D2.1 Text Mining Technical Specifications

Project Acronym:	PoliRural	
Project title:	Future Oriented Collaborative Policy Development for Rural Areas and People	
Grant Agreement No.	818496	
Website:	www.PoliRural.eu	
Contact:	info@PoliRural.eu	
Version:	1.4	
Date:	30 May 2021	
Responsible Partner:	KAJO	
Contributing Partners:	Stein Runar BERGHEIM (AVINET); Karel CHARVAT (CCSS); Patrick CREHAN (CKA); Karel PANEK (NUVIT); Otakar CERBA (Plan4all); Antoni OLIVA QUESADA (22SISTEMA)	
Reviewers:		
Dissemination Level:	Public	X
	Confidential - only consortium members and European Commission Services	
Keywords:	Text mining, foresight, documents, social media, curated reading list.	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 818496

Revision History

Revision no.	Date	Author	Organization	Description
1.1	20/09/2019	Anne GOBIN	VITO	General review
1.2	23/09/2019	Michal STOČES	CULS	General review
1.3	29/03/2021	Tommaso Sabbatini	KAJO	Updates of the text based on the monitors' comments
1.4	30/05/2021	Milos Ulman	CULS	Final formatting and submission

Table of Contents

List of Figures

Executive Summary

This deliverable describes the role of text mining in the PoliRural project and the technical specifications of Semantic Explorer - the main tool for supporting pilots in the processes of needs gathering and policy assessment. The core semantic model will be built and "trained" on the corpus of EU texts, before being adapted to regional contexts and languages. After several iterations, fine-tuning and testing, the tool will be able to accurately extract the required information from the pre-compiled regional libraries, as well as process survey data gathered at the later stages of the Project. The aim is to create a set of services providing inputs for foresight exercises to assist in the tasks of horizon scanning, needs and challenges analysis, etc. Text Mining (TM) solutions should also serve as a tool to build regional profiles for the process of System Dynamics (SD). In addition to that, the system is intended to collect data from Social Media, in particular opinions and their relationship to various aspects of policies. This will allow for comparing the analysis with quantitative data, obtained through surveys, and to provide "one version of the truth" through interactive visuals.

1 Introduction

Rural areas make up 44% to 80% of every European Union (EU)'s country (EC, 2019). These diverse territories are home of wide range of endemic animal and plant species, marvellous wildlife and rich biodiversity and constitute part of the European heritage. Rural areas are of particular importance with respect to the agrifood and Environmental sectors and should be specifically addressed within this scope. Europe's diversity of landscapes is the product of intense human intervention over millennia. Agriculture is the main land user and the resulting high visibility leads to a widespread perception that "rural" matches with "farming". EU's Common Agricultural Policy (CAP) stresses the importance of preserving the farmed landscape as traditional agricultural landscapes form part of the cultural and natural heritage, the ecological integrity and the scenic value of landscapes make rural areas attractive for the establishment of enterprises, for places to live, for tourism and recreational businesses. The ecological integrity of a landscape is an important element of its attractiveness and perceived value. Rural areas should not only be seen as a territory for food production since farming is shaping European nature and, increasingly, is taking the role of upholder of European cultural heritage (Daugstad et all, 2006).

However, European rural areas are facing new challenges. New generations prefer to move to cities for better-paid jobs. Consequently, services such as schools, hospitals and public transport decrease and local economies suffer. The result of this vicious circle is that, without policymakers' intervention, rural areas become progressively less attractive for younger generations and local economies suffer (Margaras, 2019). Secondly, climate change is forcing farmers to deal with the higher probability of extreme weather events such as droughts, flash floods and heat waves. Finally, a more difficult access to new technologies, such as high-speed Internet as well as technological innovation in general, is increasing the so called digital divide between rural areas and cities (Eurostat, 2017).

One of the objectives of PoliRural project is specifically to bring solutions to policy-makers in order to support European rural areas in responding to contemporary challenges. Decision makers have the ability to steer change and in so doing reduce the negative impact. However, this requires advanced knowledge of how a particular action, or inaction, will affect people and places, at present and in the future. In order to explore the solutions an analysis of the common situation is needed and text mining can support stakeholders, researchers and expert in this assessment. In fact, most of the world's information today is textual, included in large libraries and archives, scientific papers, Internet with its websites and social media (Hradec et all., 2019) and text mining technology is able to "process textual data in a largely automated manner", extracting and analysing for patterns and dependencies (Kayser, 2016) with the ability of discovering new knowledge at a low cost.

While in this deliverable the first part is dedicated to defining the role of text mining in PoliRural (Task 2.1), its main objective is to outline the technical specification of the text mining technology that will be used in the PoliRural project. After this introductory part, in the second chapter it will be clarified how Work Package 2 (WP2) will interact with the Foresight process (WP4 and WP6) during policy evaluation (Task 4.5) (section 2.1) and needs gathering (Task 4.3) (section 2.3), for example, by providing curated reading lists, topics and subjects categorized by priority and relevance. Furthermore, in section 2.4 it will be explained how Text Mining and System Dynamics (WP3 and WP5) will work in synergy in order to understand the current situation of rural areas,

but also to analyse how the parameters evolve over time, for example. Moreover, section 2.5 will introduce the role of text mining in the PoliRural Digital Innovation Hub (DIH) (WP3), platform that will manage big datasets in the project.

Chapters 3 to 8 are of technical character since they mainly describe the technical specification of text mining in PoliRural. The third chapter will explain which text mining solutions will be provided in PoliRural, while in the fourth chapter the authors will outline the sources that will be selected for extracting information. The fifth chapter will present the technical characters of the Semantic Explorer, a tool that will be developed in order to use text mining in the PoliRural project. In the sixth chapter the authors will clarify the interactive visual outputs of the text mining tool and in the seventh chapter there will be an outline of the infrastructure foreseen needed for the well-functioning of the tool. Chapter eight underlines the challenges and list the possible solutions so that text mining brings valuable results into the PoliRural project.

In the concluding chapter the authors attempt to sum up the whole deliverable and to shed light on the future steps and developments of the Text Mining tool and of Work Package 2 in general, within the PoliRural project.

Finally, this document serves as the Developer's Guide for the creation of the deliverables D2.2 "Prototype Text Mining Solution and D2.2 "Final Text Mining Solution". Both D2.2 and D2.3 are going to be products of Machine Learning (ML) and Natural Languages Processing (NLP). However, the models and techniques to be used in the creation of those products are not described in detail in this document, due to the experimental character of the process of their selection, training and evaluation of the results. Maximum attention is given to the statements of the end-users needs and specifications of the output of the system that should cover those needs. The exact models ML and NLP and detailed description on their implementation will be carefully selected and tested during the process of development. They will be described in details in D2.2 ‘Final Report’.

