



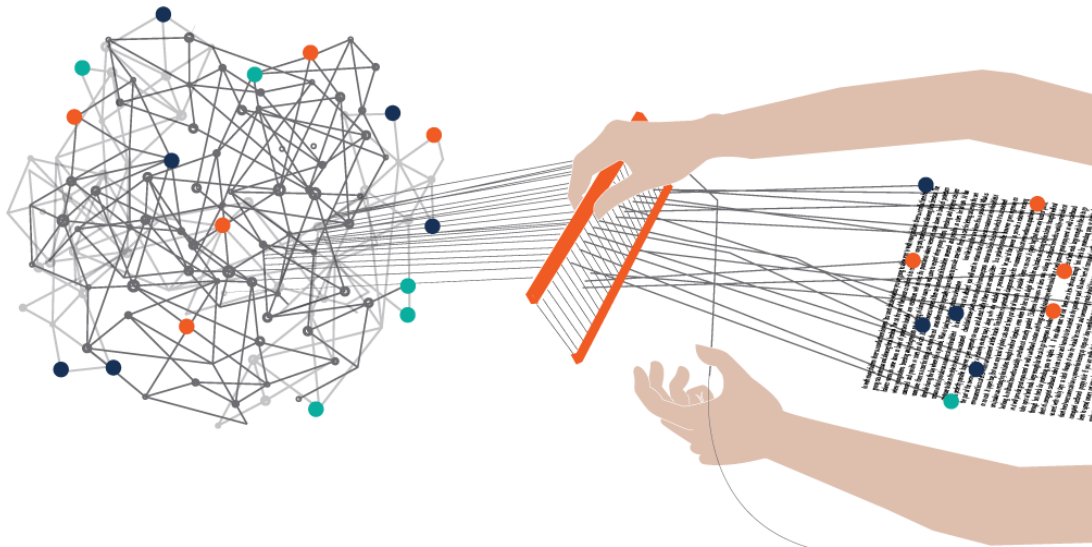
SEMEX - SEMANTIC

Polirural Seminar - Latvia
EXPLORER



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 818496. This document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

Text Mining and Semantic Explorer in Polirural

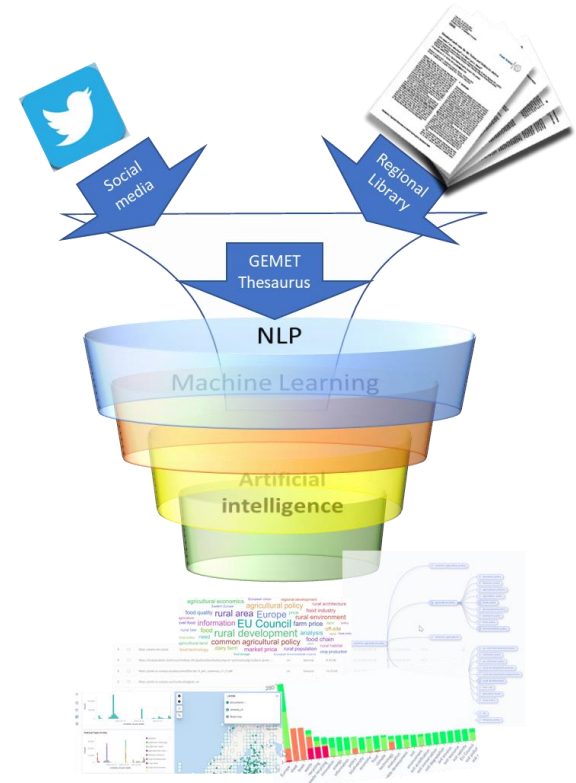


Plan of the presentation

- What is text mining
- What is Semantic Explorer (Semex)
- Semex development
- Main components
- Results and practical use
- Current and future developments
- Your inputs

