POLIRURAL

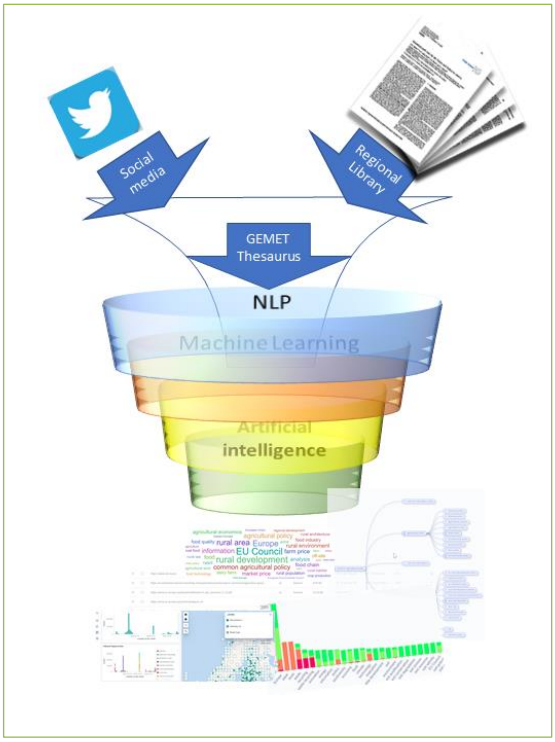

TEXT MINING
Semex.io
Semantic Explorer User´s Guide
V.1.: 02.03.2021

Contents

1. Introduction

Semantic Explorer is one of the outputs of the Polirural research project financed by the Horizon2020 Program of the EU. One of the objectives of the Polirural project is to bring solutions to policymakers to support European rural areas in responding to contemporary challenges. The tool has been developed to provide support to researchers and facilitators by reducing the cognitive load related to tasks that are essential to policy processes, such as Foresight and Policy evaluation.

It is an open-source web app tool based on cutting edge text mining (TM) technology, capable of extracting information from unstructured data and displaying the results through graphical and textual outputs. The tool is based on NLP functions including Topic Extraction, Named Entity Recognition and Sentiment Analysis. Through these NLP processes users can access analyzed text and visualize the result through clear graphical representations.

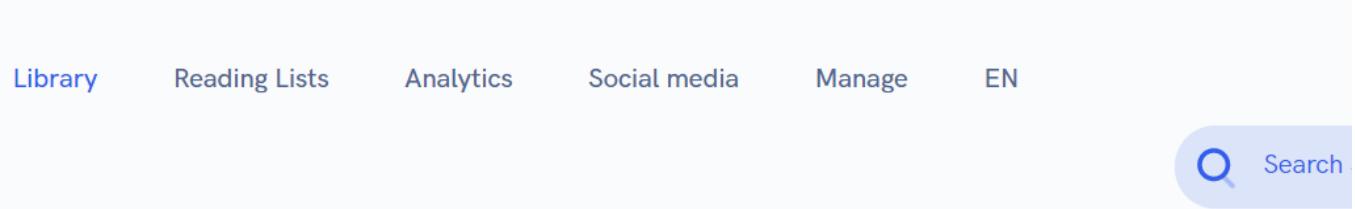


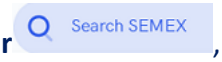
The portal is the gateway to a large library of more than 4000 documents related to European rural areas. Sources have been gathered by Polirural Pilots ensuring a high level of expertise in rural related topics and wide geographical representation (12 Pilots in different parts of Europe plus Israel). The scientific articles, technical reports and policy related documents collected relate to the needs of Polirural areas as well as to the local, regional, national and European policies.

The objective of this guide it to give practical insights about the use of the tool to make it accessible to a wider range of users. After a brief introductory part about the platform`s contents and the Text Mining technology applied, the guide provides the reader with step-by-step instructions for the use of the tool.

1.1.Content Overview

This section introduces the contents as well as the structure of the website and of this guide itself.



The main **Library** is built of sources related to **needs analysis** and **policy evaluation** for each Pilot area. Sources include online news articles, discussion forums, academic papers and blogs. Here, for each source, the system provides a summary and extracts the most frequent Topics, Keywords and Named Entities. The **Search bar** 

, accessible from almost every page of the website, gives direct access to the entire library through free text search. With this application it is possible to find specific articles as well as sources that are linked to a certain subject, similar to Google.
Jump to Chapter 3 to see more details about the **Library.**

A second section of the repository is the **Curated Reading List**, containing collections of sources created by Polirural Pilots and partners. The results are similar to the ones from the previous sections (Topics, Keywords, Named Entities), but are aggregated from the multiple sources. If you want to read more details and discover how to make a Curated Reading List go to Chapter 4 of this guide.

In **Analytics** it is possible to visualise various results from the system such as **Topic Explorer** and **Polarity Scores** and to access **Kibana** for further analytical functions (dedicated to data analysts). Go to Chapter 5 and Chapter 6 to access respectively explanations about Polarity Score and Topic explorer.

The **Social Media** section displays the results of continuous streaming of messages from Twitter with information related to Polirural research topics. Messages are streamed based on specific Twitter users that are particularly active in Polirural research field, as well as on the basis of specific hashtags such as #CAPreform, #EUFarm2Fork, #EAFRD, #Greendeal etc… More details about the Social Media section are presented in Chapter 7.

Registered users from the Polirural consortium are authorised to add, remove and update sources through the **Manage** section. To learn more about this functionality go to Chapter 8 at the end of this guide.

2. Text Mining technological basics in semex.io

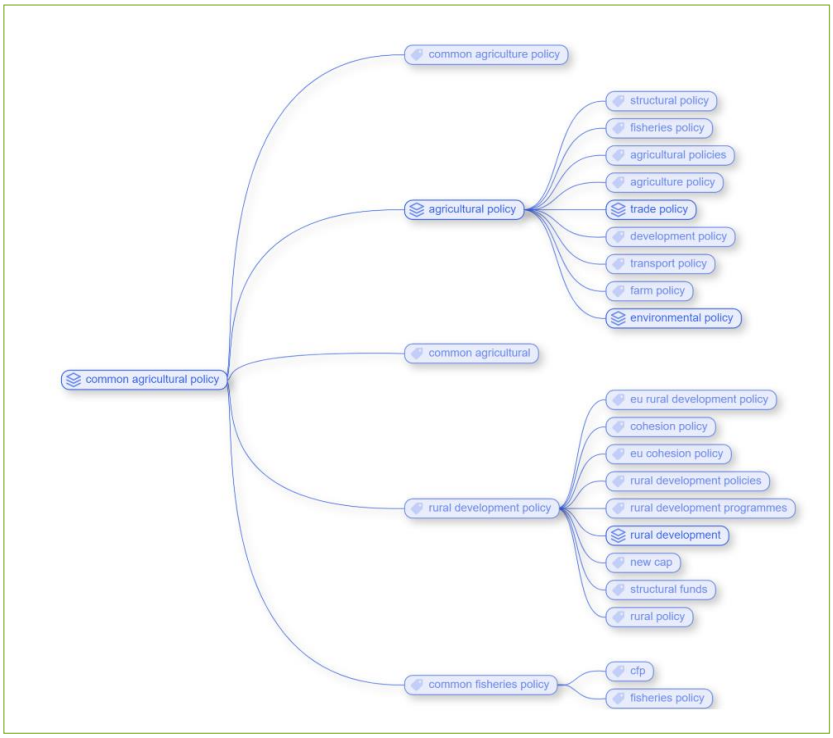
This chapter introduces the technological features of semex.io. If the reader is looking for a ´how to work with semex´ guide only jump to the next chapter.

Semex.io is the frontend application of a complex backend system based on text mining that is available through open API. Before getting into practical use it is important to understand the most basic features of the backend to better grasp the meaning and value of certain terms such as Topics, Keywords, Named Entities etc.

Natural Language Processing (NLP) breaks down language into shorter, more basic pieces, called tokens (words, periods, etc.), and attempts to understand the relationships between the tokens. In semex.io NLP is used for Keyword and Topic extraction and modelling, Sentiment analysis and Named Entity Recognition (NER). All these tasks require more basic NLP tasks such as Word and Sentence tokenization, Dependency parsing, Part of Speech tagging, lemmatization, semantic comparisons of words.

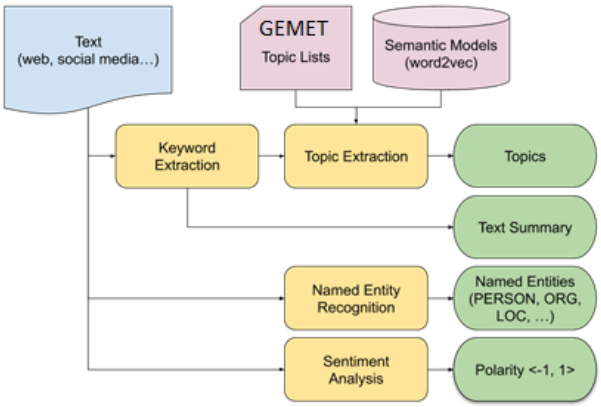
GEMET (GEneral Multilingual Environmental Thesaurus):

Semex.io uses a predefined thesaurus called GEMET. This thesaurus was developed by the European Environment Agency (EEA) and the European Topic Centre on Catalogue of Data Sources (ETC/CDS) as an indexing, retrieval, and control tool. It contains more than 5000 topics in 37 languages. Topics have hierarchical relationships (broader and narrower terms) and "related term" relationships. In Semex, we choose the topics from GEMET because it was conceived as a "general" thesaurus, aimed to define a common general language, a core of general terminology for the environmental sector, which is closely related to our main topic of interest, rural areas. The image here below, extracted from the ´Topic Explorer´ functionality, represents topics and subtopics from GEMET related to ´common agricultural policy´.