2 The role of Text Mining in the PoliRural Project

This chapter defines the role of text mining in the foresight process, taking into account the project vision and framework requirements. It clarifies how text mining can support the initial research part of policy development. It also specifies why text mining can be a valuable tool to support the foresight process (WP4 and WP6) that will be used, for example, in workshop-based meetings with the Pilots for analysing the current status of rural areas. Moreover, the collaboration with System Dynamics (WP3 and WP5) shall enhance the understanding of the common situation in order to better evaluate the challenges ahead. Finally, the interaction with the Digital Innovation Hub (WP3) tool, which has the objective of managing a large amount of data sets, providing locations in text document (geoparsing) and link the information contained in text with Linked Open Data, will be outlined.

2.1 The text mining technology basics in PoliRural

PoliRural's Core Objective 2 is to "measure prevailing attitudes toward rural policies among regional stakeholders by combining survey research, usage of existing structured data with innovative text mining techniques that are both rigorous (can produce accurate insights) and versatile (can be used with multiple online sources)" (PoliRural, 2019). In general terms, the main goal of text mining is to help process vast amounts of information from structured and unstructured sources and discover new knowledge at a low cost. The application of text mining techniques to online content found in forums or social media can provide quantification of rural attractiveness and different pressures on landscape, landscape planning, rural development planning, scenarios matching qualitative and quantitative indicators. The benefits of data analysis include better, quicker and more efficient decisions based on information which is increasingly produced in near real-time. Much information that policymakers need to make informed decisions is hidden in large amounts of textual data. It is reported that only about 20% of the data available online is numerical, with the rest (over 80%) stored as text (Giraldi, 2017). Therefore, structured and unstructured texts remain the largest, readily available source of information. Structured data has a high degree of organisation, which can be ensured through a relational database that is readily searchable by simple algorithms or other search operations. By contrast, unstructured data cannot be so easily organised, as it includes web pages, PDF files, PowerPoint presentations, blog entries, wiki pages and word documents. The sheer volume of textual data makes it practically impossible for any public policy team to perform a meaningful analysis based on human effort alone. Text mining, for all its limitations, can support this effort by offering several advantages. Besides speed they include better categorisation and reduction of large amounts of data, structured or otherwise, and better understanding of important points and interrelationships through visualisation.

The tool itself will be based on heavy-duty knowledge extraction using deep neural networks trained on the large corpus of texts (e.g. EU documents, scientific journals) and adapted to work with regional libraries and languages. One of the outputs that this text mining process produces is a semantic tree which can be explored interactively on the PoliRural platform (Digital Innovation Hub)[1]. PoliRural text mining solution shall comprise several interrelated components. Chief

[1] ANNOY & HDBSCAN will be used for clustering & novel Word Mover's Distance for sentence/paragraph similarity analysis

among them are web crawlers for scraping textual information from online sources using different protocols. Currently, these crawlers collect information from the European Publication Office, Bookshop, EURLEX, CORDIS, DG JRC PUBSY via a number of interfaces (SPARQL, SOAP, OAI, FTP, HTTP) with the help of bespoke harvesters.

2.2 High-level use cases in the foresight and policy development life-cycle

At first glance, introducing text-mining into the tool-box of policy developers might seem like trying to run before learning to walk. Policy development, particularly on regional level, has traditionally been a manual discipline where data are considered and assessed by professionals who use their subjective skills and experience along with factual information to build strategies and policies.

That being said, upon closer inspection, it is how text mining could be a positive contribution in both the fact-finding analysis, design, implementation, monitoring and evaluation stages of a policy development process is not difficult.

Below, common questions that policy developers might ask themselves, are discussed and related to the generic capabilities of the PoliRural text mining system. This explanation is the fruit of the discussion between text mining and policy making experts and has been over-streamlined to allow text mining developers to have a clearer picture of the general process to develop a focused tool.

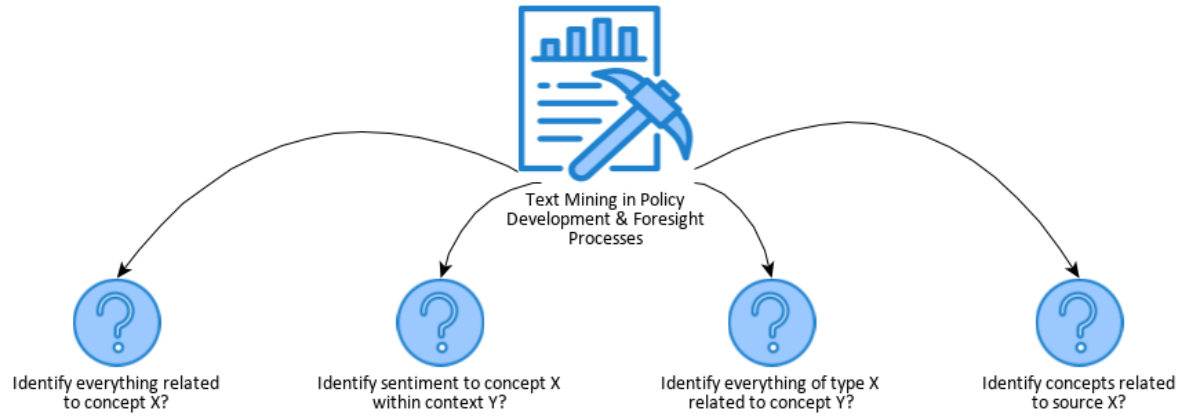

Figure 1 High-level use cases in the foresight and policy development life-cycle

2.2.1 How can I learn something about what I am about to do?

A task that is essential to any policy development process is the collection of information and data that will form a baseline for the work to be done.

A typical question posed by the professional actor could be: “I am developing a policy and am interested in relevant reading material to get into the subject.”

Text mining capabilities provided by PoliRural can be used to generate curated reading lists for topics by invoking methods to extract “everything related to concept X” from the knowledge graph that is described in greater detail further along in this document.

2.2.2 Has someone else done something like this before?

Another near universal step in problem solving is to determine if the problem has in fact been solved *before, by whom* and *where*. For this point the it is extremely important that inputs to the system are generated directly by regional actors. This is the reason why in case of Semantic Explorer the Regional Library is created and curated by local Pilots.

The text mining approach in PoliRural extracts and a rich set of metadata from the corpus of text it processes. That means that subsequent queries can be applied not only to concepts alone but a combination of semantic concepts and spatial distribution.

A typical question posed by a professional actor could be: "I am interested in a particular policy problem and am interested to find other places that (1) have the same problems or (2) where people are concerned about the same things or (3) where the characteristics of the location are similar? By using the tools provided by PoliRural, all these questions may be answered.

