Text mining to explore Food Systems

13.5.2021

Tuula Löytty, Smart & Lean Hub Oy, Finland

Introduction

Semantic Explorer (SemEx) is one of the technical and innovative outputs of the PoliRural[1] research project financed by the Horizon2020 Program of the EU [1]. SemEx's ambitious vision is to provide support to researchers and facilitators by reducing the cognitive load related to tasks that are essential to research, development, and innovation actions.

CITIES2030 partner P14 SLEAN[2] is also a PoliRural consortium partner. SLEAN exploits SemEx in the CITIES2030 work package 3 and task 3.4 which is about systems thinking, system modelling and food systems.

This document demonstrates the results of the experiment. The aim of the experiment is to test the power, features, and outputs of SemEx, and also limitations, to assess its usability and usefulness.

The main chapters of the document are:

- Main Library: added sources
- Curated Reading List (CRL): Food Systems
- Three tools for analytics: Topic Explorer, Polarity Scores and Social Media (incl. KIBANA)

Main Library

The Main Library is built of sources related to needs of knowledge. Sources include online news articles, discussion forums, academic papers and blogs.

For this experiment, about 800 new digital sources were added to the Main Library. They were sources that were addressed on **food, food system, urban food system, food value chain, system thinking and system modelling.** The sources were found by using the above keywords in Google search. The human operator selected the best sources (technical and content quality) and she/he manually added the source links into the Main Library.

Curated Reading List (CRL)

Curated Reading List

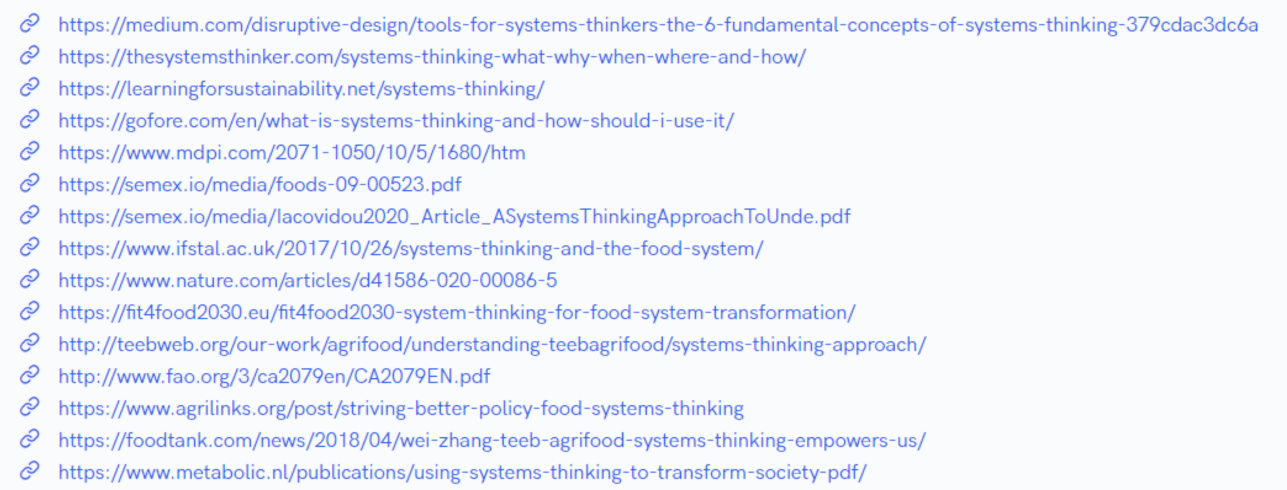
A second section of the repository is the Curated Reading Lists, containing collections of sources created by solution users. The results are Topics, Keywords, Named Entities and Word Count which are aggregated from multiple sources. The additional information of the Curated Reading List is in the SemEx Hand-on-Manual Chapter 5 [2].

Curated Reading Lists are collections of sources that are connected to the specific area of interest. In this experiment, the specific area of interest is linked to the following keywords: **food, food system, urban food system, food value chain, system thinking and system modelling**. The name of the aggregated Curated Reading List is “**Food Systems**”.

In this experiment, Food Systems CRL was created semi-automatically based on the added circa 800 new sources. It means that ICT experts’ intervention is needed. Because there were diverse technical reasons, all 800 new sources were not available. The reading list for “Food Systems” also contains only 547 sources. Nevertheless, the created reading list is an impressive collection of sources relevant to the specific area of interest, that is Food Systems. The screenshot 1 shows an example of the sources.