1.3 System workflow and core functionalities

The core of the NLP system consists of three main processes - Topic extraction, Named Entity Recognition and sentiment analysis (Polarity) as shown in the Figure below.



1.3.1 Topic extraction and modelling

This is the most complex NLP task in the system and is based on multi-label classification. The system uses a restricted list of topics which can be assigned to the text. The process can be broken down into smaller tasks which are described below.

1.3.2 Keyword extraction

The most important words or noun phrases are identified in the text. Graph-based TextRank algorithm (based on Google's PageRank algorithm) was selected as the best method to extract a list of keywords from the text. These keywords are also used to produce the summary of each source and of Curated Reading Lists.

1.3.2 Topic similarity comparison

Extracted keywords are compared to a set list of topics in GEMET and the most similar topics are selected. The comparison is done by Word Mover's Distance1 (WMD) algorithm which is one of the most accurate algorithms available to semantically compare documents.

1.3.3 Named Entity Recognition (NER)

It is the identification of real-world entities such as persons, organisations, locations and others (we can identify 19 different categories of entities). Extracted entities can link the text to specific geographical location, to an organisation or concrete persons.

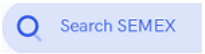
1.3.4 Sentiment analysis (opinion, emotion)

It is a technique used to identify or classify the polarity of text. The opinion polarity can range from negative (-1), through neutral (0) to positive (1). The methods used for sentiment analysis can be based on a dictionary of polarity words (terms such as 'waste' are considered negative) or training a machine learning or deep learning model.

The Semantic Explorer can process texts in the 10 languages. It was not possible to cover Hebrew and Macedonian because of no availability of trained models and lack of linguists in the project.

3. Welcome Page

The home page provides a brief description of semex.io and how it fits in Polirural research project. It also provides several statistics about Library´s content such as languages, number of sources, number of Tweets etc.

The top Menu gives direct access to all the sections of the website, to Sign in and set languages preferences. It is also possible to directly look for documents available in the Library through the Search toolbar 

on the top right.

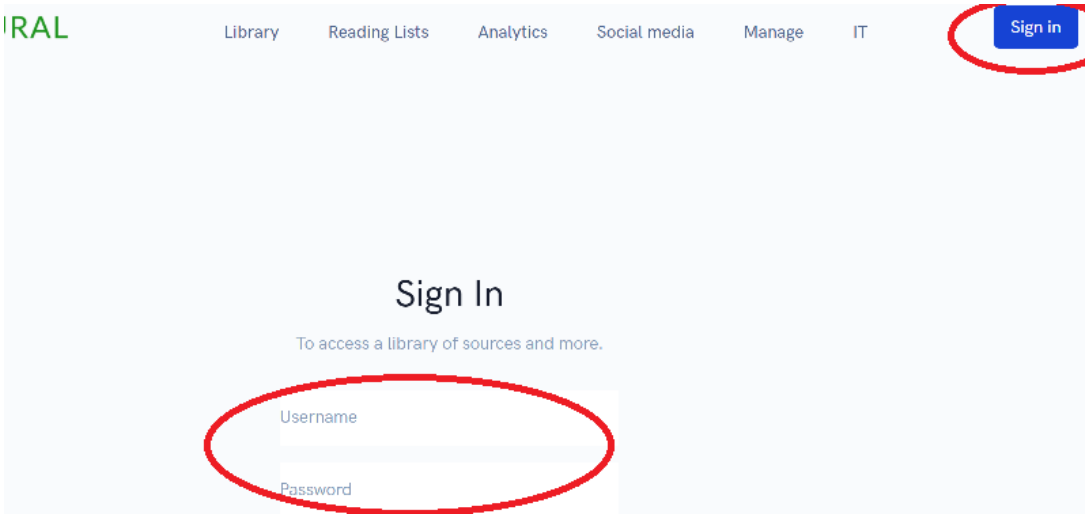
3.1 Basic Settings

This section explains how to login, how to set a language and what are the statistical information present in the home page.

2.1.1 Sign-in

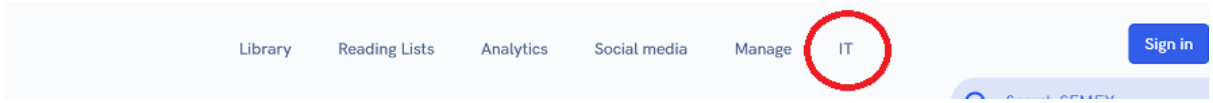
Access to the website is allowed to public (unregistered) and registered users. Registration is necessary for adding new sources in the Library, for creating new Curated Reading Lists and for accessing Kibana. Sign in credentials is restricted to Polirural partners. If you are interested in getting authorised access please write an email to KAJO team with your request.

Once you have your semex.io credentials you can click on the Sign In button on the top left menu and insert them in the relative fields:



2.1.2 Language

The semex.io website is available in English only. However, you can select a language preference on the top right menu which will pre-set filters for the text to be analysed during your navigation in semex.io.



2.1.3 Statistics

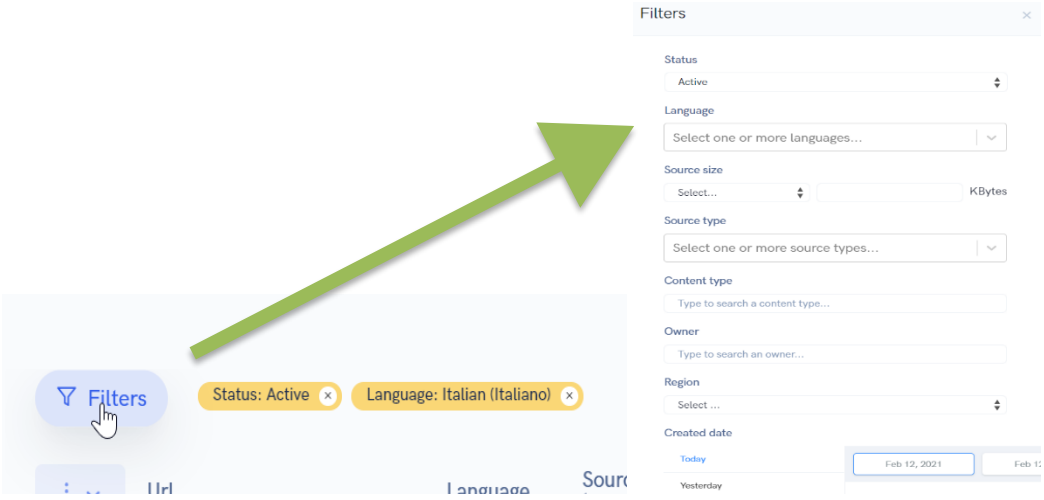
The homepage presents various statistics regarding semex.io, such as number of organizations involved, number of regions concerned, number of sources in the Library, languages, number of Tweets analysed etc… These statistics are dynamic and updated daily.

4. Library

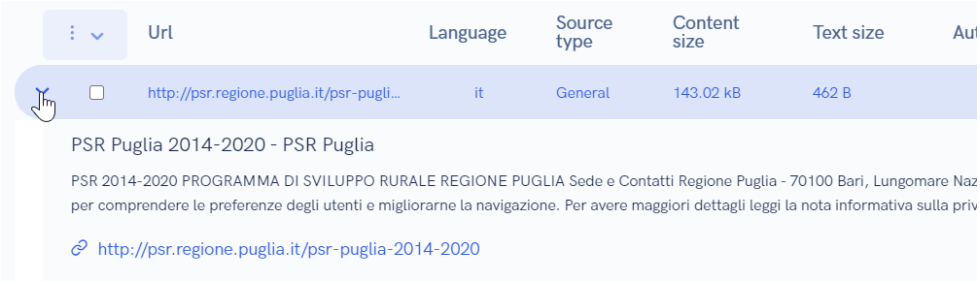
The Library section is the entry point to visualise the entire list of sources collected by Polirural Pilots related to needs analysis and policy evaluation. The Library also contains contents from the blog www.CAPreform.eu used as a test for the policy evaluation exercise.

To find a particular source, user can use the filters by clicking on the Filters button on the left sliding panel.