2.2.3 What existing policies or policy instruments must I take into account?

A third question that is very common is to determine which explicit professional and legal references must be used as absolute requirements for the work to be carried out.

A typical question posed by a professional actor could be: "I am interested to find all pre-existing policies or policy instruments that concern the same subject".

Once again, by using combinations of the four different high-level queries shown in Figure 1 above, it will be possible to find the answer to this question.

2.2.4 What is the perception of the topics I am about to design a policy for in society?

An area where the manual policy development process has a great potential for improvement is in its relationship to evolving public opinion. Traditional policy processes are linear, take a long time and require a lot of bureaucracy.

The points on these processes where public opinion is taken into account and sounded is usually at the initial fact-finding stage and then again after the policies have been formulated as a part of a public consultation. At that stage a lot of decisions have been made that limits the practical consequence of public interaction and thus, the policy may well be in tune with what was known and understood at the inception stage - but may be at odds with the current perception.

Of course, technology cannot compensate for our human trait of drawing conclusions based on sparse evidence and then changing our minds over time as we learn, understand and form more solid opinions. Technology, text mining and sentiment analysis in particular, can however enable us to keep a finger on the pulse of society throughout the policy development process.

A typical question that a professional actor might want to know the answer to is "I wish to determine the public perception of a concept in order to either develop a policy that corresponds to public sentiment or that seeks to identify need for behavioural change in order to affect a desired impact?"

Sentiment analysis functions included in the text-mining system of PoliRural will be able to assist professionals in sounding out general sentiment to concepts by analysis of social media streams/knowledge graphs.

2.2.5 How effective is a measure and under which conditions?

Having moved on a bit in the policy development process the policy maker will be trying to determine which measures should be stimulated to achieve the objective of the policy. At this stage, it is interesting to be able to identify how a given *measure* has been received and what effect it might have had since it was introduced.

A typical question that a professional actor might ask could be: "I am trying to determine the impact of a particular concept and measure regionally, nationally, globally?"

A combination of sentiment analysis and identification of sources related to a concept and measure may be used to answer this question for the policy development professional.

2.2.6 Does a specific source or collection of sources deal with my issue?

Finally, as is demonstrated by literally all text currently being produced, the present deliverable included, we are using too many words. That means that identifying if a source is relevant to the planning process or not might require a disproportionate investment of reading time that might be entirely fruitless.

A typical question that a professional actor might ask could be: "I am trying to find if *a concept* is mentioned in *this* source?"

This might sound somewhat similar to how search engines work. However, as most of us will have experienced, however relevant a source we know to exist - we may be unable to find it even in our own inbox however much we search.

Semantic searches expand the search terms automatically and are able to retrieve relevant information not only based on letter-by-letter text matches but also on related concepts. Querying the knowledge graph resulting from PoliRural analysis will enable planners to answer this and similar questions.

2.3 Text mining in the foresight process

Understanding of a generic Foresight Process will help the reader grasp how text mining can be applied. Foresight can be defined or described in many ways, but it is essentially a change-management process. It differentiates itself from other strategy processes by its emphasis on stakeholder engagement, the role of collective learning and the co-design of a desired future.

Foresight is about understanding change, how it happens and what causes change. It is about designing the future based on a desired outcome using typical management tools such as

- Vision;
- Roadmap; and,
- Implementation plan.

Foresight requires those taking part to develop an understanding of the change they want to see happen and how to make it happen. For example, change that will add up to greater prosperity for a region as a whole. It requires those involved to understand how change happens, trends must be accepted and what trends can be broken or changed. The process must engage adequately with the agents of change (CEO teams or policy makers and other deciders). It must demonstrate the feasibility of the desired change and the existence of broad-based support for such change.

Ultimately, Foresight is a collective learning process. It must create a space and time for changing minds concerning issues that may have seemed irrelevant or undesirable at the onset, but which gain acceptance as the process moves from early an exploratory phase to a final phase in which concrete recommendations are laid down.

A typical Foresight process moves through phases that include

- An initial mobilisation or design phase led by a small core team.
- An exploratory phase that explores issues and options while expanding the scope of engagement.
- A final phase in which recommendations are agreed and codified, usually in the form of a vision, roadmap and implementation plan.

The project will use the pilot Foresight activities of the PoliRural project as a living laboratory for exploring the role of text mining or TM, and system dynamic modelling or SDM in the context of a regional Foresight activity.

The pilot “model” should also indicate where and how techniques based on TM and SDM might be applied to support the work required in the implementation of the pilot.

There are clear risks involved in the use of TM and SDM. These tools are unproven, and many open questions remain in relation to

- How they can be used and in what tasks of the Foresight initiative;
- The ability of the TM system to operate with different languages (Finnish, Hebrew, Serb, etc.); and,
- The availability of suitable datasets, especially in the application of SDM.

In order to manage these risks and maintain the motivation of those involved in the regional pilots, it is therefore important to ensure the ability of the pilots to advance and produce useful results, independently of progress with the development of the TM and SDM tools.

2.3.1 Potential Application of TM and SDM as Part of the Foresight Process

A well-run Foresight initiative requires a lot of group-work structured debate and other forms of purposeful engagement with stakeholders. The group work, however, requires good preparation based on research and lots of reading. One of the ways of providing guidance to the groups is to provide curated reading lists. This reading can then provide the basis for useful debate leading to the prioritization of challenges, progressively deeper analysis, the drafting of position papers, the development of vision statements and the listing of desirable policy options. The data intensive research techniques employed in Foresight include the following.

- **HORIZON SCANNING (HS)** is an open and exploratory activity intended to understand what is happening elsewhere, in society, in the world of business, in the world of technology, in the world of research and innovation. This is often done in waves, as the orientation and emphasis of the work of Foresight evolves, as interaction with stakeholders reveals new ideas and concepts that were not adequately anticipated or explored at the start. One way to organise the output of a horizon scanning exercise is to create a series of topical lists. It is recommended to carry out a preliminary HS exercise at the earliest stage of each regional pilot, in order to identify relevant trends that "may" impact the region, the drivers and enablers that are reinforcing or mitigating those trends, and the issues and challenges that "may" arise as a result of those trends. Horizon scanning can also be applied to the search for new products or technologies, business practices or economic models, new questions worthy of scientific inquiry, new and emerging thinking about policies and how to address key challenges. Many different tools can be employed to help in this, but it mainly boils down to reading and talking to people who are knowledgeable or experienced. An important tool is Google, or other more specialised search engines. The role of the researcher is very important in selecting experts to talk to and material to read that might be relevant, even if the reason is not immediately obvious. The biggest risk in this work is the risk of missing out on something important. This has to be weighed against the risk of doing too much work and never finishing. Tools to help researchers ensure that their search is wide-enough include techniques based on mnemonics such as **PEST** or **PESTLE**. One can hope that the use of TM will reduce the con git vie load on the research helping them to search more material and identify what is interesting in an increasingly large corpus of available online content.

