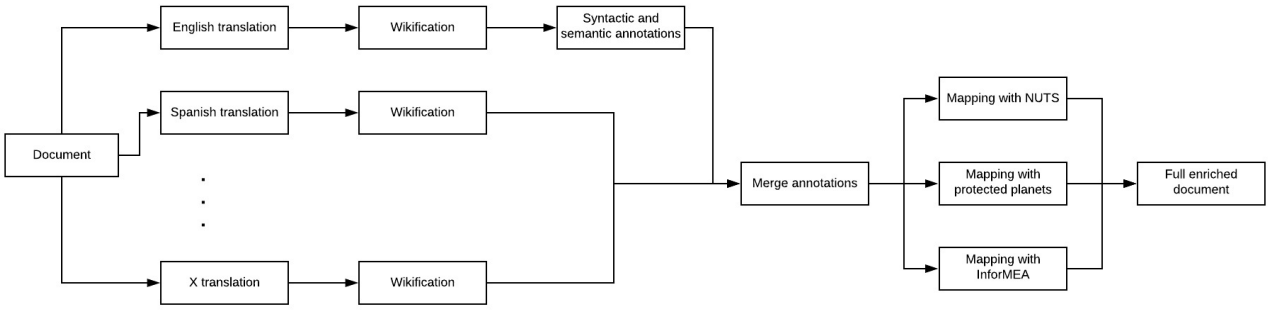

Figure 2. The annotation pipeline. It is able to annotate the legal documents with syntactic and semantic annotations, Wikipedia concepts and environmental and geospatial metadata.

The annotation component is used when new documents are added to the database. In addition, the component can be set as a stand-alone microservice, making it available to also annotate documents outside the legal domain. An extensive description of the annotation methodologies and their analysis can be found in deliverable 4.3 eLENS Semantic Toolbox and deliverable 4.4 eLENS Knowledge Extraction Components.

2.2 Search Engine and Query Expansion

The search engine component enables the user to search through the annotated legal documents. The user can provide different query parameters such as a free-flown text of the topic of interest, the document source dataset (e.g. EURLEX or ECOLEX), the list of locations, the specific document languages and the list of environmental terms available in informea. Based on these parameters, the search engine finds the relevant documents and returns their metadata. The component integrates Elasticsearch[5], a search and analytics engine, for finding relevant legal documents.

In addition, the search engine performs query expansion, e.g. extending the query text with additional terms that improve the search results. This is done by using word embeddings, e.g. word representations that

[2] http://wikifier.org/
[3] https://ec.europa.eu/eurostat/web/nuts/background
[4] https://www.protectedplanet.net/en
[5] https://www.elastic.co/elasticsearch/

capture their semantic information. Using word embeddings, we are able to expand the query by finding semantically similar terms to those in the text query. These are then added to the original query before retrieving the relevant legal documents.

The search engine is accessible through the eLENS Miner System API. The description of the query expansion and its analysis can be found in deliverable 4.4 eLENS Knowledge Extraction Component.

2.3 Document Comparison Component

The document comparison component is designed to find similar semantically similar documents. Here, the similarity is measured using only the documents' content, e.g. the document's text, descriptors and title. Doing so we are able find documents that have similar content but are not restricted to their geospatial location.

To measure the similarity of the documents we first use word embeddings to create the documents vector representation. Afterwards, we calculate the Cosine distance between documents and use it as a similarity measure. This measure is then stored in the database and used to rank the most similar documents to the user selected document. This functionality is accessible through the eLENS Miner System API.

The methodologies and their evaluations are described in detail in deliverable 4.3 eLENS Semantic Toolbox and deliverable 4.4 eLENS Knowledge Extraction Component.

2.4 Text Embedding Component

The text embedding component was developed to support the search engine and document comparison component. It is designed to provide a vector representation for the document comparison methodologies and to find similar terms for the query expansion part of the search engine. It is designed to support multiple languages, enables automatic language detection of the input text and return the embedding in a format that can be read by any programming language. The methodology used in this component is described in deliverable 4.3 eLENS Semantic Toolbox.