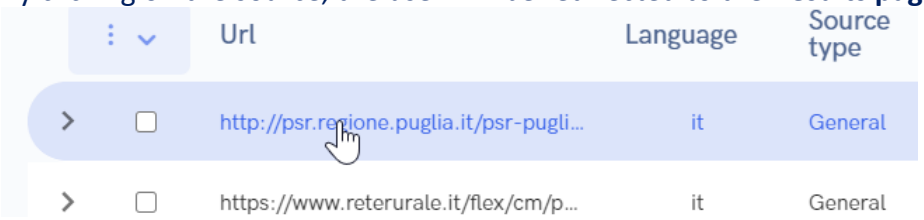
User can filter the library by: Status, Language, Source size, Source type (most of the sources will be found in General type), Content type, Owner, Region, Created and Updated date etc…



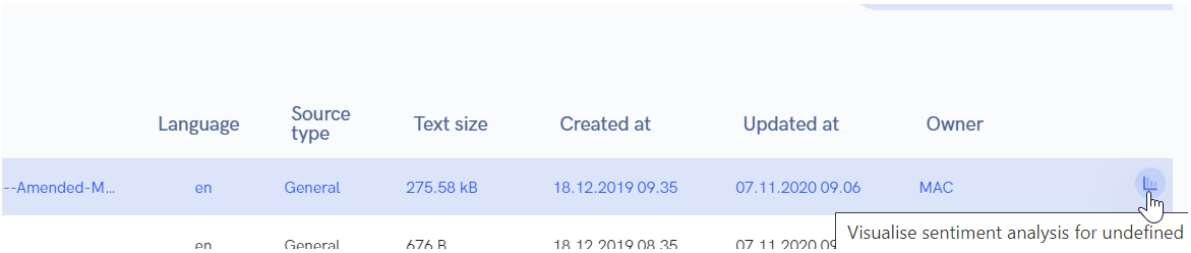
For each source, the system provides the possibility of opening a sliding panel by clicking on the left menu as in the image below. The sliding panel shows the **title, a short summary and the link to the original source**.



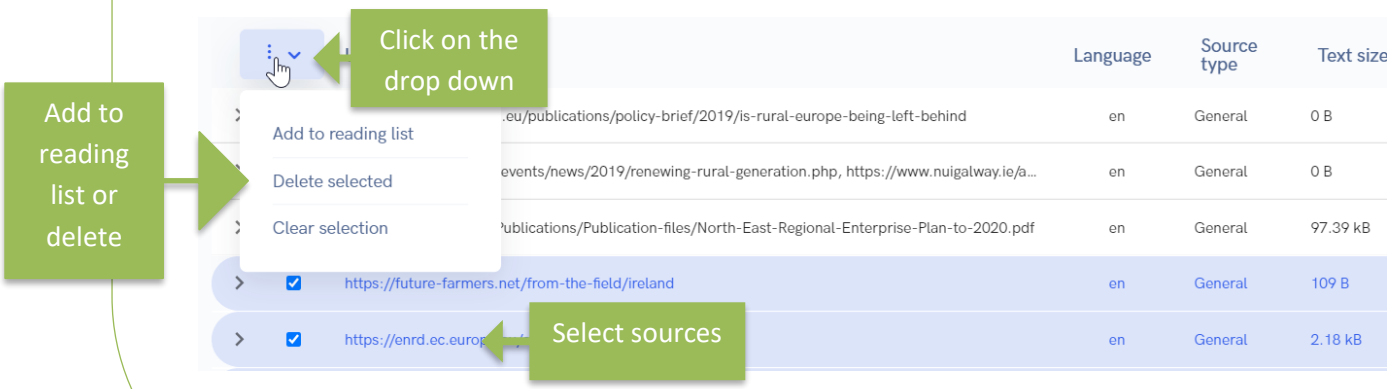
By clicking on the source, the user will be redirected to the **Results page**.



By **clicking on** the icon on the right of the source line, user is redirected to the **sentiment analysis histogram** view (see section 3.3 of this guide).



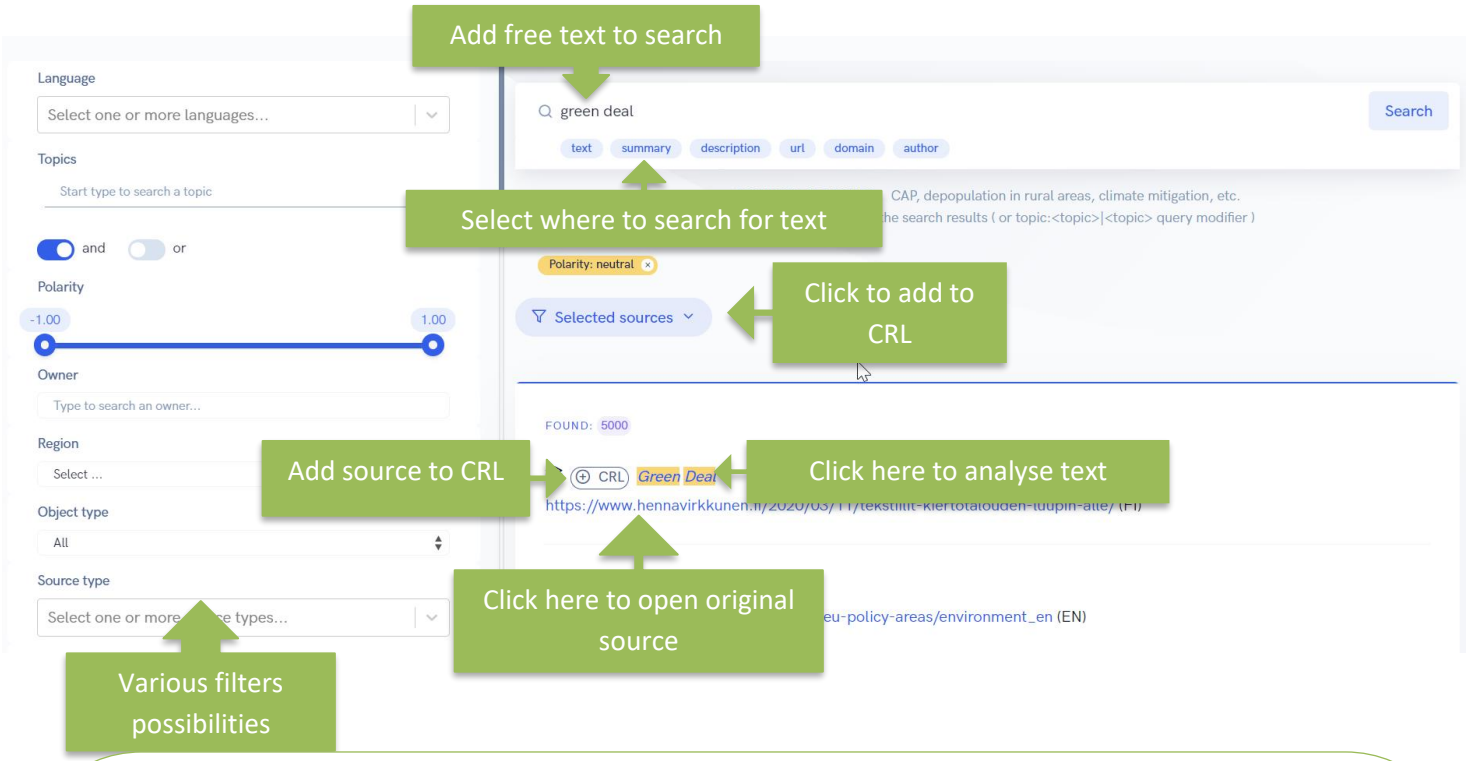
From the Library section it is also possible to select some of the sources and add them to a Curated Reading list or to Delete them.



4.1 Search Bar

The **Search bar** Search SEMEX, accessible from almost every page of the website, gives direct access to the entire library through free text search. With this application it is possible to find specific articles as well as sources that are linked to a certain subject, similar to any library search tool.

Once the user has inserted the free text in the search bar the system redirects to search page as illustrated below.