[1] PoliRural H2020 project web-page: www.PoliRural.eu

[2] Smart & Lean Hub Oy, Finland, www.smartlean.fi



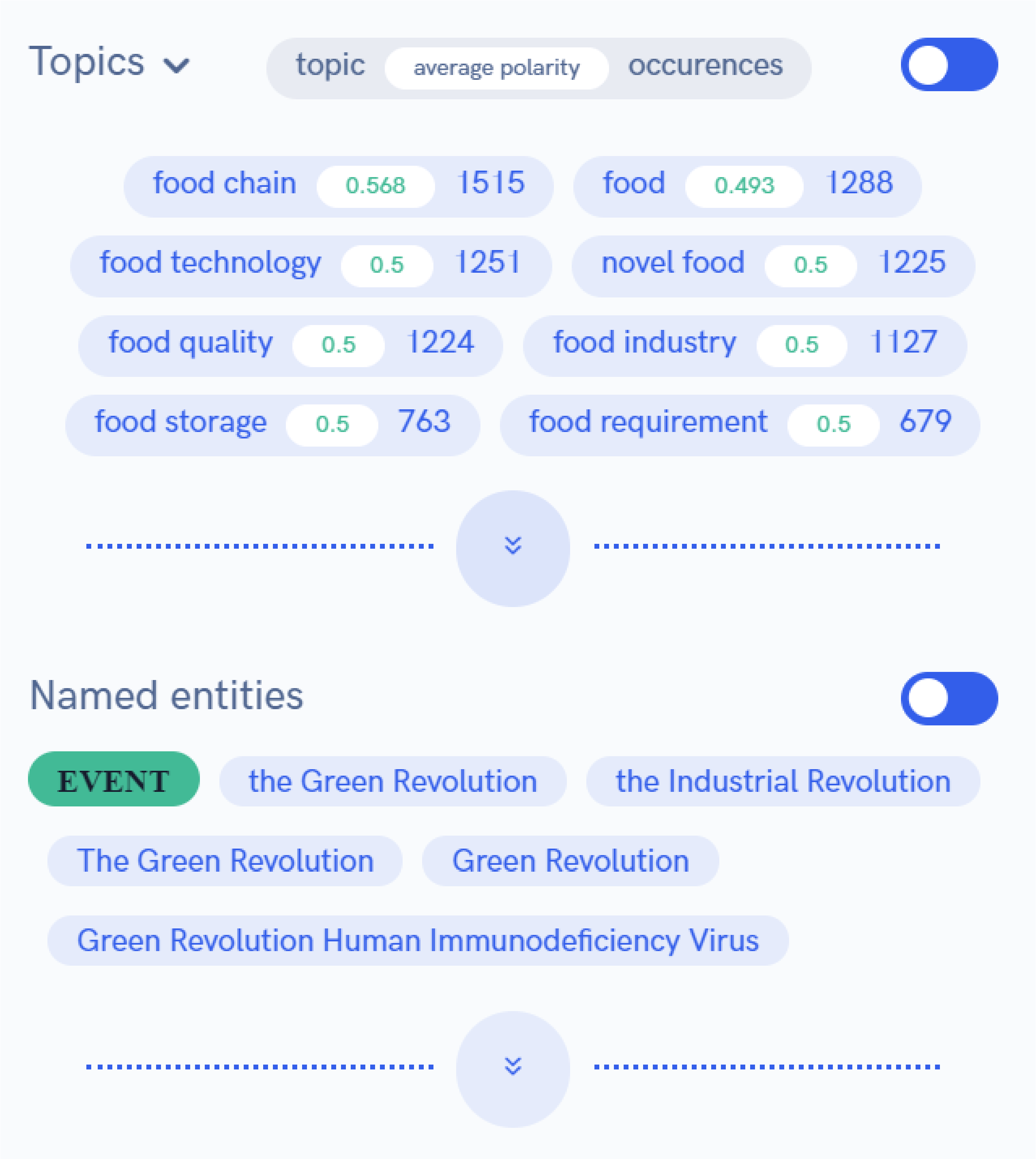
- https://medium.com/disruptive-design/tools-for-systems-thinkers-the-6-fundamental-concepts-of-systems-thinking-379cdac3dc6a
- https://thesystemsthinker.com/systems-thinking-what-why-when-where-and-how/
- https://learningforsustainability.net/systems-thinking/
- https://gofore.com/en/what-is-systems-thinking-and-how-should-i-use-it/
- https://www.mdpi.com/2071-1050/10/5/1680/htm
- https://semex.io/media/foods-09-00523.pdf
- https://semex.io/media/Iacovidou2020_Article_ASystemsThinkingApproachToUnde.pdf
- https://www.ifstal.ac.uk/2017/10/26/systems-thinking-and-the-food-system/
- https://www.nature.com/articles/d41586-020-00086-5
- https://fit4food2030.eu/fit4food2030-system-thinking-for-food-system-transformation/
- http://teebweb.org/our-work/agrifood/understanding-teebagrifood/systems-thinking-approach/
- http://www.fao.org/3/ca2079en/CA2079EN.pdf
- https://www.agrilinks.org/post/striving-better-policy-food-systems-thinking
- https://foodtank.com/news/2018/04/wei-zhang-teeb-agrifood-systems-thinking-empowers-us/
- https://www.metabolic.nl/publications/using-systems-thinking-to-transform-society-pdf/

Screenshot 1: The slice of the automatically created Curated Reading List for Food Systems

The SemEx 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.

Curated Reading List - Result page

The CRL result page aggregates e.g. extracted TOPICS, NAMED ENTITIES, KEYWORDS and WORDCOUNT from sources of Curated Reading List. The result page also presents the summary of each source. See the screenshots 2-5.



Topics

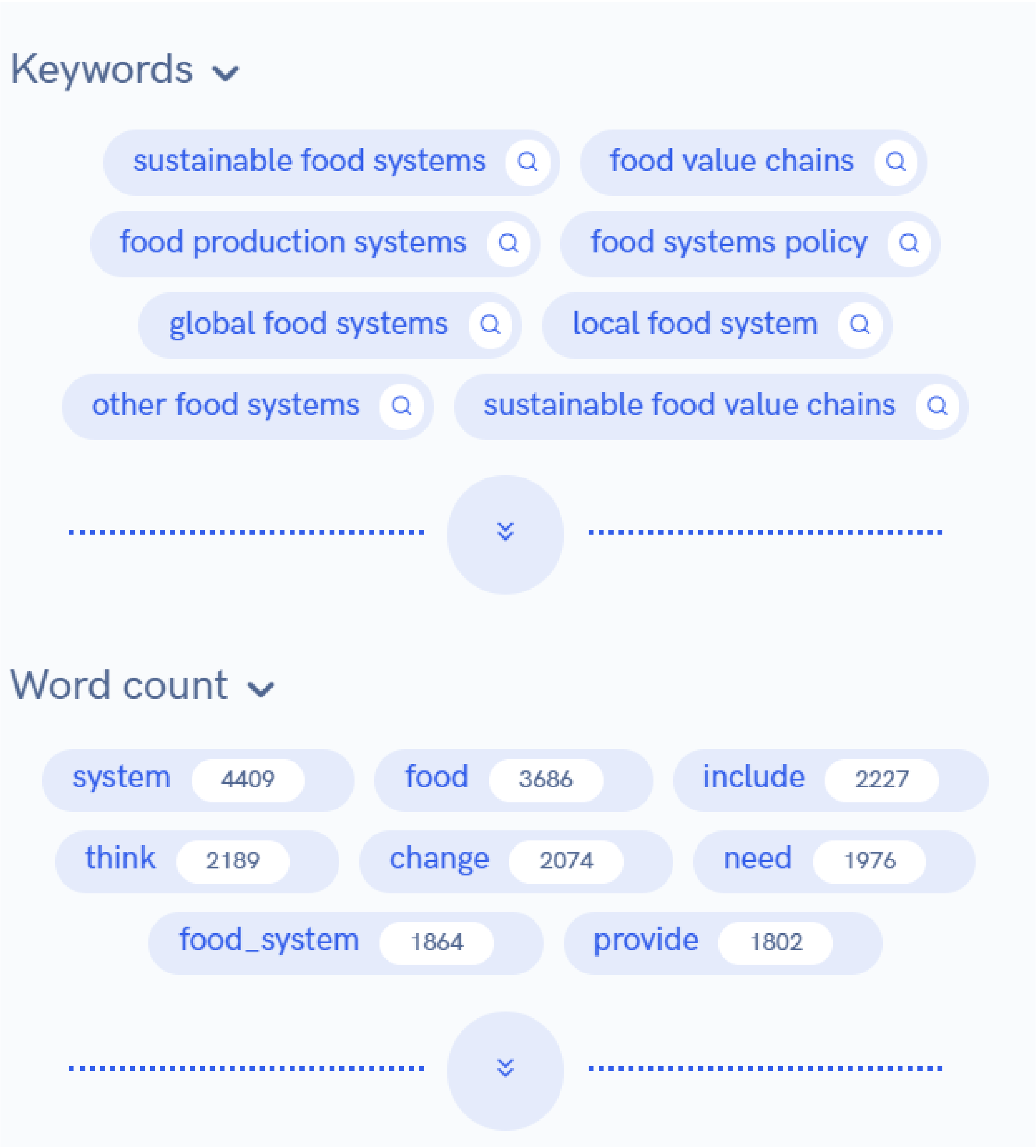
topic	average polarity	occurences
food chain	0.568	1515
food	0.493	1288
food technology	0.5	1251
novel food	0.5	1225
food quality	0.5	1224
food industry	0.5	1127
food storage	0.5	763
food requirement	0.5	679