This component can be also run as a microservice, enabling it to be integrated into other text processing and analysis architectures. Its code and documentation are available on Github[6].

[6] https://github.com/ErikNovak/python-text-embedding-microservice

3 Evaluation

The eLENS Miner System is integrated into the eLENS Portal – more precisely in the Legislation Discovery Tool. There, the user can select the environmental topic of interest and the geospatial Area of Interest (AOI), which are then sent to the eLENS Miner System. The system processes the user query and returns the legal documents which the system identified as relevant. While the integration process is out of scope for this deliverable, it will be presented in WP6 deliverable 6.6 eLENS Services.

We asked our legal project partners, namely DLA Piper and IUCN, to evaluate the relevance of the query results returned by the eLENS Miner System. This section provides the aggregated list of feedbacks provided by both partners.

Aggregated Feedback

- The search results are relevant for the EU (most of the times) and are sensitive to the user selected Area of Interest (AOI) and the project use-case environmental topics, e.g. deforestation, pipeline, and illegal infrastructure.
- Using the advanced search, which allows the user to create a more specific request, provides more relevant documents.
- The search results do not show any directives and regulations. In addition, the results do not show any local legislation. Including both the country-level regulations and legislations can be very relevant for the user.
- If a user selects an Area of Interest outside the EU, the eLENS Miner System returns documents that are not relevant for that area (returning documents associated with the EU).

3.1 Evaluation Analysis

The initial feedback provided by the partners provide good insight into the performance of the eLENS Miner System. The system is able to find relevant documents for the selected AOI and the environmental topic, as long as the AOI is within the EU borders. When the AOI is outside the EU it still finds documents that are relevant to the selected environmental topic, but are not applicable to the selected area. This problem can be solved by configuring the search engine to have a stricter policy when calculating the relevance of a document to a selected AOI. This is something we intend to do in the future.

Most of the feedbacks are related to the availability of documents, e.g. country-level regulations, legislations, etc. This is associated with the document sources that are currently indexed in the eLENS Miner System. The documents are gathered from EUR-Lex[7] and ECOLEX[8] datasets. While ECOLEX contains legal documents that are applicable across the globe, the bigger EUR-Lex dataset is focused only on the EU law. In addition, the ECOLEX dataset can be only used for research purposes and hence cannot be included in the

[7] https://eur-lex.europa.eu/homepage.html
[8] https://www.ecolex.org/

final product. Because the eLENS Miner System currently does not include country-based legal documents it cannot provide them as the search result.

During the eLENS Miner System development we were also looking for alternative legal datasets than can be included into the final product. Unfortunately, we were not able to identify such datasets and will continue to look for them in the future.

4 Global Digital Twins Initiative

In recent years, many initiatives that introduced different digital twins in various domains have been started and many results are already visible in everyday life. For example, in Industry 4.0 digital twins have taken an integral role within several H2020 projects and help to solve various cognition, modelling and optimization problems.

Earth Observation domain is an ideal domain for creation of global digital twins, reflecting the environmental situation of the world. Moreover, digital twins do not only contain data about the current state of a particular system, but also introduce operators. The operators are able to simulate and predict the situation within the digital twin in the near/far future.

EnviroLENS project has explored the data fusion from "two separate worlds". The world of Earth Observation data and the world of text (legislation). World news are an essential part that is needed for building a global digital twin. EventRegistry[9] is a JSI related spin-off based our own technology that captures online global news in real time. An example dashboard for EO related news is depicted in Figure 3. Moreover, the EventRegistry can aggregate the news into events and is able to produce additional metadata on the topic (from timeline, etc.).