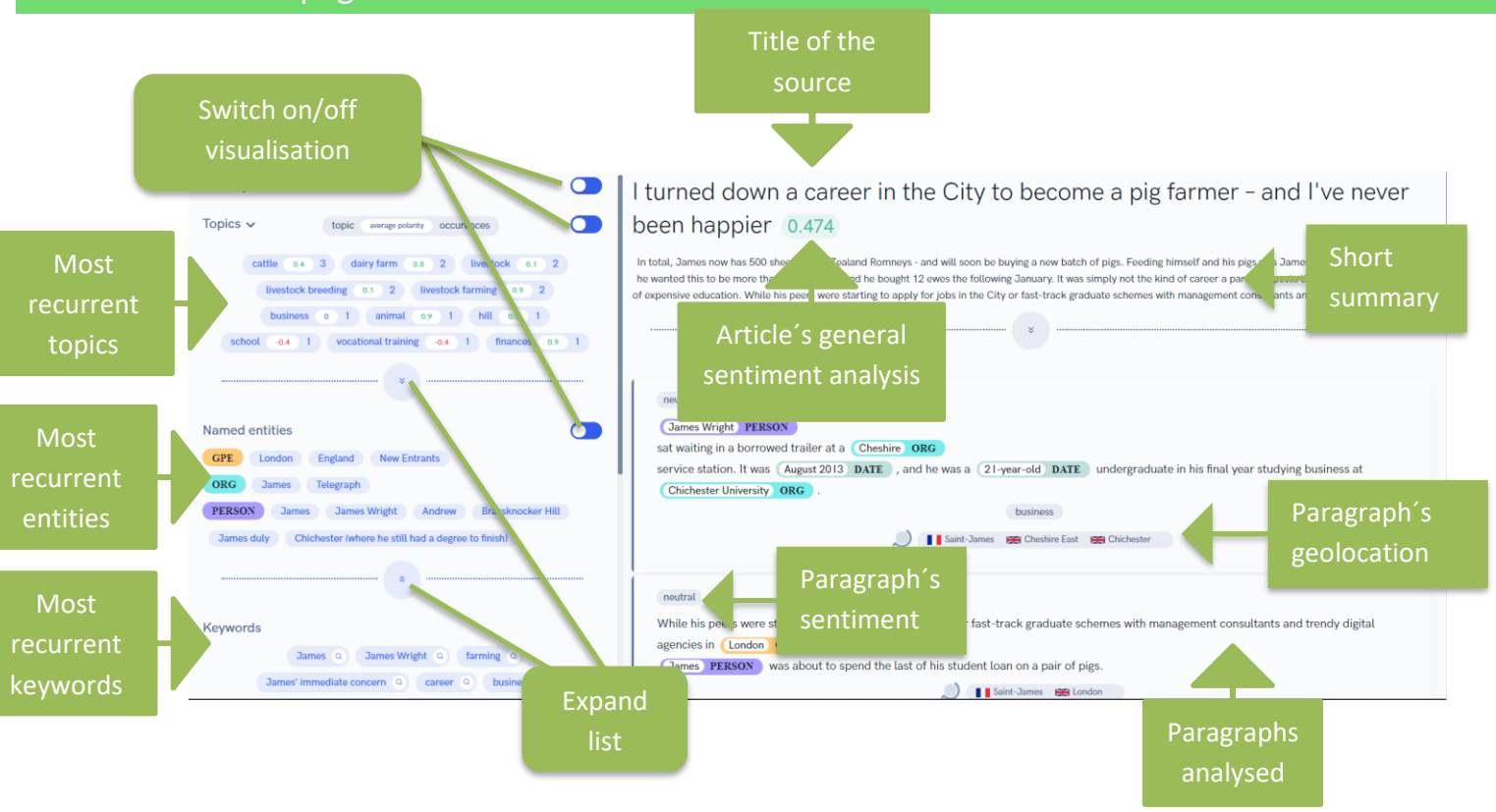


The left panel offers the possibilities of applying various filters in order to narrow and focus the text search in the Library. It is possible for example to set a specific language as well as a specific Topic (from GEMET). This option is especially interesting because it gives user the possibility of performing a Boolean search using the AND/OR operators. For example, if one is interested in finding the sources linked with the Topic 'common agricultural policy' which contain the word 'EAFRD' one should insert `EAFRD` in the search bar and `common agricultural policy` in the left panel Topics section and choose the operator AND.

In the filter panel it is also possible to add a specific polarity to look for text with a negative or positive sentiment. For example, by setting the polarity to -1 to 0 the system will display only the negative paragraphs linked to the Topic `common agricultural policy` and containing the term `EAFRD`. This can be useful in order to collect negative and/or positive text about a certain topic.

From this page it is also possible to add sources in Curated Reading List by clicking on the CRL icon. For more information about Curated Reading List jump to Chapter 4.

4.2 Source result page



The source result page provides the semantic analysis results for one source.

The **right panel** displays the original text extracted from the source, providing from top to bottom:

- the Title;
- the general sentiment of the text;
- a short summary;
- all the paragraphs with related:
  - sentiment analysis,
  - named entities,
  - geolocations,
  - Topics

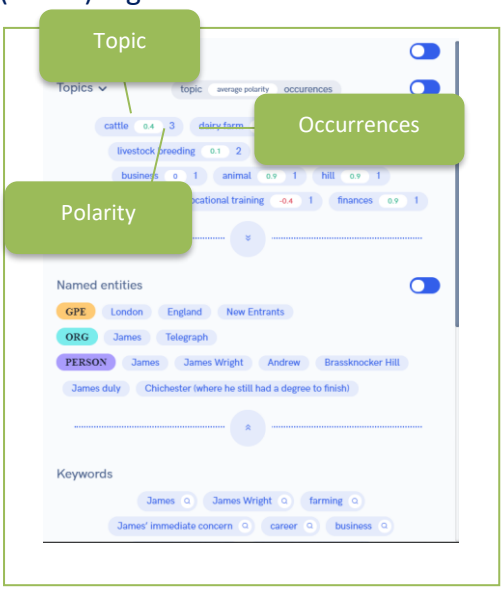
It is possible to switch on/off the visualisation of these extracted information by clicking on switchers on the left menu.

By clicking on the icon user can expand the Topics, Named Entities, Keywords and Summary sections.

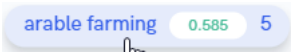
The left menu displays for each source the extracted Topics, Named entities and the results extracted from the original text.

4.2.1 Topics

These are the Topics that best represent the specific paragraph. The system automatically extracts keywords and compares them to a list of GEMET topics, selecting the ones that are most similar. The comparison is done by Word Mover's Distance1 (WMD) algorithm which is one of the most accurate algorithms available to semantically compare documents.
Thus, Topics are not necessarily words that are present in the text. Topics represent semantically the meaning of the text. Moreover, each paragraph is usually linked to more than one Topic.



By clicking on one of the Topic on the left menu the system will display only the paragraphs that are linked to that paragraph. It is also possible to click on more than one Topic in order to view the paragraph(s) linked to ONE or ALL the Topics, depending on the filter specification.

For each Topic the icon 

provides 3 information:

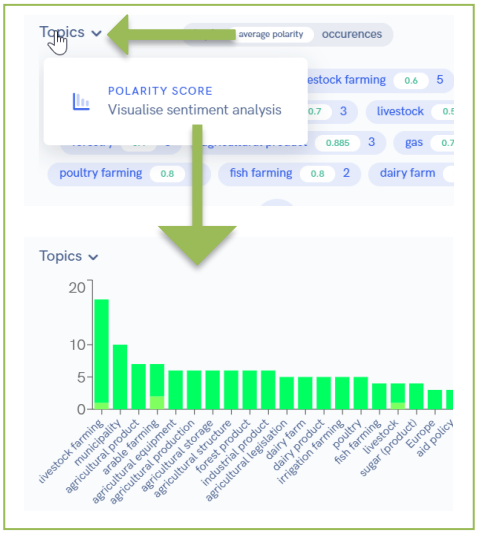
- the Topic;
- the average polarity and
- the number of occurrence(s) in the original source.

It is also possible to visualise the occurrences and polarity in a small histogram:

Click on Topics>POLARITY SCORE

To get back to the list of Topics:

Click on Topics>Back to topic list



[1] "From Word Embeddings To Document Distances." http://proceedings.mlr.press/v37/kusnerb15.pdf. Accessed 13 May. 2020.

4.2.2 Named entities

One feature of text mining is that it can automatically detect in the text some entities such as persons, locations, organizations, laws etc… In the case of semex.io the named entities extracted are the following:

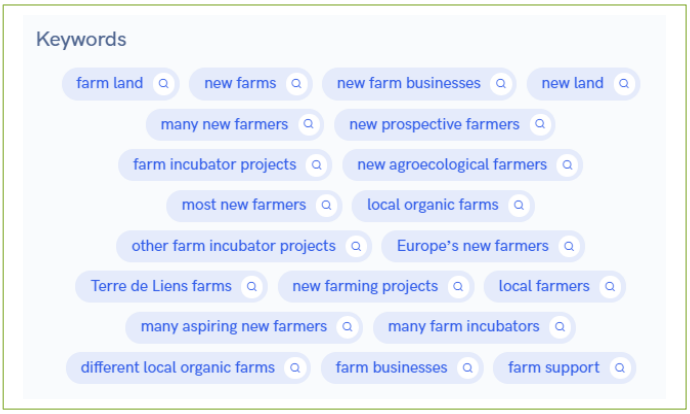
- PERSON: People, including fictional.
- NORP: Nationalities or religious or political groups.
- FAC: Buildings, airports, highways, bridges, etc.
- ORG: Companies, agencies, institutions, etc.
- GPE: Countries, cities, states.
- LOC: Non-GPE locations, mountain ranges, bodies of water.
- PRODUCT: Objects, vehicles, foods, etc. (Not services.)
- EVENT: Named hurricanes, battles, wars, sports events, etc.
- WORK_OF_ART: Titles of books, songs, etc.
- LAW: Named documents made into laws.
- LANGUAGE: Any named language.
- DATE: Absolute or relative dates or periods.
- TIME: Times smaller than a day.
- PERCENT: Percentage, including ”%“.
- MONEY: Monetary values, including unit.
- QUANTITY: Measurements, as of weight or distance.
- ORDINAL: “first”, “second”, etc.
- CARDINAL: Numerals that do not fall under another type.
- MISC: Miscellaneous.

The ‘Named entities’ section in the left panel displays all the entities extracted from the original source. By clicking on one entity of the right menu the system will display on the left section of the page the paragraphs where the particular entity is present.

4.2.3 Keywords

The system automatically identifies the words that are most important in a given text. These keywords are extracted and displayed in the left menu. They are also used to create the summary of each source.