Named entities

EVENT

- the Green Revolution
- the Industrial Revolution
- The Green Revolution
- Green Revolution
- Green Revolution Human Immunodeficiency Virus

Screenshot 2: Topics and Named entities of the Food System CRL



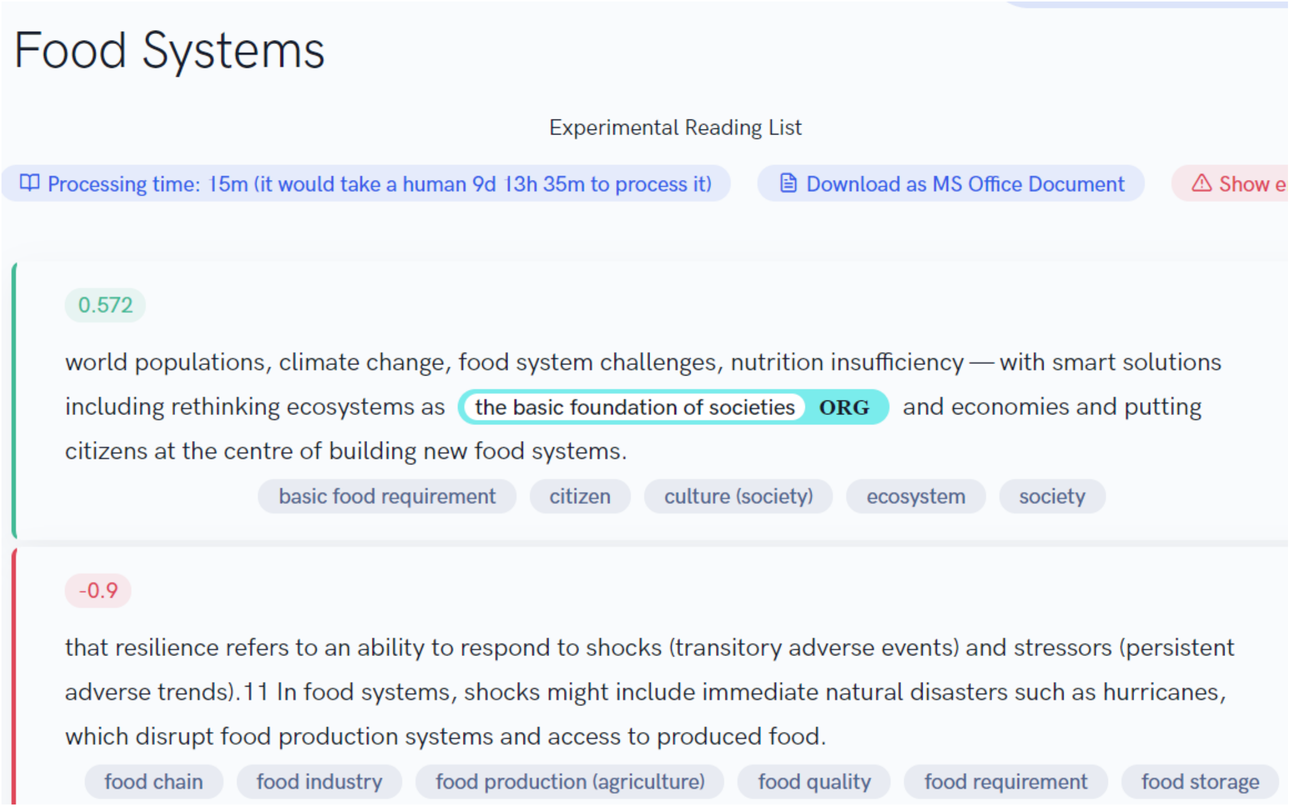
Keywords

- sustainable food systems
- food value chains
- food production systems
- food systems policy
- global food systems
- local food system
- other food systems
- sustainable food value chains

Word count

system	4409
food	3686
include	2227
think	2189
change	2074
need	1976
food_system	1864
provide	1802

Screenshot 3: The Keywords and Word count of the Food Systems CRL



Food Systems

Experimental Reading List

Processing time: 15m (it would take a human 9d 13h 35m to process it)
Download as MS Office Document
Show e

0.572

world populations, climate change, food system challenges, nutrition insufficiency — with smart solutions including rethinking ecosystems as the basic foundation of societies ORG and economies and putting citizens at the centre of building new food systems.

basic food requirement
citizen
culture (society)
ecosystem
society

-0.9

that resilience refers to an ability to respond to shocks (transitory adverse events) and stressors (persistent adverse trends).11 In food systems, shocks might include immediate natural disasters such as hurricanes, which disrupt food production systems and access to produced food.