What is text mining

- Natural Language Processing (NLP) transforms **text into numbers** so that **computing power can analyse data** for us.
- Complex algorithms, Artificial Intelligence and Machine Learning to **detect patterns in vast amounts of text**
- Text mining can:
 - Evaluate the **sentiment** of discussions in social media about a certain policy
 - **Summarize long texts**
 - **Find patterns and dependencies**
 - **Categorize text by locations and/or by topic**
 - Provides **straightforward graphical results**



www.Semex.io - the dev team in Polirural



Denis Kolokol



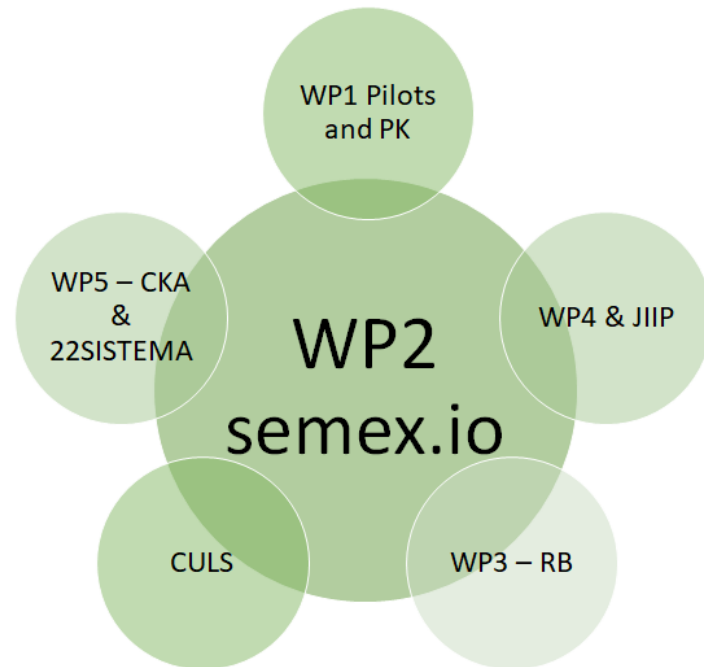
Jan Mazanec



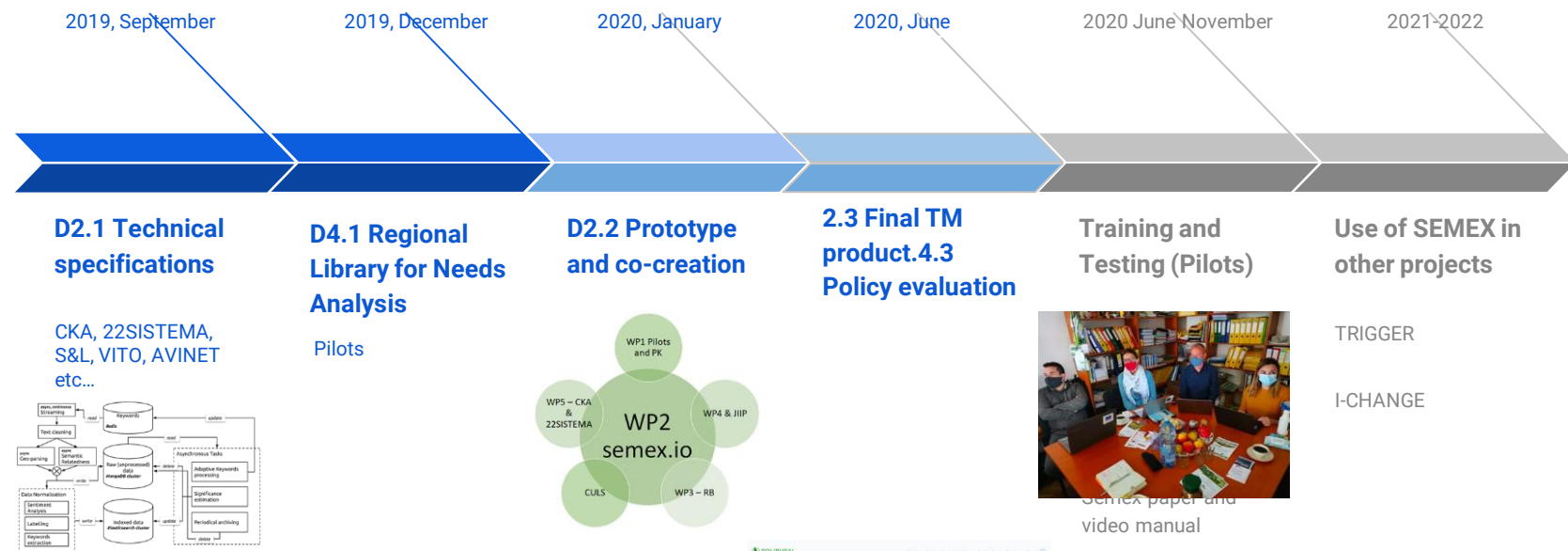
Milan Kalas



Tommaso Sabbatini

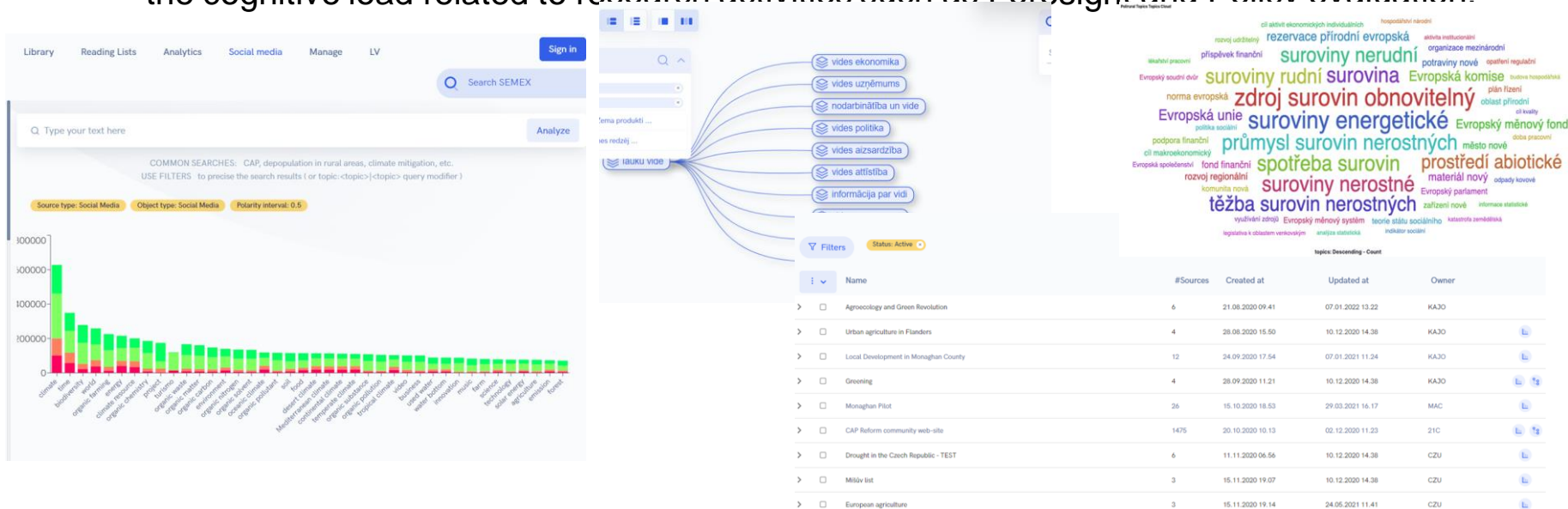


Semex development timeline



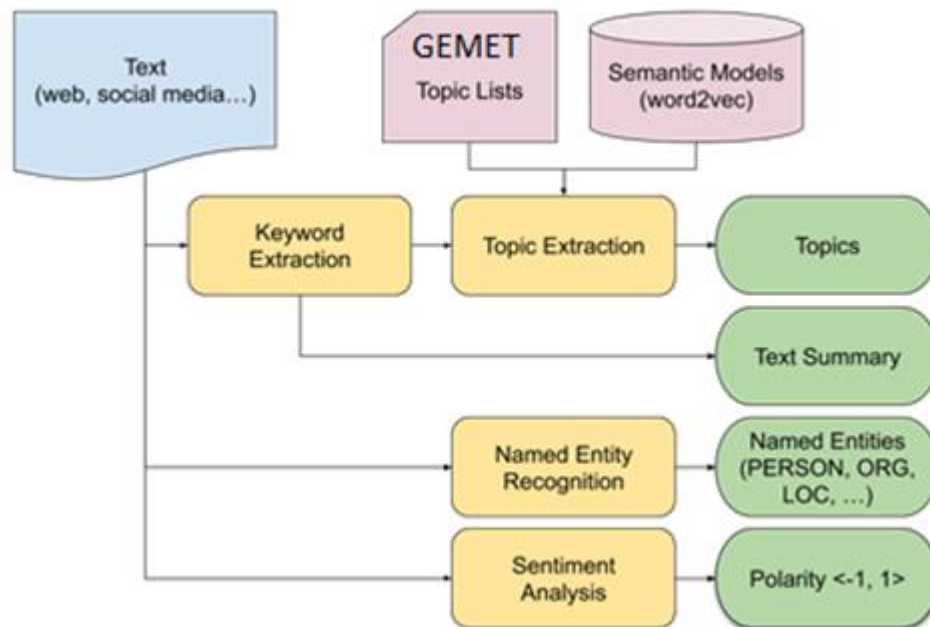
What is Semantic Explorer - Semex.io