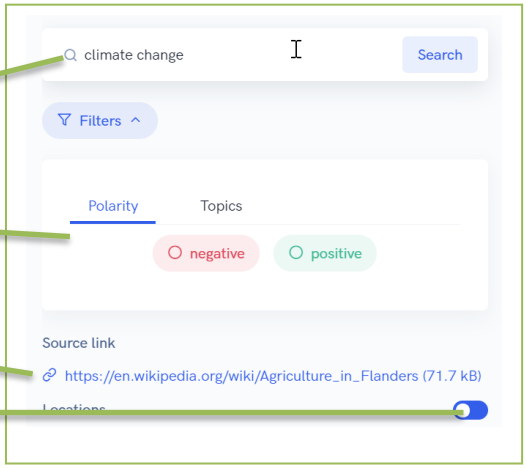
By clicking on one keyword user will be redirected to the Library's search page to search for that particular keyword in all the sources contained in the Library.



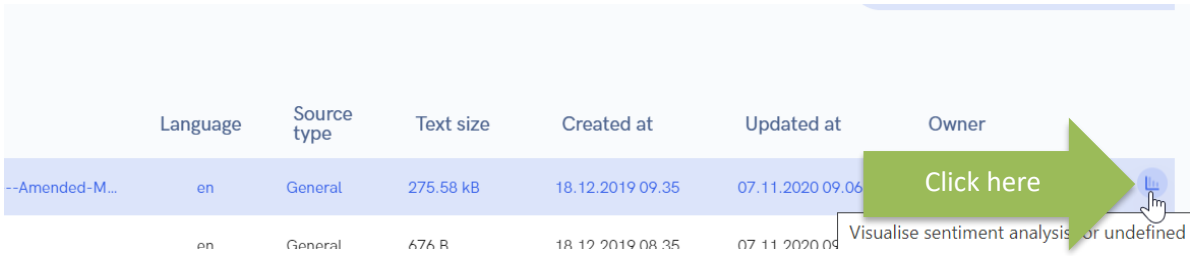
4.2.4 Other options

By scrolling down in the left menu some other options are available.

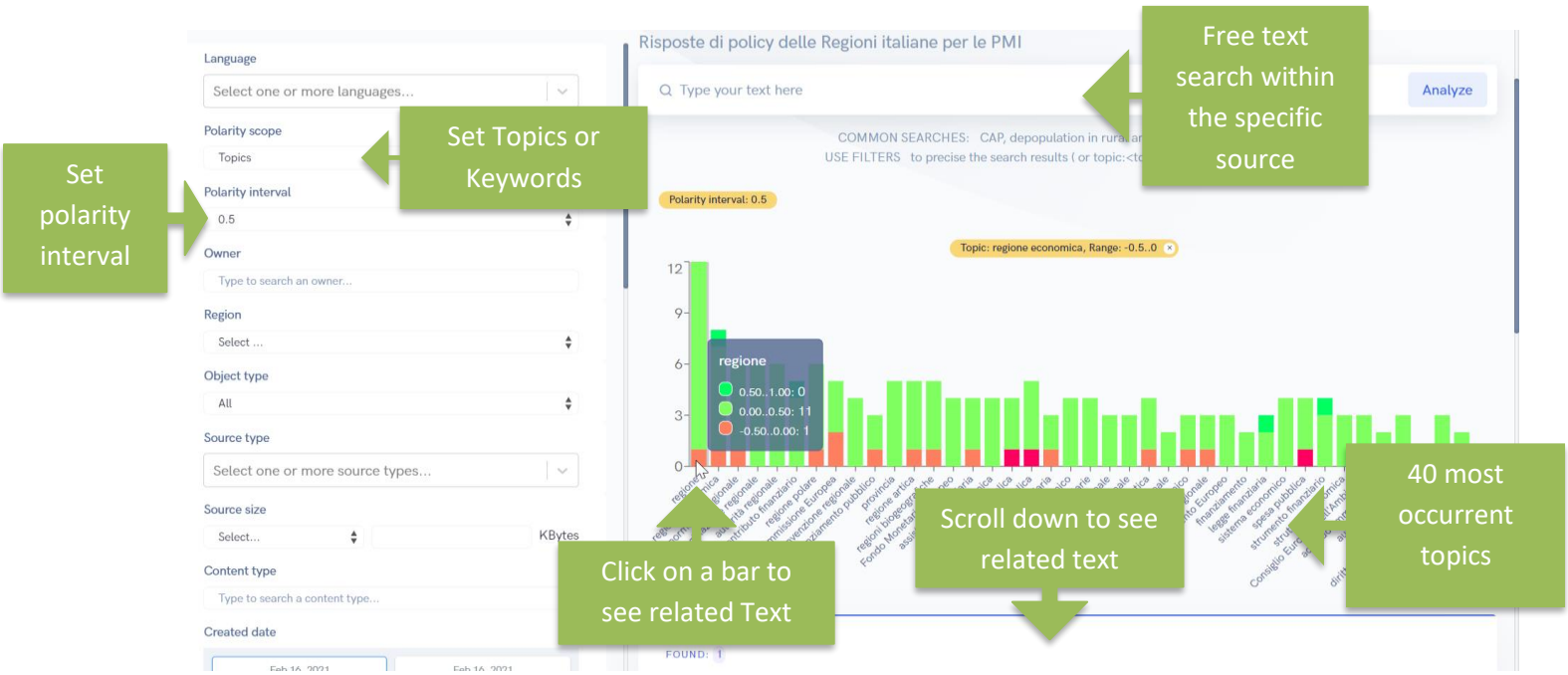
- It is possible to search free text within the source. In the example we searched for 'climate change'.
- It is possible to set some additional filters for the text to be displayed such as Negative or Positive Polarity; ANY or ALL Topics
- Source link provides the original link
- It is possible to switch on/off the display of geolocation



4.3 Sentiment analysis histogram



By clicking on the icon on the right of the source line, user is redirected to the sentiment analysis histogram.



The sentiment analysis histogram is a visual representation of the results from Topics and Keywords extraction from one source. The histogram bars show the number of occurrences for the 40 most recurrent topics or keywords. The various colours represent the sentiment polarity (-1 to 1) for each topic with full red indicating a very negative sentiment and full green very positive sentiment. Polarity is pre-set with an interval of 0.5, but it is possible to change it in the left menu in the Polarity interval option to have more granularity.

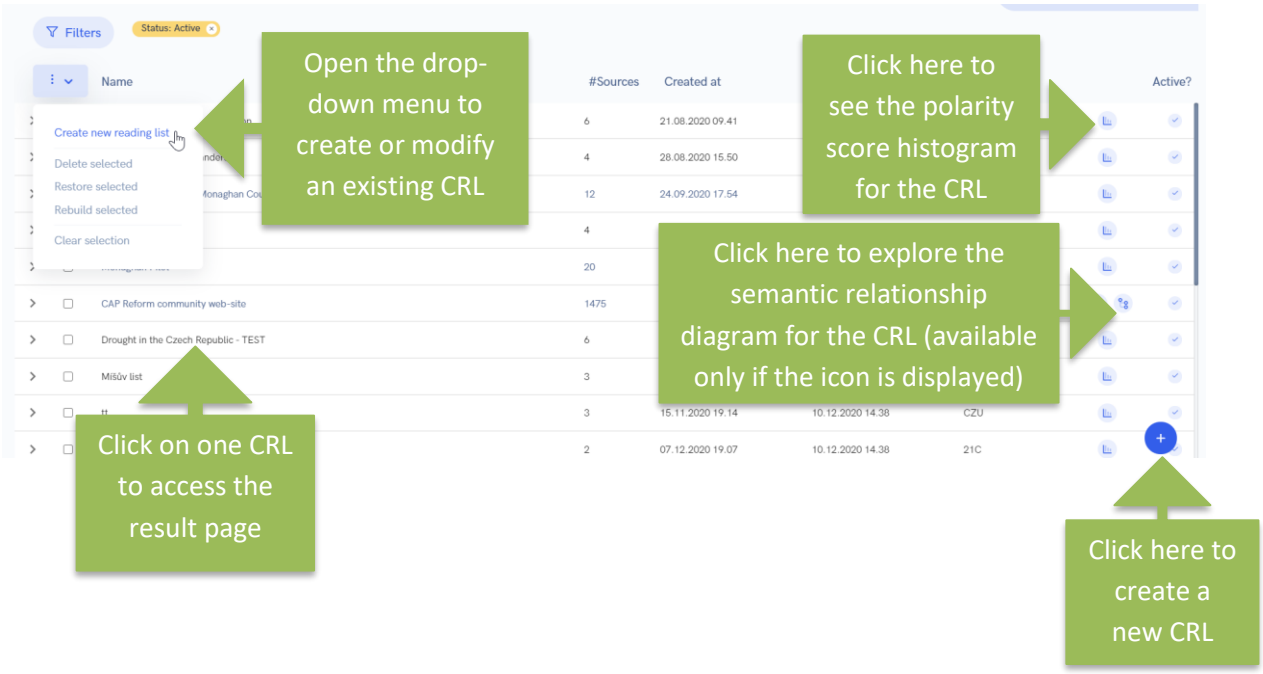
Users can click on a specific bar and a specific polarity, as defined by the colours, to see the related paragraphs by scrolling down in the right section of the page.