Screenshot 4: The brief summaries of the Food System sources

Analytics

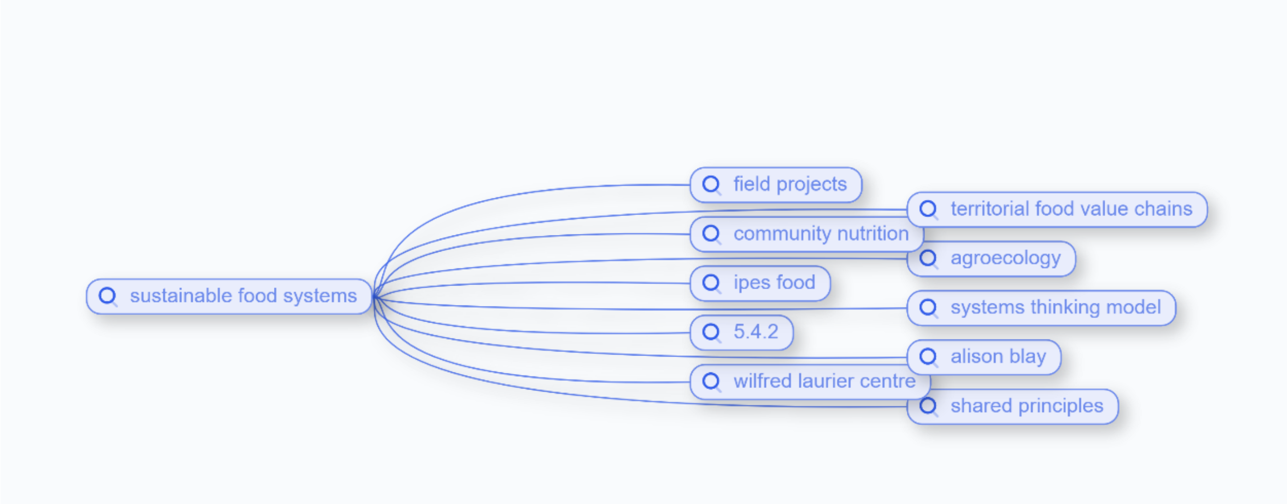
Analytics tools visualize various results from the system such as Topic Explorer and Polarity Scores, and to access Kibana for further analytical functions. SemEx Hand-on-Manual's chapters 6 and 7 respectively provide explanations about Polarity Score and Topic Explorer [2].

Topic Explorer

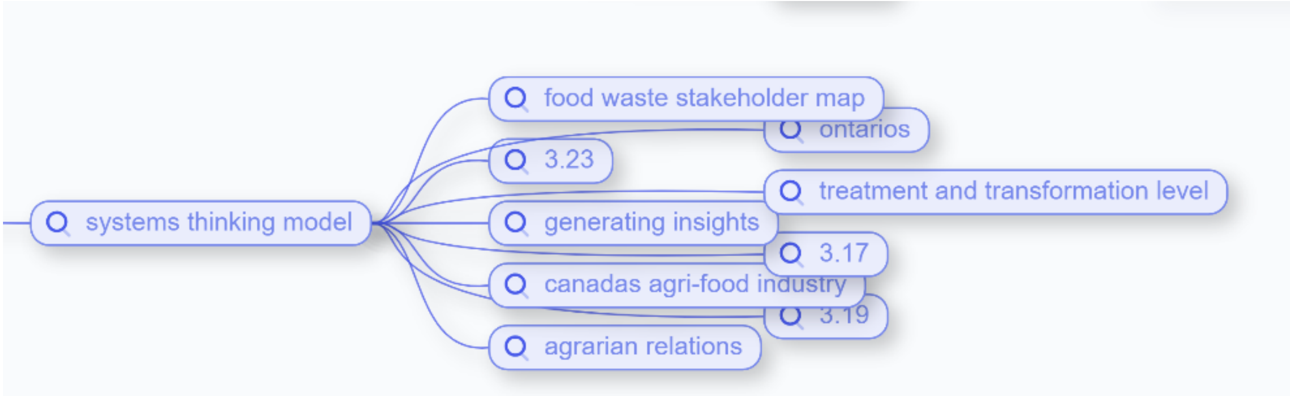
Topic Explorer is a semantic tree representing nodes labelled with relations, topics and subtopics.

In this experiment and analysis the semantic tree is generated from a large dataset of sources listed into the Curated Reading List for Food Systems. Based on the text contained in the sources the system creates a specific semantic tree model (see screenshots 5-7).