- Semantic Explorer (www.SEMEX.io): technical product from [Polirural](#). The tool has been developed to provide inputs to analysts and researchers supporting policy makers, by reducing the cognitive load related to research activities such as Foresight and Policy evaluation.



Semex - System workflow and core functionalities

- **Text in 10 languages:**
 - **regional library** (5600 articles about rural development in 12 regions)
 - Continuously crawled **Tweets (48M)** about rural development in 12 regions in 10 languages
- The system extracts keywords and compares them (word2vec algorithm), to the topics from the General European Multilingual Environmental Thesaurus (**GEMET**)
- The main results are:
 - **Topics**
 - **Text summary**
 - **Named entities**
 - **Sentiment analysis/Polarity**



Semex - main results from library

- Summary
- Topics
- Keywords
- Sentiment analysis
- Named entities

The screenshot displays the Semex interface with several filter categories and their corresponding results:

- FAC**: Gaitan, CIAP
- GPE**: France, Italy, Belgium, Brussels, Rome, Romania
- LOC**: Europe, Northern France, the Franches Terres
- NORP**: European, French, Belgian, D.
- ORG**: CIAP, Europe's, Anselm, Kindling Trust, The European Parliament, Ama Terra, University of Girona

Search results include:

- Cavicchioli (S3, D., Bertoni, D., Tesser, F., & Gianfranco Friso, D. (2015))
- Bertoni (D., Tesser, F., & Gianfranco Friso, D. (2015))
- Tesser (F., & Gianfranco Friso, D. (2015))
- Spain (Maria Diaz, Laia Batalla, Vanesa Freixa Rurbans) United Kingdom
- Moss the Pyrénées Orientales
- Llissui (the "mountain rights", i.e. the rights that allow you to use the mountain comm
- Italy (i, in this publication.) Loire Atlantique Spain El Viver
- UK (Kindling Trust FarmStart) Strasbourg
- the Next Generation of Farmers
- D. (Tesser, F., & Gianfranco Friso, D. (2015)) D. (2015) Somme
- Graines de Paysans (www.grainesdepaysans.be/fr/bienvenue/) Basque
- Kindling Trust
- The European Parliament Ama Terra University of Girona