User have also the possibility of searching in a specific text thanks to the upper bar that allows to free text search.

On the left menu there are various other possible filters that a user can try.

5. Curated Reading List

The idea of creating a Curated Reading List (CRL) section is the result of active dialogue between text mining developers and other partners of Polirural project. In particular, the idea of having a repository for interesting sources rose from the necessity in Foresight activities of going through a big amount of literature to prepare workshops such as the Deep Dives. Curated Reading Lists can therefore be collections of sources about a specific area of interest. The system provides, not only the possibility of storing sources in a determined repository accessible whenever, but also the aggregated analysis such as summary, Topics, NER, Keywords, Wordcount and extracted URLs.



5.1How to create a reading list

There are various ways to create Curated Reading Lists in semex.io. A CRL can be created with sources that have been previously uploaded in the Library or with completely new links to new sources. It is also possible to create CRLs from various sections of the website and in particular from:

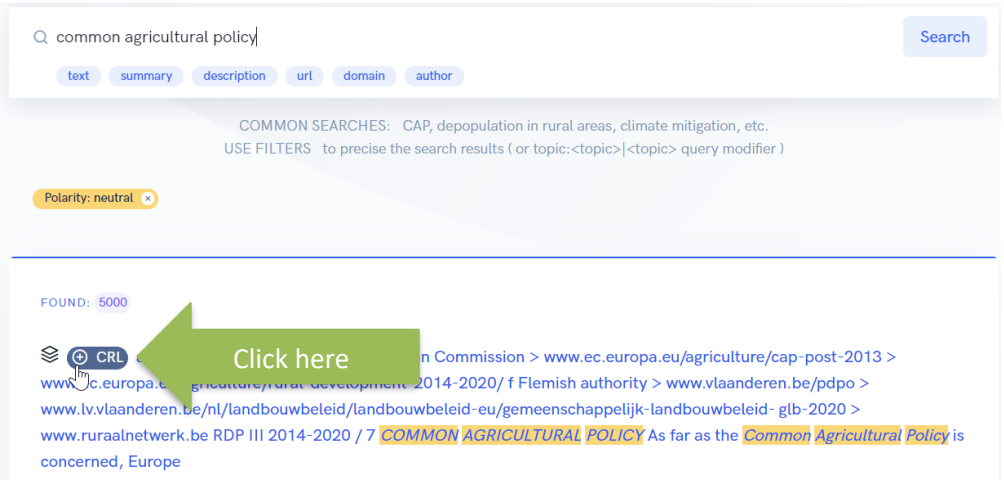
- Library
- Search results
- Reading List

5.1.1 CRL from Library

As explained in the related section it is possible to add sources to a CRL directly from the Library. In Library select the required sources, open the down menu on the left of the screen, choose the `Add to reading list` option.

5.1.2 CRL from Search result

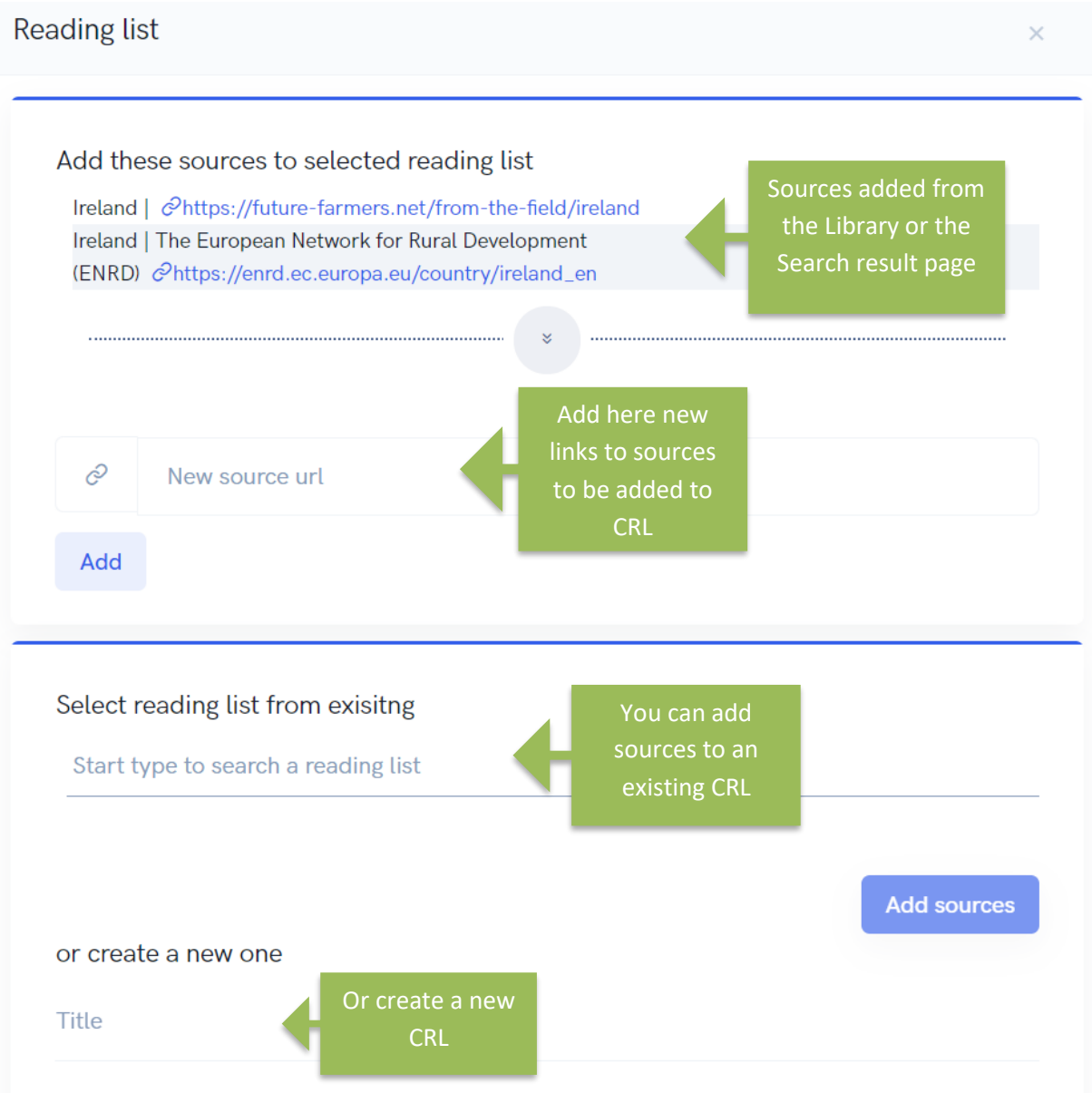
As explained in the Search option section it is possible to add sources to CRL directly from the result page of the search function by clicking on the CRL button.



5.1.3 CRL from Reading List section

The Reading List section is the repository where all the CRL created can be accessed. However, it also provides the possibility of creating new ones by clicking the + button or by opening the drop-down menu on the top left of the page.

The three options described above will open the following window to complete the creation of the CRL.



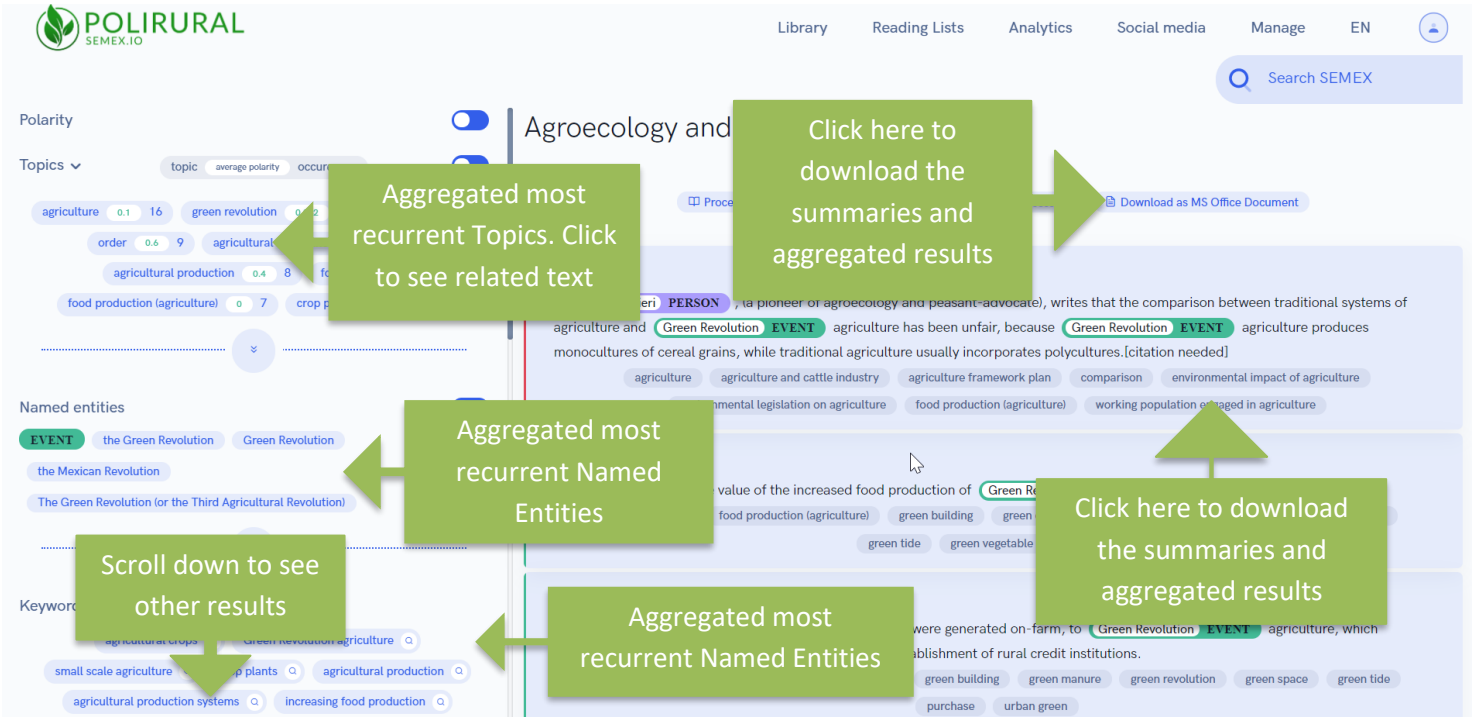
In order to add new sources it is necessary to:

- type the url in the `New source url` space
- click on Add
- Repeat this action for all the sources that you want to add to the CRL

You can add the new sources to an existing reading list or create a new one by adding a new title and press the Create & Add button on the bottom left of the page (scroll down if you cannot see the button). The system will inform the user that is creating the reading list and that some time is needed for processing. The user will need to wait for the indicated time to access the newly created CRL.

5.2 CRL Result page

The CRL result page is very similar to the Source Result Page (see section 3.2), but while in the Source the results are relative to one source only, in CRL the results displayed are the ones aggregated from the several sources composing the CRL itself.

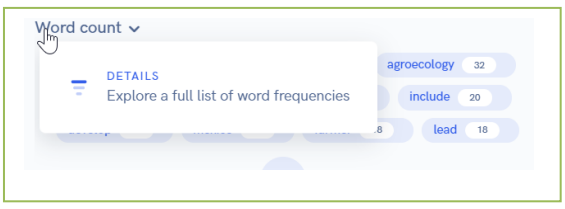


The left panel is dedicated to extracted Topics, Named Entities, Keywords and single words from the different sources. The system extracts all this information from each source and then it aggregates the results that occur most frequently.

The right panel displays the text linked to some of the extracted semantic features of the left panel. For instance, by clicking on the Topic 'agriculture' on the left panel the system will display the paragraphs linked to that 'Topic'.

Due to the big amount of text of some of the CRL it is only partially possible to retrieve text linked to specific named Entities. However, the aggregated result can be used to have a general understanding of what the text is about with a few clicks.

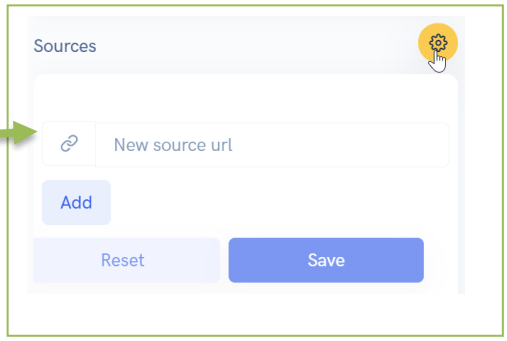
Following internal request, in CRL it is also possible to see which single words are the most recurrent through the Wordcount section in the left panel. It is also possible to expand the list and open a specific panel by clicking on Wordcount>DETAILS.



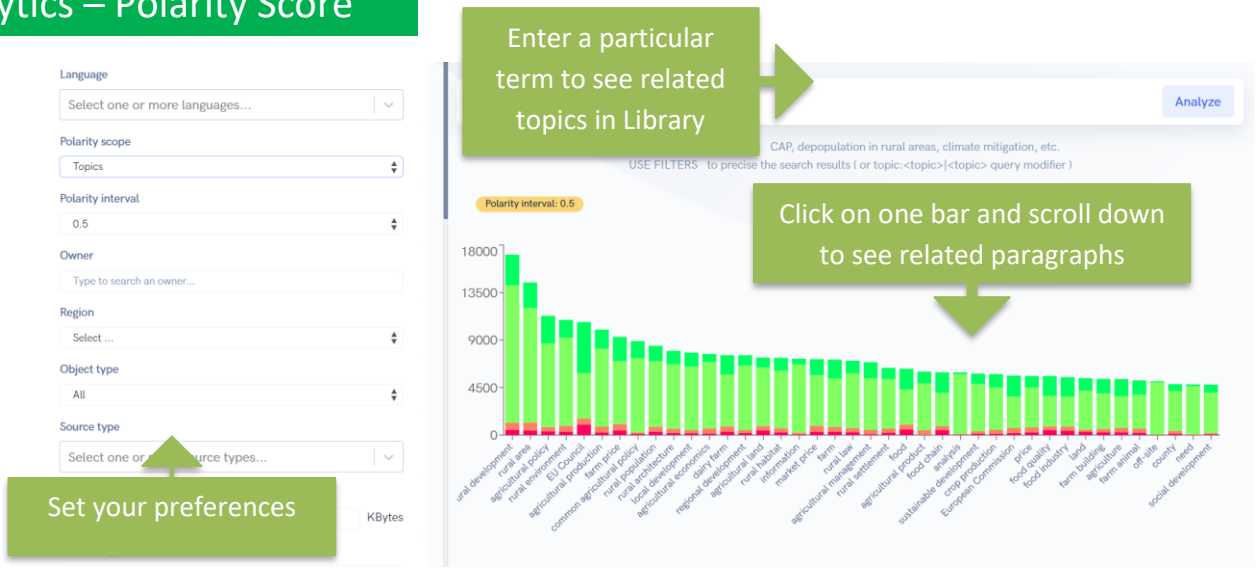
All the results, including the summaries of each source and an overall summary, are available in a downloadable MS Word document by clicking on the relative option, as illustrated in the image.

In the bottom of the left panel user can add some additional filters for the text to be displayed such as Polarity and Topics (click one or more Topic to see related text).

Finally, it is possible to access the links to the original sources as well as adding new sources to the existing CRL by clicking the button.