- **DRIVERS Analysis** to understand how social and economic phenomena are being shaped and to run what-if scenarios to explore policy options. This is an important possible domain of application of SDM. Successful implementation of SDM may require the use of TM to identify data sets and models required as inputs to useful SDM simulations.
- **ISSUES and NEEDS Analysis** based on recognised local challenges, and on the exploration of needs and challenges emerging elsewhere. The search to identify issues and needs should capture well-known or recognised needs that are already mentioned in policy documents and local development discourse, but it should anticipate future local needs based on needs that are known and recognised at a global level or in regions elsewhere.

This requires search and research relating to local and international sources of information, insight, opinion and social commentary. There is great scope for the application of TM techniques to this kind of task.

- **DEEP DIVES** are sense-making activities intended to understand the relevance of phenomena identified through horizon scanning or needs analysis as well as the potential and feasibility of possible solutions. This requires research to identify relevant reports and documentation, as well as experts with deeper understanding of the phenomena under discussion.

- **COMPARATIVE POLICY ANALYSIS,** an exploration of old and new policy options based on formal evaluations or lived experience with policies implemented in the region, other EU members states or elsewhere. The intention being to identify elements of a policy response that might be effective given the context of the region and the constraints faced by its actors.

It is both unreasonable and inefficient to expect everyone involved in a Foresight initiative to do their own research. many of those involved will have little or no ability or experience in this kind of research, and most will simply not have the time. Those who are responsible for the Foresight exercise will need to ensure the availability of relevant knowledge and steer those involved towards suitable reading material. This can be done by the provision of **CURATED READING LISTS** or discussion papers based on the reading of such lists. The quality of such lists will be an important factor in establishing the credibility of the team leading the Foresight initiative and in motivating the participation of stakeholders.

It is not easy to do good research, and this is getting harder over time due to the increase in sources of information and its variable quality. In the sciences this has reached a stage where it is physically impossible for experts in a field to keep up with all of the research published in their area. The same is happening with more anecdotal but nevertheless important sources of information from the press, popular blogs and social media. This is why it is important to explore how tools based on TM and SDM can be usefully integrated into the overall Foresight process.

2.3.2 The Application of Text Mining (TM) in a Generic Foresight (FS) Process

This chapter mirrors the discussions that text mining and foresight experts entailed in order to create a TM tool that could support FS activities. Both TM and FS experts are aware that TM cannot replace human work. However, Polirural provides a good opportunity to experiment the various options to start using TM as a support tool for desk research related to FS. This chapter includes basic explanations about Foresight that may help TM developers to design specific solutions for FS exercises.

2.3.2.1 Horizon Scanning, Where Exploration Starts

Although the initial lists generated by the Horizon Scanning (HS) process may be long, they will be incomplete. New sources of data, information and insight will be discovered as the work proceeds, and as new stakeholders may need to be inspected in successive waves of TM. This will

grow the list and improve it in terms of relevance and in terms of the quality of the accompanying analysis. As the list grows, it may emerge that not all issues have the same level of urgency, that not all issues are actionable at regional or national level. Factors such as these indicate a need to establish priorities. This ranking of priorities should not be done too early in the process, however. The risk is that important priorities may get overlooked or culled due to lack of understanding. It is also possible that priorities of importance for one group may seem unimportant to another. At some point it will be decided that the search has gone on long enough and that there are diminishing returns on further efforts to scan the horizon. now is the best time to carry out a ranking exercise. Care should be taken to involve all major stakeholder groups in the ranking process. Once priorities are established, there may be a need to repeat tasks such as HS and the compiling of CRLs. Each time however with different objectives, a better focus on the priorities and a clearer sense of what is relevant or important for the region.

2.3.1.1 Drivers Analysis and the Language of Change

One of the activities that people typically do in a foresight initiative is to examine “drivers of change” using a structured brainstorming approach supported by reading, research and inputs from local experts. The goal of such an exercise is typically to get a good feeling for what will shape the future (in terms of employment, prosperity and a rural renaissance etc.)

Identifying drivers of change is based on reading and talking to experts. It often includes brainstorming phenomena identified on the basis of Horizon scanning, with a view to asking if they are relevant for the Foresight pilot and how they are they likely to manifest at regional level.

Both TM and SDM can be useful in this process, playing complementary roles and reduce the burden of work on those charged with the research needed in all 3 phases of each pilot. TM could help in identifying drivers and enablers, whereas SDM could be used to explore the dynamics of how they interact.

2.3.1.2 Text mining to support the creation of curated reading lists

Much of the work that is required in the early stages of a FS exercise is of an exploratory nature and can be described as research to understand how change happens, what is driving the changes that will affect us (the region that is the subject of the FS exercise). Text mining can be of great help in this. In particular as a tool to support the work of the foresight experts supporting the foresight process. The text mining could help by supporting the development of «curated reading lists». Providing cues to the Foresight team, identifying issues that may be of relevance for the region and the foresight process.

2.3.1.3 Iterative use of the Text Mining tool in Sense Making activities and Deep Dives

One of the features of Foresight is that it starts out being open and exploratory in nature, looking at many possibilities. The process suspends judgement on issues, until later ... the goal being to help people expand their ideas about the world and how it works, or could work at regional level, often based on examples from elsewhere. This broad based, exploratory and open-minded approach is usually called “horizon scanning.”

The curated reading lists will provide a basis for communication with stakeholders as well as for workshop-based activities that can be described as “sense making” activities. The sense making activities may involve list-ranking activities to establish priority or relevance lists, also 2-dimensional BCG style box ranking, or even more complex multi-criteria decision analysis.

The sense making activities may involve list-ranking activities to establish priority or relevance lists, also 2-dimensional BCG style box ranking, or even more complex multi-criteria decision analysis. The sense-making activities may involve further exploration but on more specific issues, using for example deep-dive workshops. For example, horizon scanning may identify climate change as an important general issue. If it is retained as a priority area for further study, then a deep dive may be required to understand how change will play out or is already playing out in the region that is the subject of the FS activity. A second wave of text mining work may be required to dig out a second curated reading list that is more clearly relevant for the region, for example on flooding or volatility of agricultural yields etc.

Thus, in this way, the TM work is an iterative process that starts general and progressively digs out context to inform discussion that are increasingly region-specific and relevant for the FS process.

Ideally it should be possible to use TM as part of an iterative process that starts at a high level with general concepts, and with guidance from the researcher progressively digs deeper identifying content that will better inform the participants of the sense-making workshops, with information that is of increasing relevance for the region and for the challenge being addressed.