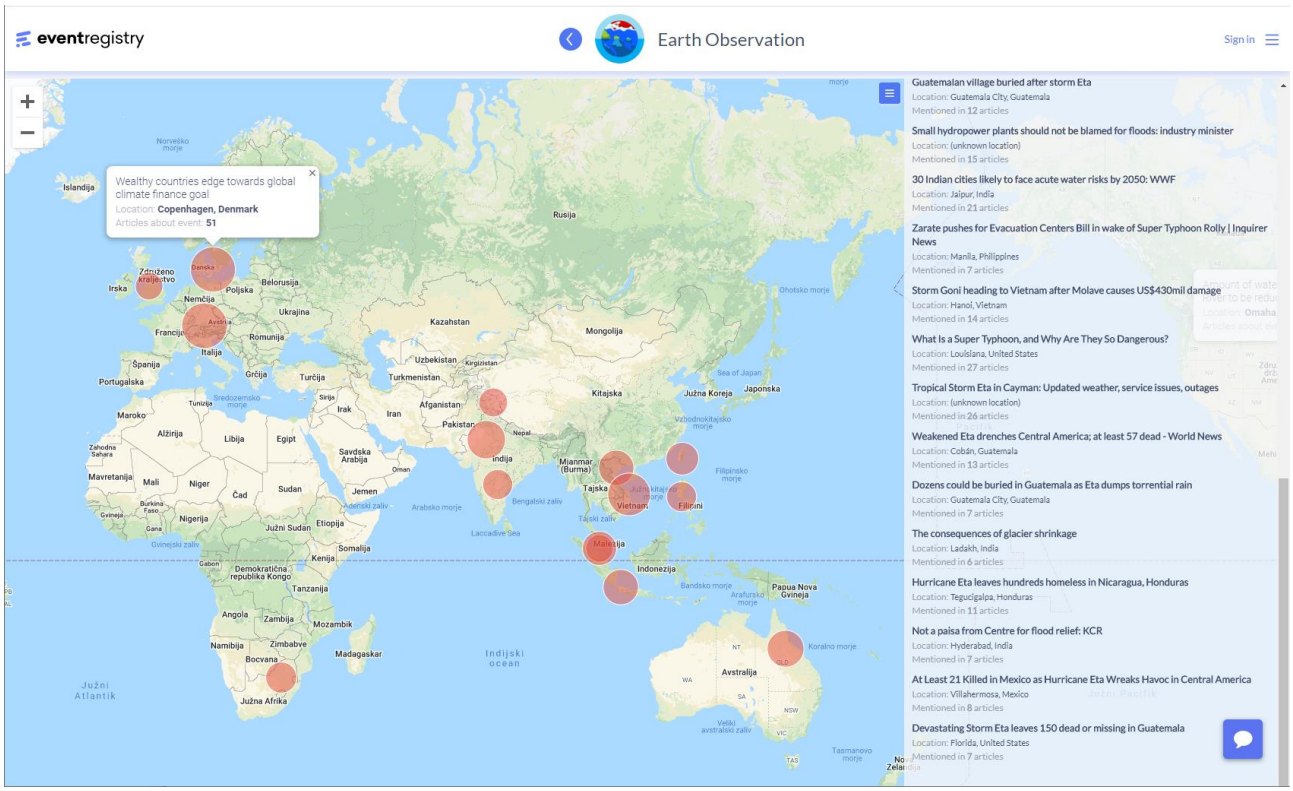

Figure 3: Earth Observation feed created using Event Registry data.

[9] https://eventregistry.org/

EventRegistry functionalities can be upgraded with advanced text-processing technologies developed within EnviroLENS (enrichment, search and similarity). And a great additional value would be provided by connecting this source to the actual EO capabilities (as demonstrated in EnviroLENS).

JSI has therefore started an initiative (and has acquired funding) with ESA including Sinergise and EventRegistry to provide an "Earth Observation Digital Twin of the News". What if we could produce dossiers automatically to link current news items to Earth Observation imagery which complements social and traditional media?