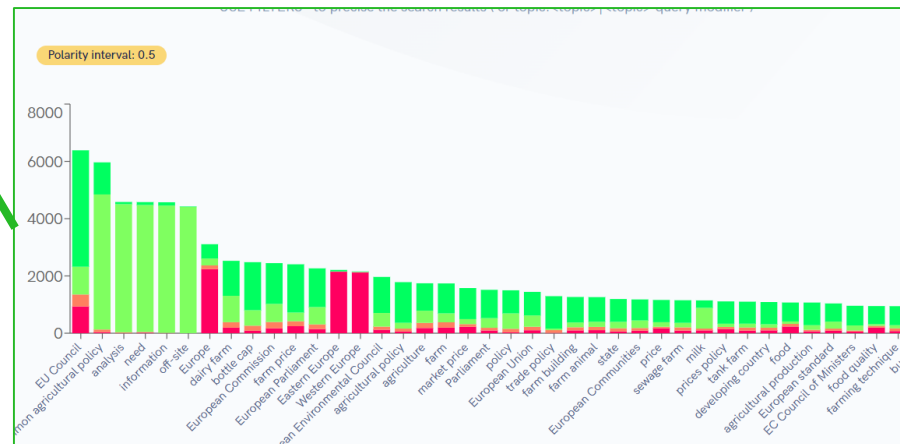
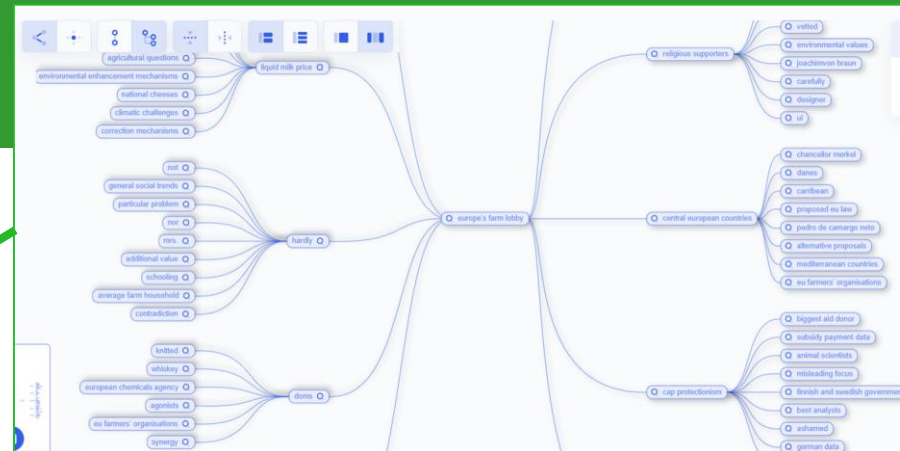
The screenshot shows a detailed analysis of a text snippet: "I turned down a career in the City to become a pig farmer - and I've never been happier". The interface includes several analysis modules:

- Switch-on/off visualisation**: A toggle switch to control the visibility of different analysis modules.
- Most-recurrent-topics**: A list of topics related to the text, such as "livestock breeding", "livestock farming", "school", "vocational training", and "research".
- Most-recurrent-entities**: A list of named entities, including "GPE", "LOC", "PERSON", and "ORG".
- Most-recurrent-keywords**: A list of keywords, including "James", "James Wright", "Nanning", "James' immediate concern", "James", and "James'".
- Article's general-sentiment-analysis**: A sentiment score of 0.474.
- Paragraph's geolocation**: A map showing the location of the text, with markers for "London" and "Chester".
- Paragraph's sentiment**: A sentiment score of 0.474.
- Paragraphs analysed**: A list of paragraphs that have been analyzed.



Semex - Curated reading lists

- Curated Reading Lists is a collection of sources about a specific area of interest
- Aggregated analysis of Summary, Topics, NER, Keywords, Wordcount and extracted URLs.
- Topic explorer diagram
- Possibility of analyzing the most frequent topics or keywords on the polarity score histogram
- CRL has been tested in Foresight exercise providing summaries for more than 200 texts about COVID in less than one working day. The quality of summaries was of high quality. Paper published by Patrick Crehan describing the experiment




Semex - results from Social Media

- Extracted Tweets:
can give an idea
about people's
sentiment about a
certain policy or topic




Results and use in Polirural

- **Creating a text mining tool for 10 languages in 12 months** was an exploit.
- Internal development of specific **crawlers** for big data and for various types of files (HTML, Word, PDF etc...)
- **Use of text mining** in:
 - **Policy evaluation**, tested
 - **Foresight**, tested (CRL)
 - **Rural topics** tested
- Development of a framework for the use of text mining in research projects: using it in 2 new Horizon projects.

 30
Organizations

 20
Regions

 12
Languages

 5600
Documents in Library

 41564326
Tweets

Current and future developments of SEMEX

- SEMEX is been used in the Horizon project **I-CHANGE** to support the analysis on European and **local perspectives on citizen science (social media)**.
- SEMEX will be used to **analyse EU legislation in the climate/health domain (Trigger - Horizon project)**
- The solutions are being fine tuned in order to focus on future users. We are also exploring new directions for more meticulous results.
- SEMEX will remain available for Polirural use for at least 24 months.
- Access to the restricted areas (CRL) for external users is granted on a case by case basis. In case of interest please send me an email at: tommaso.sabbatini@kajoservices.com

Questions?