2.3.1.4 Use of Text Mining (TM) and System Dynamics Modelling (SDM) to Support Policy Evaluation (PE)

Policy evaluation (PE) is already done on an industrial scale, but it is difficult to understand how it actually helps in the context of Foresight (FS). However, a form of policy evaluation could be of considerable help in FS, and TM could play an important role in this context. For example, TM could be used (in principle) to identify other regions, that have confronted similar challenges to the current one, but in the past, so it is possible to look at this, what was done and the impact it had, as an input to a discussion on policy options. Thus, text mining could provide a useful basis for exploring policy options for the region that is the subject of the FS activity. It could help provide «curated reading lists» to support the discussion on policy options, based on the experience of regions that have confronted similar challenges in the past.

Will it one day be possible to look at historical data of a region, and run scenarios intended to explore the impact of policy choices, for example based on narratives structured as follows:

- Our region is here today…
- The following regions were in a similar situation 10 years ago…
- Look what happened to them based on data from 10 years ago until today…
- Let’s explore the factors behind their evolution, the decisions they made, the policies they tried to implement

In this way SD can also become a tool for policy evaluation.

Annex I provides examples of additional use cases expressed by the final users that go beyond the scope of these technical specifications, but will be considered case-by-case depending on the feasibility of the option. These examples, however, provide a clearer understanding of how text mining can be used in the foresight process.

2.4 Text Mining and System Dynamics

In the PoliRural project, Text Mining and System Dynamics (WP3 and WP5) will work in close collaboration in order to analyse more in-depth the global reach and the challenges ahead. If the text mining can help to identify significant parameters of change, for example related to population dynamics, value creation, employment, cost of living, quality of life etc. the system dynamic modelling can help people understand the ways in which they interact and evolve over time, and how they can be manipulated by policy intervention and community action. The idea is not to make predictions but to help people involved in the foresight exercise to gain a more in-depth qualitative understanding of how the policy works, what can happen and how they can influence the future. The objective of this section is therefore to start defining the products coming from text mining (TM) that may allow the design of system dynamics (SD) models.

It is important to understand the normal course of action when modelling with local agents or communities, so that everyone may see the global reach and the challenges ahead. The objective of system dynamics is firstly to help understanding the current situation as a trend or behaviour over time, produced by the interactions of a given structure. The model tries to identify the structure, so that past and current realities are explained. Once the model is defined and agreed, policy exploration can be built driven by sensitivity analysis and scenario design.

The extent to which direct human intervention (via field work, interviews, workshops, etc.) may be replaced or supported is something to be adjusted by the ongoing process of the project.

When dealing with local communities the preliminary step will be gathering all the information (coming from literature, expert sources, workshops, etc.) needed to answer the following general questions:

Which are the main discourses and narratives?
Which are the local drivers or most rapid trends?
Major conflicts and problematic indicators?
What are the main capitals, thresholds and risks?
What is the institutional context (involved in the specific narrative)?

This will lead us to identify trends or behaviours over time, and then the firsts Causal Loops Diagrams previous to the models.

From this approach, one can find a procedure proposal including information sources needed and outputs expected from TM to start designing models for the pilot areas.

2.4.1 Inputs expected from text mining to build System Dynamics models

The inputs expected are divided in three categories:

a) First characterization of the area

b) Issues to be addressed

c) Technical and behavioural parameters

In the following sections, for every case, an explanation including SOURCES and EXPECTED RESULTS will be provided.

2.4.1.1 First characterization of the area

The first characterization is a general framework to approach the pilot area. Depending on this first idea the model will take a different form. One of the clearest features will come from the settlement model. The example of Flanders has to be understood and analysed as an urban region with a rural population very much intertwined. The approach of the population module will adapt to this foundational feature.

Sources

The documents needed to identify this first general characterization are very general documents or sources. In the case of Flanders, the Factsheet on 2014-2020 Rural Development Programme of Flanders (Belgium)[2] is a good example.

In the extract below one can find some key indicators of the urban nature of the area and a proposal of indications to individuate them (marked in green).

> **1. SITUATION AND KEY CHALLENGES**
>
> Flanders is the northern region of Belgium. It covers an area of 13 521 km2 and counts approximately 6.35 million inhabitants. The region has a very high population density (475 inhabitants per km2) which is more than four times the average density of the European Union.
>
> Only 7% of the area is rural and 2.5% of the population lives in the rural area. The Flemish countryside is highly urbanised. It has a very fragmented landscape with strong links between countryside and cities. From the geographical, functional and cultural points of view, rural and urban areas are increasingly interlinked.

These are general, often introductory documents or chapters of a more specific subject. Some other links giving answer to the general characterization are *Wikipedia definition of agriculture in Flanders*[3] and *Prospects and challenges for agricultural diversification in a peri-urban region (Flanders – Belgium)*[4] in chapter number 2 titled *The rural area and agriculture in the Flemish region*.

> The Flemish countryside is highly urbanized. It is not only characterized by a high population density and a highly fragmented landscape, quite often the relationship between the countryside and the urban environment is also very strong. The

2 https://ec.europa.eu/agriculture/sites/agriculture/files/rural-development-2014-2020/country-files/be/factsheet-flanders_en.pdf

3 https://en.wikipedia.org/wiki/Agriculture_in_Flanders

4 https://lv.vlaanderen.be/sites/default/files/attachments/prospects_and_challenges_for_agricultural_diversification_in_a_peri-urban_region_flanders_-_belgium.pdf

> countryside and the urban environment are becoming increasingly intertwined in a geographical, practical and cultural way. The present communication and transport facilities favour economic, social and cultural interaction. As a result, differences are shrinking: the urban environment is an attraction pole for employment, services, education and entertainment. But city dwellers rely on the surrounding countryside for ecosystem services, such as a green area, calmness and recreation. As far as food production, water resources, energy and biodiversity are concerned, rural areas provide important services to the whole society, including the urban environment. Because of the strong linkage between the countryside and the urban environment, dynamics in the rural area are partially determined by the urban environment.

Some other sources include text like this:

> Flanders can be portrayed as a peri-urban area, in which agriculture is still a significant economic sector.

In these texts we are searching for structures like

`<name(s) of the region or synonyms like the XXX region, landscape, countryside… > IS, HAS, IS CHARACTERIZED BY, PRESENTS, CAN BE PORTRAYED AS…`

> Titles are also a clue, like in the case of *The rural area and agriculture in the Flemish region*.

`... IN (THE) <XXX name of the region> KEY INTRODUCTORY WORDS like Description, Introduction`

Figures and numbers are another source of quantified information of the first characterization, with a structure similar to

`FIGURE of the region/area/population/…`

Expected Results

What we expect from this first analysis is a **literal definition of the region**, that is going to allow us to work on a typified approach to the modelling exercise.