The goals of the project are quite similar to initial goals within the EnviroLENS project:

1. Encourage fact-based reporting and decision making while raising awareness for EO and Copernicus.
2. Provide journalists, analysts, and interested citizens with quality background material.
3. Generate automatically supporting material (i.e., customised satellite data) as trends are identified.
4. Attractive visual and interactive user experience.

The two main envisioned implementation principles include also the workflows envisioned in EnviroLENS:

1. Analyse current news and social media streams for topics such as fires, earth quakes, floods, air pollution etc.
2. Find EO products related to the events, identify relevant parts and highlight them (using current AI methods).

Example Workflow:

If forest fires in Siberia trend in the media, AI scans the news for more exact locations, scans S1/S2/S3/S5P for signature information, and generates a dossier of imagery which links back to the original news items and tweets. Once finished the dossier can be pushed to users via social media.

The methodology, developed within EnviroLENS has been directly transferred to another environment (in this case, the main target audience are the journalist and not lawyers).

5 Conclusion

In this document we present the eLENS Miner System architecture and provide a brief description of its components. The architecture was designed and developed to 1) annotate the legal documents with syntactic and semantic annotations, Wikipedia concepts, environmental and geospatial metadata, 2) enable searching through the documents by providing geospatial and environmental information as well as free-flown text, and 3) to identify similar documents. Since the components were described and analysed in previous deliverables, we provide only a brief overview of their functionalities.

The eLENS Miner System is integrated into the eLENS Portal. It was evaluated by our legal project partners and provided feedback on its performance: while the system performs well in locations for which we have the relevant documents, going outside that area provides irrelevant results. This is accredited to the datasets available within the project, e.g. ECOLEX (can be used only for research purposes) and EUR-Lex (limited to the European Union), and to the loose rules of the search engines when calculating the relevancy of the document. We intend to improve the search engine to provide more relevant results and continue searching for legal documents that can be included to the eLENS Miner System.

Additionally, we identified and intend to collaborate within the Global Digital Twin Initiative. While the initiative focuses on news documents, the eLENS Miner System components were developed general enough to be also used in domains outside of law.

Appendix A: API Documentation

This appendix contains the API documentation of the eLENS Miner System.

Method	Route	Description
GET	/document	Retrieves the requested documents
GET	/document/{document_id}	Retrieves the requested document
GET	/document/{document_id}/similar	Retrieves the most similar documents
GET	/document/search	Search for relevant documents
GET	/embeddings/create	Create the text vector representation

Retrieves the requested documents

GET /documents

Query parameters				
Attribute	**Type**	**Title**	**Optional**	**Description**
document_ids	Array of Number	List of document IDs	False	IDs of the documents we want to retrieve
Response body attributes				
Attribute	**Type**	**Title**		**Description**
documents	Array of objects	Documents		The array of objects containing the document metadata

Document properties			
Attribute	**Type**	**Title**	**Description**
document_id	Number	Document ID	The document ID
name	String	Document Name	The document name
document_source	String	Document Source	The document source
category	String	Category	The documents legal category
date	String	Date	The documents date
fulltextlink	String	Full Text Link	The link to the documents full text
sourcelink	String	Source Link	The link of the document on the source site
sourcename	String	Source Name	The name of the documents source

Retrieves the requested document

GET /documents/{document_id}

Route parameters				
Attribute	**Type**	**Title**	**Optional**	**Description**
document_id	Number	Document ID	False	The ID of the document to be retrieved

Response body attributes			
Attribute	**Type**	**Title**	**Description**
documents	Array of objects	Documents	The array of objects containing the document metadata. **Contains only one object.**

Document properties			
Attribute	**Type**	**Title**	**Description**
document_id	Number	Document ID	The document ID
name	String	Document Name	The document name
document_source	String	Document Source	The document source
category	String	Category	The documents legal category
date	String	Date	The documents date
fulltextlink	String	Full Text Link	The link to the documents full text
sourcelink	String	Source Link	The link of the document on the source site
sourcename	String	Source Name	The name of the documents source