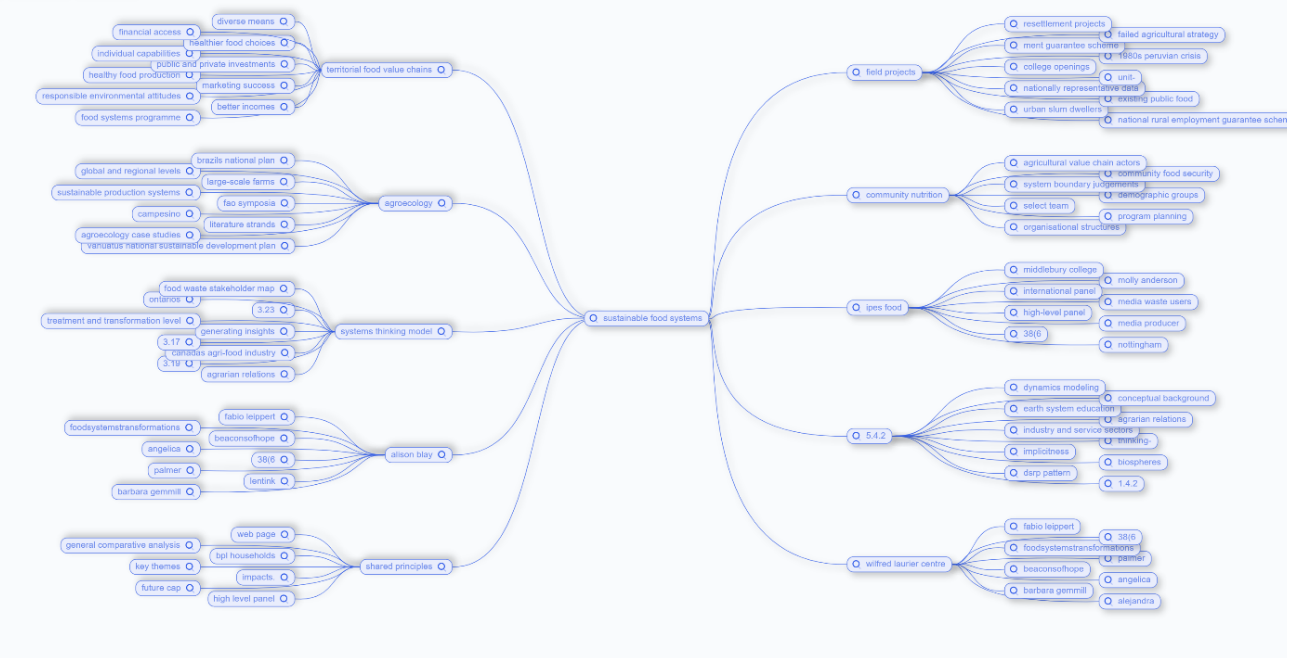
The results at screenshot 5-7 are **before "cleaning" the model**. The screenshots 5-7 show that the semantic tree includes strange topics and subtopics e.g. number and non relevant sub-topics. The model will require trimming and cleaning which is manual work made by PoliRural ICT-expert. The work will be executed at time which will be dependent on the expert's other tasks.



Screenshot 5: The first level branches for "Sustainable Food Systems"


systems thinking model
food waste stakeholder map
ontarios
3.23
treatment and transformation level
generating insights
3.17
canadas agri-food industry
3.19
agrarian relations

Screenshot 6: The second level branches for "Sustainable Food Systems" and "Systems thinking model"



Screenshot 7: The full semantic tree based on CRL for Food Systems sources

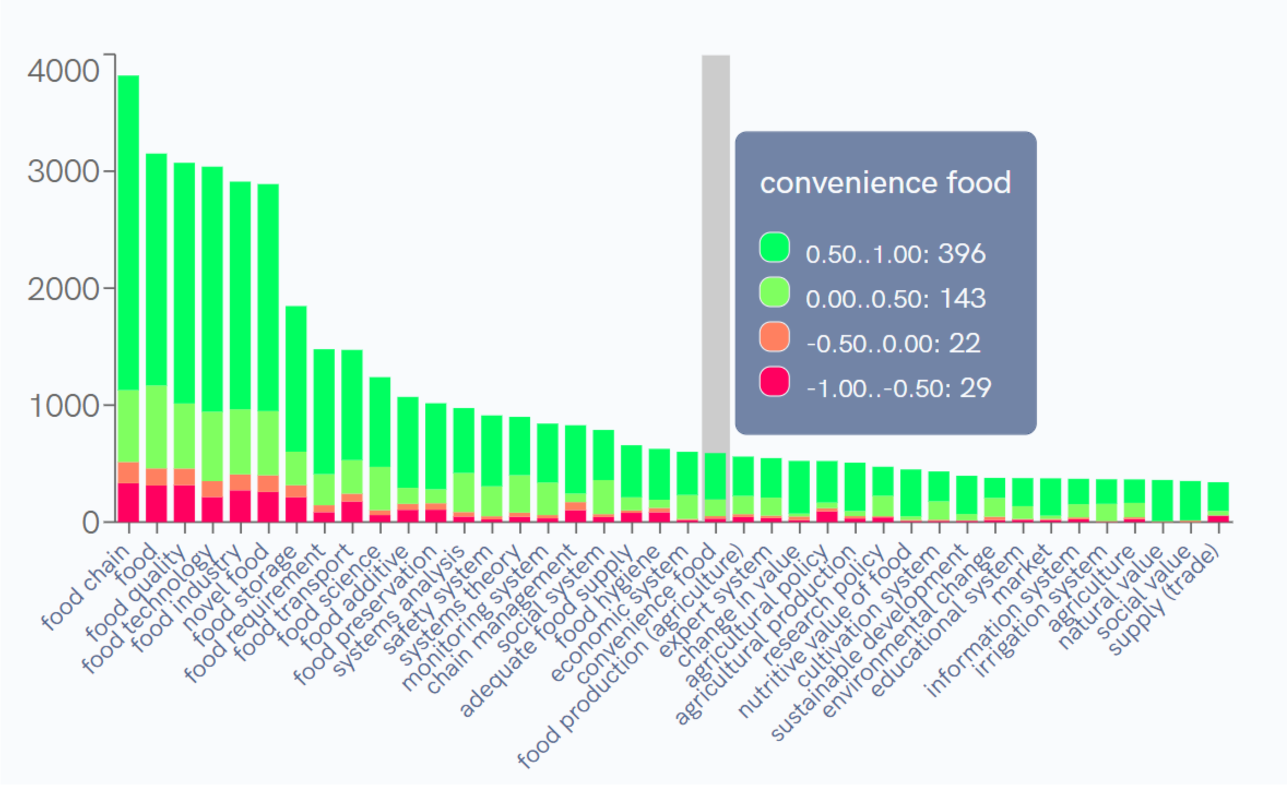
Polarity Scores

The polarity score section is a visual representation of the contents of semex.io based on the Curated Reading List for Food Systems (screenshot 8). The panel displays a histogram with 40 bars, which are the most recurrent topics. The various colours represent the sentiment of the text with the red colours indicating negative polarity and green colours suggesting positive sentiment.

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. Semex.io uses a predefined thesaurus called GEMET[3]. It contains more than 5000 topics in 37 languages. The topics of GEMET are closely limited to the PoliRural main topic, that is rural development.

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. GEMET topics seem to cover the scope of Food Systems i.e. the topics in the image (screenshot 8) are mostly well understandable in the context of Food Systems.