The definition may be completed both with a quantitative analysis coming from the texts (figures included in the text) but also a quantitative analysis of the text itself.

Questions

To study the possibility for the pilot areas to give some kind of hierarchy of the sources, so that the results could be analysed differently. This could be a valid hierarchy for the sources:

1. General introductory text
2. Specific (addressing one topic or keyword: agriculture, landscape, population, infrastructures…)
3. Super specific (addressing subtopic levels: organic agriculture, commuting, climate change effect on agriculture…)

Considering 2 and 3 types may also include some introductory chapters (type 1).

2.4.1.2 Issues to be addressed

Once we have a first characterization of the region, we need to fix the boundaries of the modelling exercise, finding the issues to be addressed. These are the main discourses, narratives, trends and changes and also local drivers affecting them.

The issues will come from some of the following topics:

- Population dynamics
- Natural resources and cultural heritage
- Land use
- Territorial structure
- Governance
- Knowledge and R+D+i
- Socio – economic system

Institutional framework and policies:

To help identify the topics find below a proposed list of terms relating each of the topics. This list can be updated as a work in process.

Thesaurus

- **POPULATION DYNAMICS**: depopulation, aging, young, immigration, rural population, urban population, average lifetime, birth rate.
- **NATURAL RESOURCES AND CULTURAL HERITAGE**: natural resources management, water use, soil, soil management, landscape, biodiversity, natural risks, conservation of the heritage, community implication, environmental quality, waste management.
- **LAND USE**: land use intensity, transformation trends (urban to rural; natural to not natural; protected land; urban uses and planning), territorial polarization, land use normative, forest management, agriculture, crop cover, crop diversity, crop rotation, agricultural activity.
- **TERRITORIAL STRUCTURE**: urban development, urban sprawl, urbanization, urban fringe, peri-urban areas, horsification, gardenification, rural development, coastal areas, inland areas, polarities and networks, mobility and transport infrastructures, territorial equilibrium, insertion in superior territorial units (region, state).
- **GOVERNANCE**: shared views, identity, strategies, integrated management, corporate participation, public participation, policies, opinion diversity, common entrepreneurship (common initiatives), social cohesion, quality of life, politics about rural development, housing policies, population dynamics.
- **KNOWLEDGE AND R+D+i**: knowledge management, shared knowledge, knowledge generation, agrarian training, information flows, information culture, literacy, generation of patents, doctorates, technology acquisition, skilled labour, advanced start-ups and spin-off, R+D expenditure, higher education diploma, higher secondary diploma, higher secondary vocational diploma.
- **SOCIO-ECONOMIC SYSTEM**: structure of the competitiveness for the different sectors (exogenous, endogenous; long term, short term), income polarity, profitability, environmental externalities of the economic activity, the structure of the family and public income and expenditure, labour market, environmental principles and rules of the local

economy, prices, incomes, outsource, labour migrants, agricultural sector, industrial sector, specialised services, tertiarization.

- **INSTITUTIONAL FRAMEWORK AND POLICIES**: decision making process, strategic and planning mechanisms, leadership, legal framework, provision of public services, objectives, goals, strategic goals, strategies, measures, policies, priority aspects, policy recommendations.

Sources

The main source documents will be type 2 and 3 studying topics or subtopics affecting the region.

Find below examples of text excerpts coming from the sources provided. You can find highlighted in green the general rules to identify the issues to be addressed:

> The fall from 42.282 farm managers in 1999 to 29.394 in 2009 coincided with an increase in the average age of farm managers from 46.2 years to 49.5 years. Moreover, there is only a small number of young managers: in 2009 only 2.3 % of Flemish businesses had a manager aged under 30, and 7.9 % were older than 65 (LARA, 2011).
>
> The context of Flemish urbanisation causes cities to grow beyond their boundaries to form urban regions with a city centre, an agglomeration and a suburb. These urban regions are very large compared with other countries. As a result, the majority of the Flemish population in the highly urbanised Flemish region actually lives outside the city centres (Boudry et al., 2003).
>
> Increasing urbanisation not only results in potentially less available open space; a growing demand also results in higher land prices.
>
> In Flanders, climate change is expected to manifest itself primarily in a marked temperature rise with an increase in frequency of extremely hot summer days and in high precipitation variability, with an increase mainly in winter precipitation.
>
> Under a high climate change scenario, harvest losses of up to 30 % are likely due to drought stress for shallow-rooted summer crops such as sugar beet, grown in sandy soil.
>
> Potential consequences in the field of animal production are higher wind chill temperatures, leading to production losses, new illnesses and plagues, lower energy demand for heating and higher energy demand for cooling (Gobin et al., 2008).
>
> While direct water consumption (total water consumption excluding cooling water) may have decreased considerably in Flanders over the past decade (-10 % in 2009 compared with 2000; MIRA, 2012), pressure on water resources remains high. With a value of approx. 32 % in 2007 (EEA, 2009), the Belgian Water Exploitation Index (WEI, actual water consumption expressed as a percentage of water availability) exceeds the 20 % threshold, which is considered as alarming (Alcamo et al., 2000).
>
> In Flanders, too, numerous claims are putting pressure on the limited available space. Typical local elements of such pressure are the growing demand for construction and industrial sites, 'horsification' and 'gardenification' (Bomans et al., 2009; Bomans & Gulinck, 2008).

> One of the causes of the successor problem in agriculture is the fact that lower prices and falling incomes often force farmers and/or their partners to take on extra work outside the business. Only 13.8 % of Flemish farm managers aged over 50 have a potential successor.
>
> Moreover, certain tasks are outsourced, either to contractors for specialised operations that require expensive machinery, or to labour migrants from Eastern Europe (mainly Poland, Romania and Bulgaria) for seasonal work for which local workers are hard to find. Furthermore, an increasingly greater portion of income is earned outside the business.
>
> Thus, 54% of agricultural households derive income from non-agricultural activities. A positive aspect of human capital is that starters in the agricultural sector are relatively well trained: 18% hold a higher education diploma, 28% a higher secondary diploma in agriculture, 12% a higher secondary vocational diploma in agriculture, and 33% an installation certificate.
>
> The decrease in the number of farmers, although increasingly better trained, goes hand in hand with scale enlargement, specialisation, high capital intensity, and small profit margins. This puts continuous pressure on business management: business managers are required to manage increasingly bigger units and are ever more dependent on uncertain markets (and therefore income) and capital markets to finance their business. Poverty and cash flow problems are therefore still considerable, but also difficult to estimate. In 2009, 254 farmers applied for aid from the non-profit association Farmers at a crossroads. However, farmers identify administrative burden as one of their main professional problems, which also generates a great amount of stress. No data are available on social capital, e.g. degree of integration in social life (e.g. membership of associations).