6. Analytics – Polarity Score

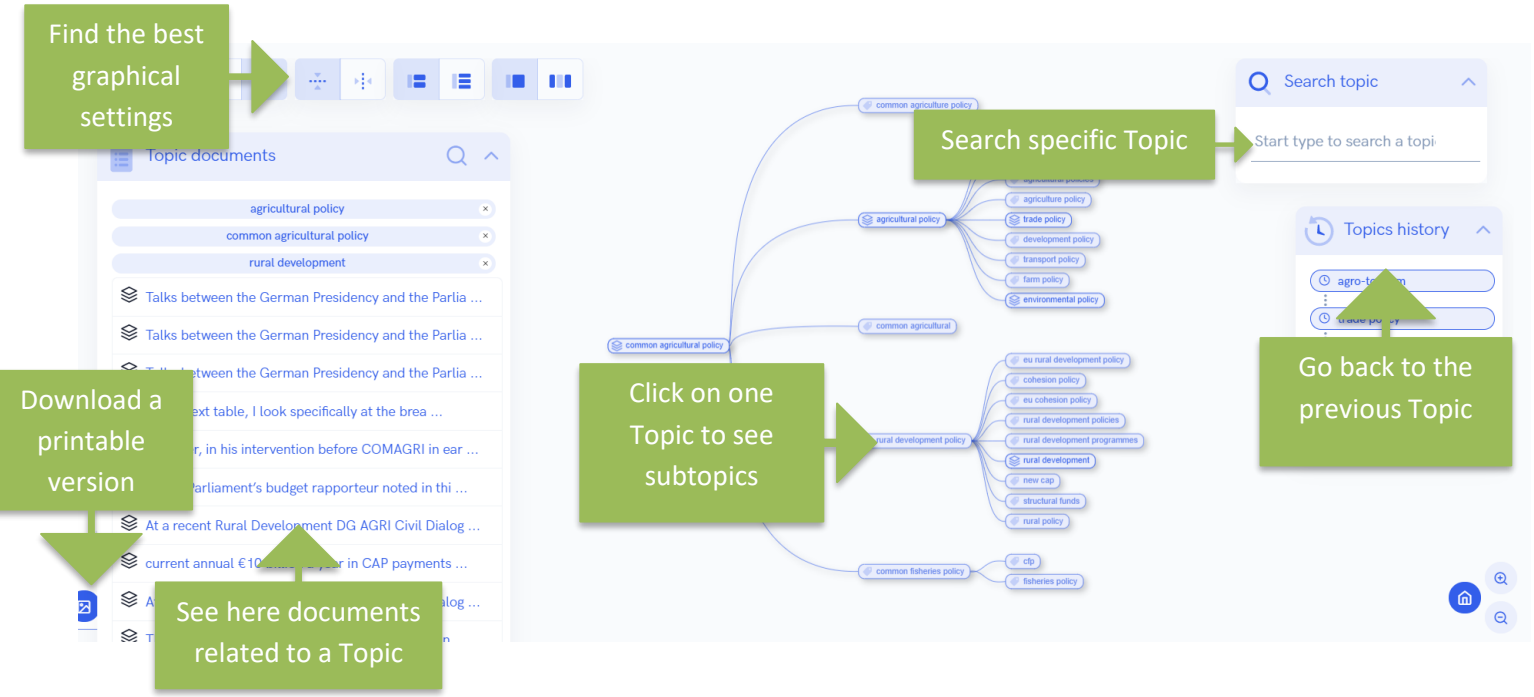


The polarity score section is a visual representation of the contents of semex.io entire library. The right panel displays a histogram with 4O bars, which are the most recurrent topics or keywords in the Library. The various colours represent the sentiment of the text with the red colours indicating negative polarity and green colours suggesting positive sentiment. As it is obvious in the picture above, most of the text is positive and this is mainly due to the fact that our Library contains mostly scientific papers and reports where the tone is rather positive. It needs to be specified that sentiment analysis has been developed and extensively experimented by businesses to detect customers satisfaction and therefore show some limitations in handling text related to policy making which usually contains a rather diplomatic language. All things considered, almost every bar in the histogram contains a small red section which represent the negative sentiment of the text which is available for visualisation by clicking directly on the bar and scrolling down.

Through the left panel a user can apply some filters and modify some of the aspects of the histogram. With the 'Polarity scope' it is possible to choose between Topics and Keywords. As already explained, Keywords are the most important words extracted from the text through the Graph-based TextRank algorithm, while to get Topics the system compares the above said Keywords to a set of topics defined in GEMET thesaurus and determines the most appropriate. Thus, keywords are words present in the text while Topics are mostly not present. Moreover, a limit of using a specified thesaurus is that it might miss some new words and topics. For example, GEMET does not include emerging topics such as COVID, 'Green Deal' and other recent concepts which were not in use when the last update of GEMET was done.

The 'Polarity interval' option allows to add more or less granularity in the sentiment analysis and to choose between 4 different options. Users might be interested in verifying the differences between 'positive' vs 'negative' or more fine grained evaluations like very positive, slightly positive, very negative, slightly negative etc…

7. Analytics – Topic explorer



Topic explorer is a semantic tree displaying links between Topics and Subtopics from semex.io thesaurus. A semantic tree is a diagram representing nodes labelled with relation and topics. In semex.io the Topic explorer utilises the semantic trees to represent the semantic relations among the Topics occurring in the GEMET thesaurus.

For humans it can be easier to find a specific field of interest in these hierarchically structured lists of topics for the following reasons:

- Relations of the topics help us to find the specific topics we are interested in faster (browsing the thesaurus);
- We can visualize the relations of the topics which can help with finding similar topics which also describe the area of our interest (topic explorer in semex);
- The thesaurus is a closed set of topics and it's easier to orientate in a limited number of topics.

By clicking on one of the topic the system will generate a new tree with new subtopics associated. Topics and subtopics are linked to specific sources in our Library. The icon indicates that a topic is linked to one or more sources displayed in the left 'Topic documents' panel.

It is possible to search for a specific topic through the 'Search topic' menu on the top right of the screen. The Topics history panel gives the possibilities of looking into search history and go back or forward to some of the user's visualisations. The icons on the top left allow user to customise the graphical settings of the semantic tree.

In semex.io it is also possible to generate a semantic tree from a large dataset (ca. 800) of sources. Based on the text contained in the sources the system creates a specific semantic model. For instance, the Curated Reading List created from the articles from the blog www.capreform.eu has its own semantic tree that is accessible in the Curated Reading List section, by clicking on the icon  of the specific reading list.

8. Social media

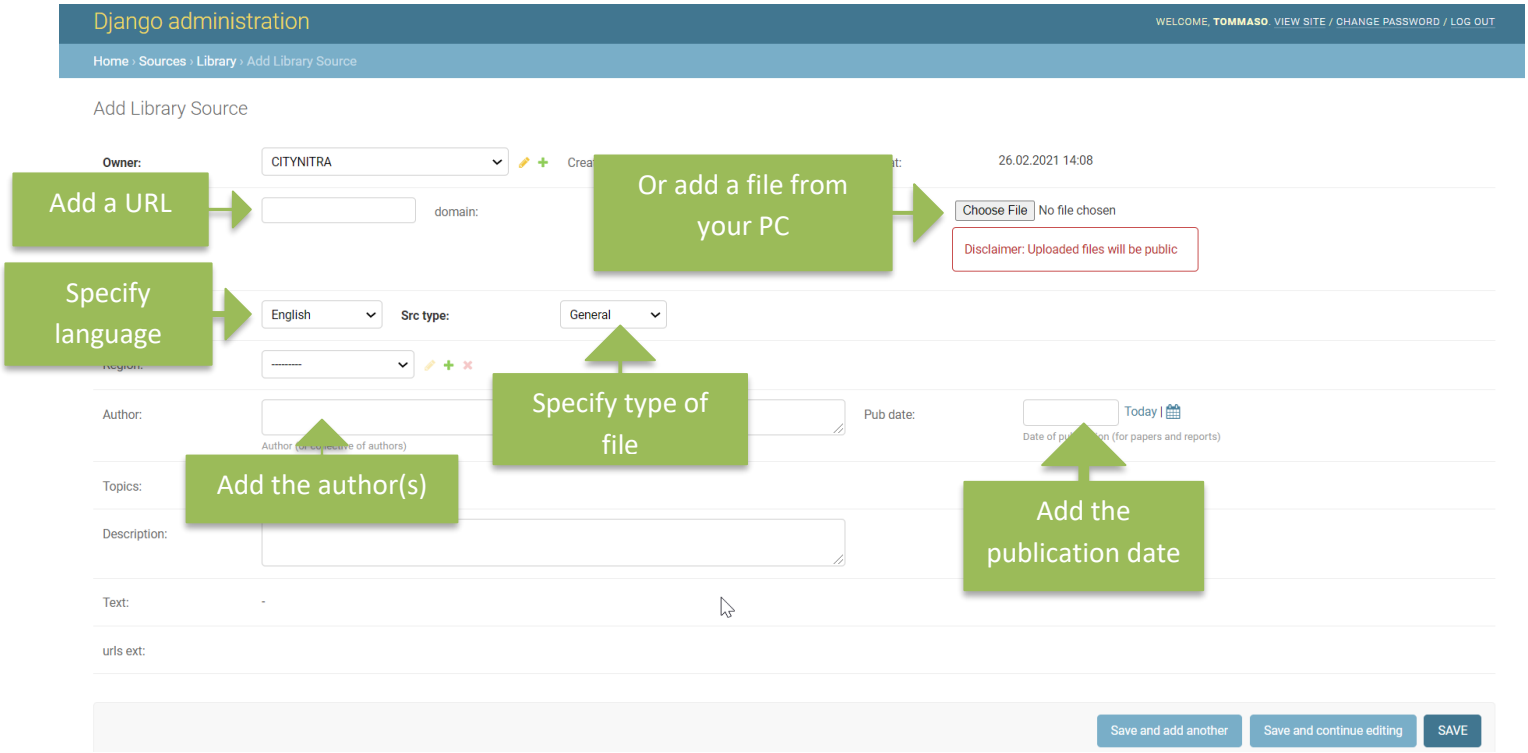
The **Social Media** section displays the results of continuous streaming of messages from Twitter with information related to Polirural research topics. Messages are streamed based on specific Twitter users that are particularly active in Polirural research field, as well as on the basis of specific hashtags such as #CAPreform, #EUFarm2Fork, #EAFRD, #Greendeal etc…
The Social Media section is accessible from the top menu by clicking on Social Media.

The left panel allows user to add filters for narrowing the results. It is possible to select a Language as well as the possibility of choosing between Topics or Keywords (differences between these possibilities are explained in section XY of this guide). It is also possible to choose between various polarity intervals in order to visualise more in detail the different polarities. In Social Media, an important filter is the Date field since from December 2020 the system is streaming particular Twitter users and hashtags. Before this date the system was streaming Tweets based on the thesaurus, but the results were not always related to Polirural field of research.

The right panel displays the results. It is possible to insert some free text in the Analyze bar to visualize the Topics/keywords that are related. The bar chart shows the number of Tweets and the various colours represent the polarity (Green: 1 to 0; Red 0 to -1). It is possible to visualise the related Tweets by clicking on one of the bars and on a particular polarity. By clicking on the red section of the bar the system will show the negative related messages.

9. Manage

The manage section is accessible only for members of the Polirural consortium and serves the purpose of adding, modifying and deleting Library's sources. By clicking on the Manage button users are redirected to a Django administration section with the complete list of sources. Authorised users can add sources by clicking on the button ADD LIBRARY SOURCE + on the top right of the page. In order to obtain username and password please drop an email to the KAJO team.



Initially the system allowed adding only sources available on the Internet by providing a link. Following internal request, the option of adding files directly from personal hard drives has been developed giving users both options.

User must specify the correct language of the text otherwise the system will not work properly.