Topics

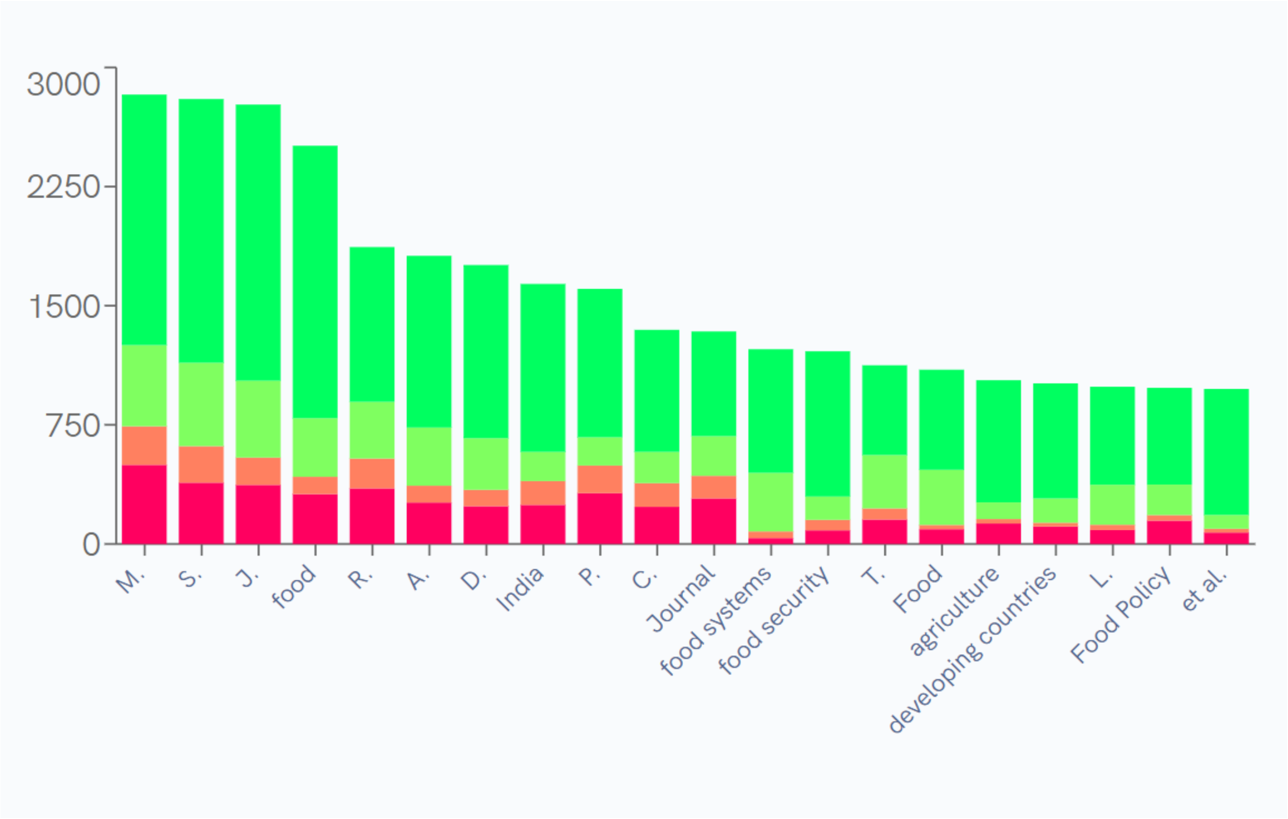


Screenshot 8: The polarity Score of the Curated Reading List sources for Food Systems

[3] GEneral Multilingual Environmental Thesaurus, https://www.eionet.europa.eu/gemet/en/about/

Keywords

As noted from screenshot 9, the keywords are strange and meaningless and do not reflect the Food System. Screenshot 9 is taken before “cleaning” the model. The model will require trimming and cleaning which is partly manual work made by PoliRural ICT- expert. The work will be executed at time which will be dependent on the expert's other tasks.