Retrieve the most similar documents

GET /documents/{document_id}/similar

Route parameters				
Attribute	**Type**	**Title**	**Optional**	**Description**
document_id	Number	Document ID	False	The ID of the document to retrieve its similar documents

Response body attributes			
Attribute	**Type**	**Title**	**Description**
documents	Array of objects	Documents	The array of objects containing the document metadata

Document properties			
Attribute	**Type**	**Title**	**Description**
document_id	Number	Document ID	The document ID
name	String	Document Name	The document name
document_source	String	Document Source	The document source
category	String	Category	The documents legal category
date	String	Date	The documents date
fulltextlink	String	Full Text Link	The link to the documents full text
sourcelink	String	Source Link	The link of the document on the source site
sourcename	String	Source Name	The name of the documents source
similarity	Number	Similarity	The documents' similarity score

Search for relevant documents

GET /documents/search

Query parameters				
Attribute	**Type**	**Title**	**Optional**	**Description**
text	String	Query Text	False	The text for which we wish to retrieve the documents
source	String	Document Source	True	The comma separated source names of the documents. When provided, returns only documents that are associated with any of the provided sources. Possible options: eurlex, ecolex. **Default value** is eurlex.
locations	String	Associated Locations	True	The comma separated location names. When present, returns only the documents that are associated with any of the locations. **Default value** is Null.
languages	String	Languages	True	The comma separated languages in English. When present, returns only the documents that are available in any of the provided languages. **Example:** languages=english,german
informea	String	InforMEA	True	The comma separated InforMEA terms. When present, returns only the documents that are associated in any of the provided InforMEA terms.
limit	Number	Limit	True	The number of retrieved documents.
page	Number	Page	True	The page of the retrieved documents.

Response body attributes			
Attribute	**Type**	**Title**	**Description**
documents	Array of objects	Documents	The array of objects containing the document metadata

Document properties			
Attribute	**Type**	**Title**	**Description**
score	Number	Score	The relevance score of the document to the query.
document_id	Number	Document ID	The document ID
title	String	Document Title	The document title (if present, otherwise None).
abstract	String	Document Abstract	The document abstract (if present, otherwise None).
link	String	Document Link	The link to the associated document or webpage containing more information.
date	Date	Document Date	The date of the document (if present, otherwise None).
celex	String	Document CELEX ID	The document CELEX ID (if present, otherwise None).
keywords	Array of Strings	Document Keywords	The document keywords array (if present, otherwise None).
source	String	Document Source	The documents source.
informea	Array of Strings	InforMEA Terms	The document InforMEA terms array (if present, otherwise None).
languages	Array of Strings	Document Languages	The array of languages the document is available in (if present, otherwise None).
subjects	Array of Strings	Document Subjects	The array of subjects associated with the document (if present, otherwise None).
areas	Array of Strings	Document Areas	The areas the document is associated with (if present, otherwise None).

query	Object	Query Parameters	The query parameters used to perform the search.
metadata	Object	Search Metadata	The search metadata used to support navigating through the results.

Metadata item properties			
Attribute	**Type**	**Title**	**Description**
total_hits	Number	Total Hits	The number of all relevant documents.
total_pages	Number	Total Pages	The maximum number of result pages.
prev_page	String	Previous Page	The URL to the previous search result page (if present, otherwise None).
next_page	String	Next Page	The URL to the next search result page (if present, otherwise None).

Create the text vector representation

GET /embeddings/create

Query parameters				
Attribute	**Type**	**Title**	**Optional**	**Description**
text	String	Text	False	Text for which we want the embedding.
language	String	Language	True	Language of the given text.
Response body attributes				
Attribute	**Type**	**Title**	**Description**	
embedding	Array of Numbers	Embedding	The array of numbers representing the text embedding.	
language_model	String	Language Model	The language of the embedding model used to create the embedding.	
text	String	Input Text	The input text	
tokens	Array of Objects	Input Tokens	The tokens used to generate the text embeddings.	