General rules for the identification can be summarized as follows:

WORDS FROM THESAURUS

VERBS (or VERBAL FORMS) INDICATING INCREASE, DECREASE, DEPENDENCE: increase, manifest, grow, lead to, decrease, put pressure, identify

ADVERBS OR OTHER FORMS INDICATING GROWTH OR DECLINE: fall from, the majority of, highly, increasingly

FIGURES OR QUANTIFICATION FORMS: the majority, threshold, greater, better, bigger.

Social networks may also be a source rather adding quantitative aspects to the identification than as identification source itself.

Additionally, social networks might work as identifiers of new issues arising (in the case of Flanders: Mercosur negotiations, specific climate change consequences…).

Expected Results

The final result expected is a **list of issues to be addressed** with any kind of quantification coming from figures from the text but also a quantitative analysis of the text itself and the quantitative inputs from social networks.

2.4.1.3 Technical and behavioural parameters

This is the last and most accurate analysis, with the idea of giving clues for the structure of the model, once the issues to be addressed have been identified.

Sources

Documents type 2 and 3, and specifically around the words identified in section b (Issues to be addressed).

The task here is to identify this text excerpts that could be easily translated into model pieces. Find below some examples found in the sources provided. Marked in purple the key words to identify the excerpts.

> The number of starting farm businesses in Flanders has fallen under 200 per year. In order to keep the agricultural sector viable, more beginning farmers are needed. The reasons for the low number of beginning farmers are multiple: an insufficient number of economically viable businesses to take over, an insecure and low income, legal uncertainty, increasing social demands, a less favourable image...
>
> The countryside is facing a rapid evolution due to changes in the Flemish agricultural sector (fewer but larger agricultural businesses, diversification, part-time farming, changing views on landscape and buildings ...).
>
> As in other regions, agricultural activity in Flanders has changed considerably, with the main trends determining its development in the second half of the twentieth century: intensification, specialisation and concentration.
>
> The research results show that active diversification (being processing at the farm, alternative markets and tourism) distinctly occurs more as the distance to the city decreases (Figure 10). In other words, active diversification activities increase as one approaches the city. On the other hand, environmental measures and the maintenance of hedges and trees occurs more frequently further away from the city.

The rule to identify is not so clear in these cases, but it has to do with either verbs (verb forms), nouns or expressions giving the idea of causality. Every excerpt of text is linked to an identified issue.

… ARE NEEDED; THE REASONS FOR…; DUE TO…; TRENDS DETERMINING…

…OCCURS MORE AS… DECREASES; …INCREASE AS… APPROACHES…; OCCURS MORE FREQUENTLY FURTHER AWAY

Expected Results

The results here are dynamics in the form of pieces of text that can be easily translated into models. The dynamics are linked to the issues previously identified, and more than one dynamic may be linked to every issue.

2.5 Text Mining and PoliRural Digital Innovation Hub

The PoliRural Digital Innovation Hub (DIH) will manage large datasets for the PoliRural project. Many of these will have a spatial character or components. Part of the data will be in the form of

Linked Open Data. For these reasons there will be two basic requirements from the side of DIH towards text mining:

- It should be able to discover locations in text document (via geoparsing).
- Link the information contained in text with Linked Open Data.

2.5.1 Geoparsing

A Geoparser Service is a network-accessible service that focuses on the geoparsing and marking of free text messages using a vocabulary, such as place names, which is possibly specified by the user. The output from a Geoparser Service is a collection of features that identifies words and phrases in the original text resource. The returned collection of features is suitable for subsequent processing, such as user-controlled geocoding. It is anticipated that this Geoparser Service will have a significant impact on the ability of applications to share multiple distributed interoperable Geoparser Services and offer a useful service to the rural community (Charvat et all, 2006). There will be two potential outputs from Geoparsing:

- Metadata about documents, which will allow to link a given document with a concrete location. There could be the possibility of using also other analysis of text like time stamp, key words subjects, etc.
- Service, which will support direct visualization of places in the text on the map in Figure 2.

Figure 2 Example of Geoparsering service

2.5.2 Text mining and Linked Open Data

Text Mining and Linked Open Data (LOD) will be used in synergies. There could be a benefit for text mining using as an additional source of entities to improve text mining. On the other hand, text mining can control or improve information in RDF LOD data. Currently, there exist a number of scientific papers related to this type of analysis (Paulheim, 2013; Lausch et al, 2015). These are the next steps to combine Text Mining and LOD:

- Entities in Linked Open Data have to be recognized as entities in the original dataset.
- Data about the entities is extracted using the previously identified URIs. It is helping to generate additional features and add them to RDF data sets.

3 The Text Mining Solution

In this document the concepts, terms and definitions are given strictly related to the project. They do not necessarily correspond to the concepts in the "real world", described by the same keyword(s).

3.1 General concepts

3.1.1 Keyword

Keywords are separate words (or short phrases that consist of maximum three words) used for filtering streams of content from Social Media such as Twitter or LinkedIn.

In the current project keywords are obtained in two ways:

1) The list of initial keywords (for each language) are prepared by experts.
2) In order to collect information as close as possible semantically to the given topics, the list of keywords should be continuously updated, based on the information collected and indexed. New keywords will be obtained using Adaptive Keywords technique (Vitiugin and Castillo, 2019) as well as extracted from **GDELT** (3.1.10).

A **keyword** can also be used as a **topic** (see below), but those terms are defined independently as they serve different purposes: keywords are used for filtering, while topics for semantic analysis.

3.1.2 Topic

Topic is an ultra-short summary of a given semantic entity or a sphere of interest, that can categorize a certain document found on a website, in a paper, or in Social Media feeds.

Any given text can contain more than one topic. Examples: "landscape development", "landscape quality", "new farming business model", "rural policy", "rural policy measures", "landscape dynamics", "leader", "biodiversity", "ecosystem services", "landscape maintenance", "landscape architecture", "historical landscapes and heritage".

3.1.3 Subtopic

Topic can include **subtopics**. Examples (in English):

- "Landscape development" subtopics:
  - "Landscape quality"
  - "Landscape dynamics"
  - "Landscape maintenance"
- "Rural policy" subtopics:
  - "Rural policy measures"
  - "Rural policy proposal"

3.1.4 Similarity Cluster

Topics and their subtopics can be represented as a *directed graph* and in the context of the PoliRural project it is called **Similarity Cluster** (see Figure 3) (Hradec et al., 2019).