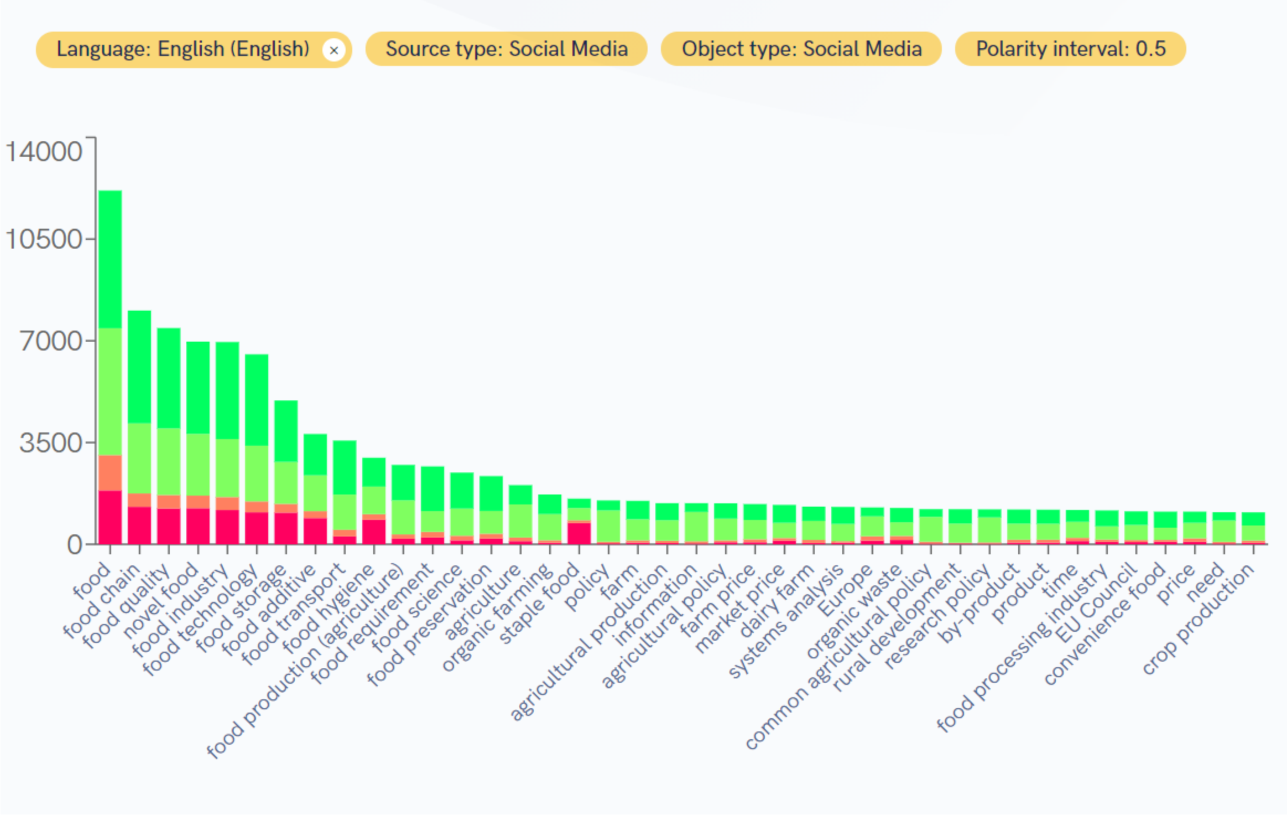


Screenshot 9: Keyword based polarity score

Social Media analysis and Kibana

The Social Media section displays the results of continuous streaming of messages from Twitter with information related to topics. More details about the Social Media section are presented in SemEx Hand-on-Manual, Chapter 8 [2].

The below screenshots 10-12 are the result of the search made by using following terms: the search word - food system, time period - last 1 year, language - English

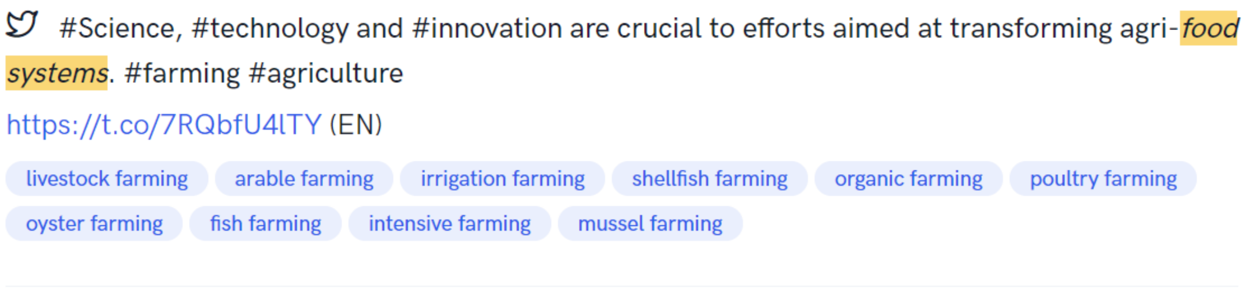


Language: English (English) × | Source type: Social Media | Object type: Social Media | Polarity interval: 0.5

Screenshot 10: A polarity ranges based on search at Social Media: food systems, all, english

When comparing Polarity Scores from added sources (screenshot 8) to Polarity Scores from Tweets (screenshot 10) we can notice that certain topics are the same, but there are also differences which originate from the used different sources. Also, the semantic analysis result (positive/negative issues) is different because of the used source.

By clicking one bar you can review the sources i.e. Tweets and the topics which are linked to the tweet. The result is in the screenshot 11.



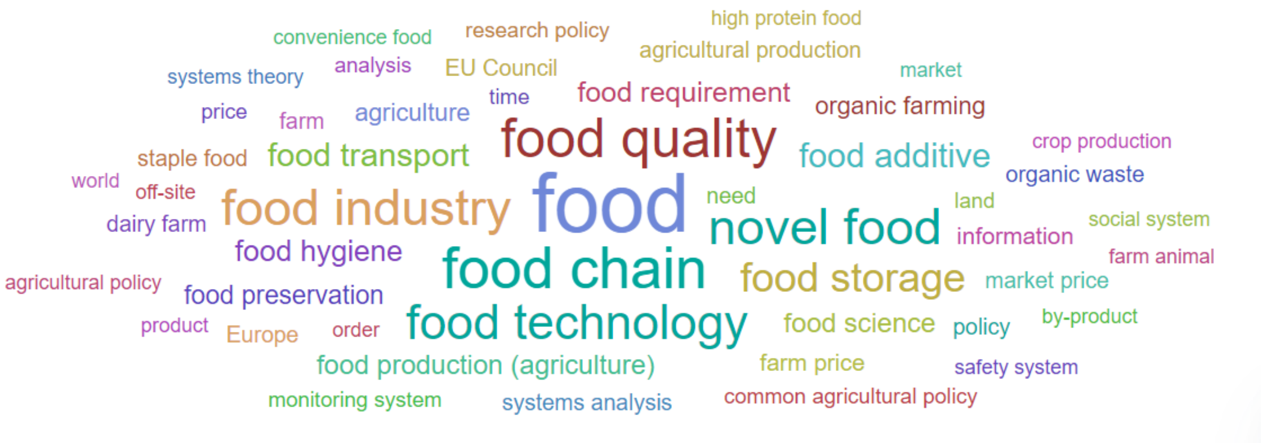
#Science, #technology and #innovation are crucial to efforts aimed at transforming agri-*food systems*. #farming #agriculture

https://t.co/7RQbfU4lTY (EN)

livestock farming | arable farming | irrigation farming | shellfish farming | organic farming | poultry farming | oyster farming | fish farming | intensive farming | mussel farming

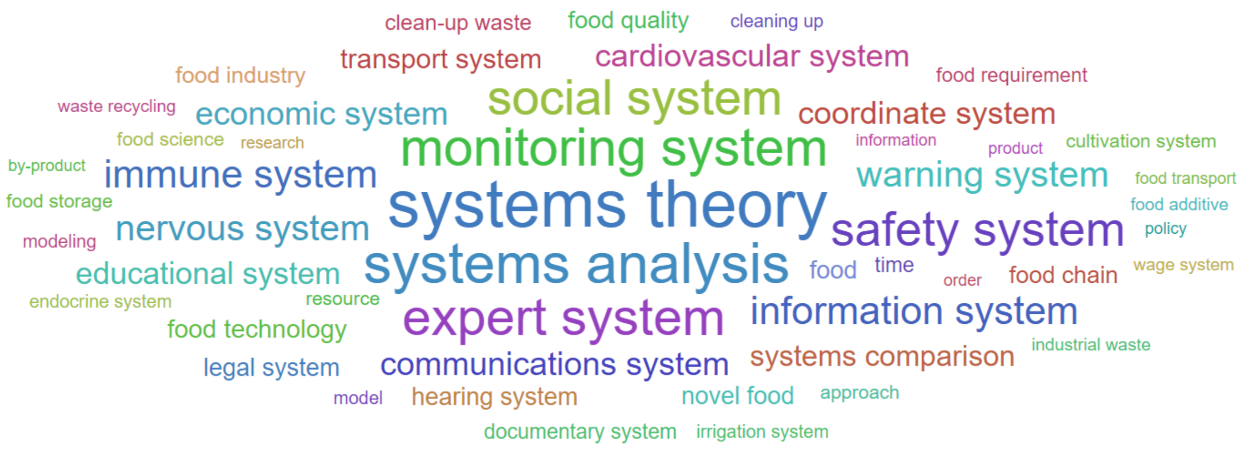
Screenshot 11: One example of a tweet that is counted into bar "organic farming".

The screenshot 12 is a word cloud from KIBANA. It demonstrates the search term i.e food systems with 67469 documents which covers both tweets and paragraphs of the sources.

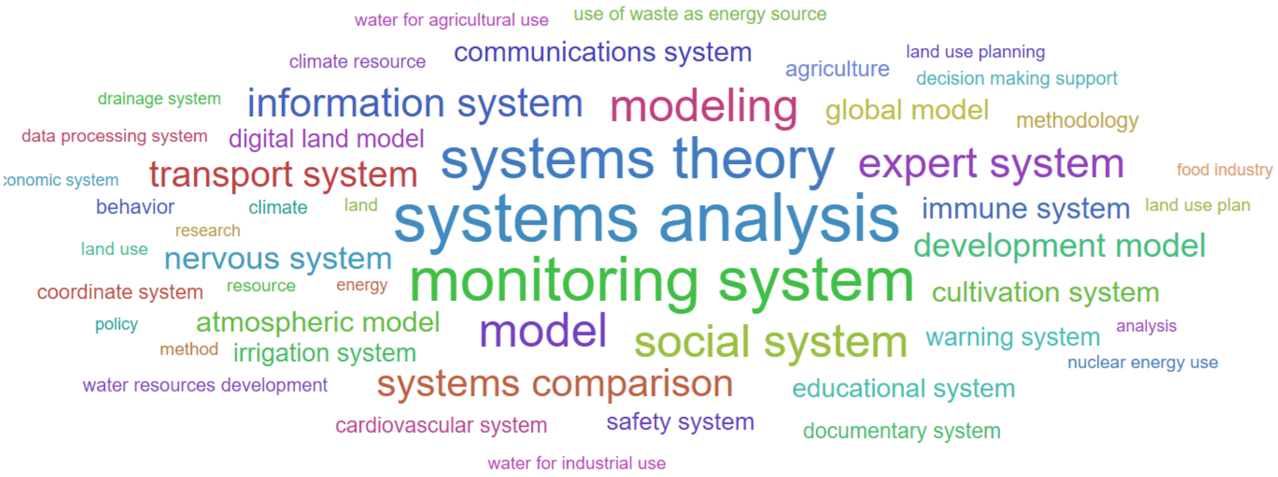


Screenshot 12: Word Cloud at KIBANA: food systems, last 1 year, english - 67469 sources.

The next two word clouds are also a visualization made at Kibana. The word cloud for the exact term "Systems Thinking" was structured from only 2891 sources, which were sources that had been added manually into the library (screenshot 13). For Kibana the sources are split into several paragraphs and that's why the number differs from the SemEx analysis that was nearly 600 sources. The latter world cloud (screenshot 14) prints visualization of the "System modelling" based only on 146 sources. Despite the limited number of sources, it's still interesting to test different keywords and topics and assess the results, which include oddities like "cardiovascular system".



Screenshot 13: Word Cloud at KIBANA: "Systems thinking", last 1 year, english - 2891 sources



Screenshot 14: Word Cloud at KIBANA: “System modelling”, last 1 year, english - 146 sources

Observations and conclusions

SemEx is a pilot application, the purpose of which is to demonstrate possibilities to analyze unstructured data that is in text format.

The application uses sources that are manually added to the library. The manual adding of sources is one limitation that may direct the usage of SemEx towards narrowly scoped topics.

The approach to create a Curated Reading List (CRL) for the new sources addressed on Food Systems was technically easy for the user, but it required extra manual work from the ICT experts. This mandatory intervention of the ICT expert is also a limitation.

The Twitter tweets as a source is a technically interesting application. It shows how an interesting topic can produce tweets and comments in the social media community. The knowledge value and reliability of the tweets may be so low that the usage of tweets is not possible. However, tweets surely tell something about our time.

The user’s minimum goal was to get visual images of the Food System: the elements and their connections. The analysis of Topic Explorer, Polarity Scores, and Social Media (incl. KIBANA) provided the main images. However, the images are only the tip of the iceberg. The more important result of the experiment is the Curated Reading List and the possibilities that it provides e.g. drill down into one source, group sources that touch the same issue. The Curated Reading List can provide the tools for the researchers and facilitators who are seeking and collecting knowledge on a certain topic. The continuous update and expansion of the Curated Reading List will rule its usability and usefulness.

The main output of the experiment was the learning. It forced the user and experimenter to concentrate on the topics of Food Systems, seek new sources and above all to play with SemEx to understand how the new sources operate in the application.

References

[1] D2.3 Final Text Mining Solution, PoliRural, 2020, available at
https://polirural.eu/wp-content/uploads/2020/08/D2.3_final.pdf, accessed 13.5.2021

[2] Semantic Explorer user’s guide, PoliRural, 2021, available at
https://polirural.eu/wp-content/uploads/2021/04/Hands-on-Manual_spi.pdf, accessed 13.5.2021.