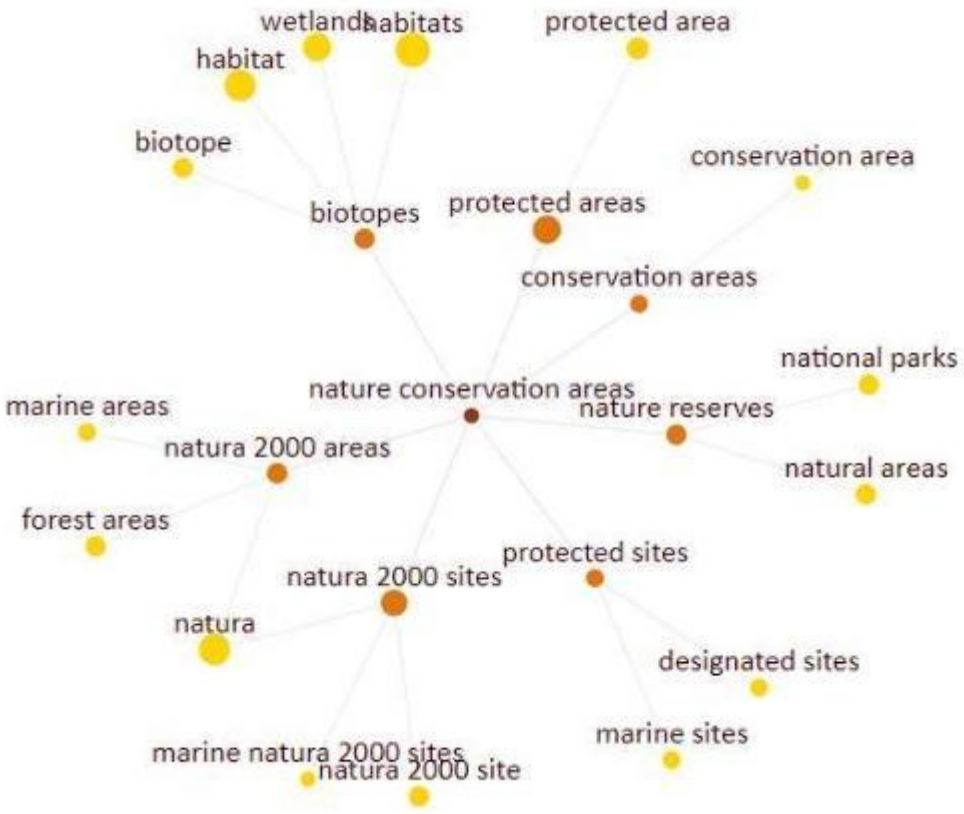

Figure 3 Example of Similarity Cluster

The topic is a root, while subtopics are nodes. If a node has children, it is considered to be a topic on its level, while his children are subtopics.

3.1.5 Shared Term Space

A network of two or more separated topics semantically linked together by subtopics form *undirected graphs*, i.e. some topics can be universally general, while others can be both independent topics or subtopics of other, more general, topics. In the context of the project it is called **Shared Term Space**.

The relation between topics and subtopics in Shared Term Spaces are different from Similarity Clusters, and Figure 4 demonstrates this difference: the incoming arrow indicates subtopic, an outgoing one indicates a topic. Therefore, a node can simultaneously be a topic and subtopic of another topic.

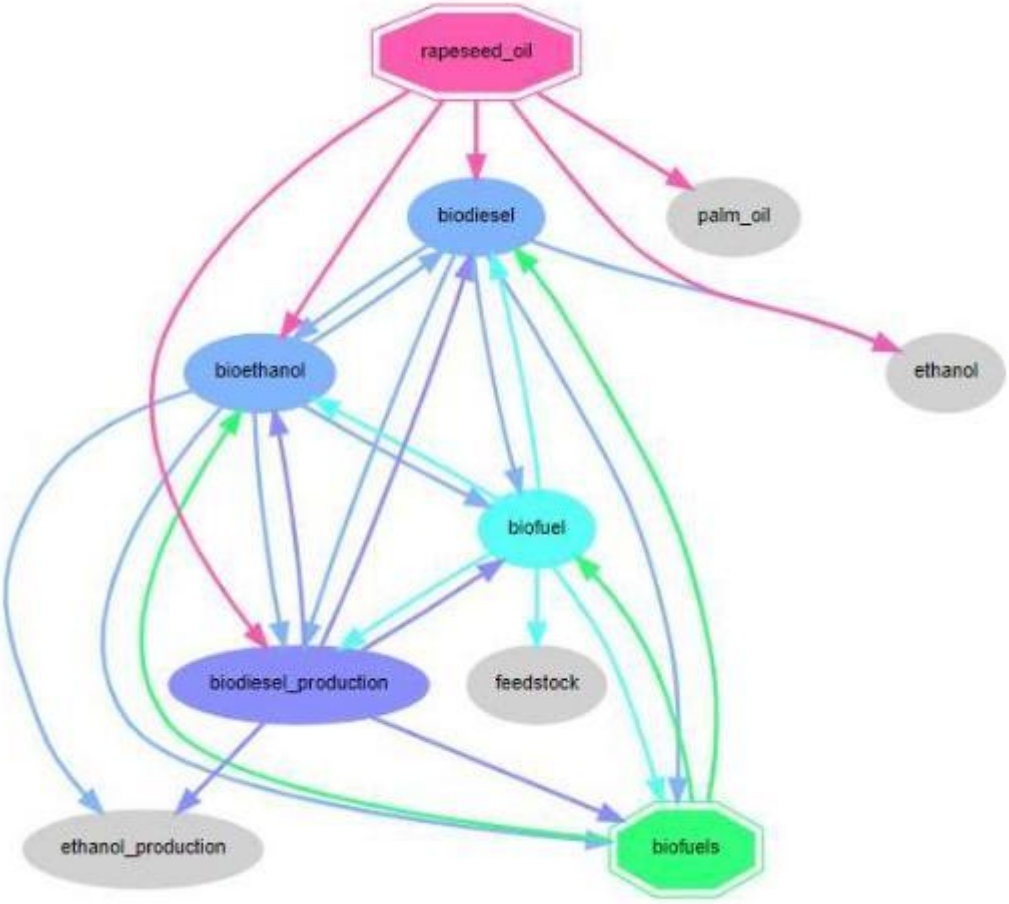

Figure 4 Example of Shared Term Space

3.1.6 Direct quote

Direct quote is a text that repeats another text, being already processed by the system, for example, re-tweet or a post on Facebook shared as a tweet. Direct quote can only be recognized as such if it is compared and acknowledged as being similar[5] to the text, already processed by the system and saved in the database. A text that is compared to another text, for which, is being considered as a "direct quote" and should be discarded from processing.

3.1.7 Geospatial Location

Geospatial Location includes:

- **Geo-point** (latitude, longitude)
- **Area**:
  - **Geo-point** together with the **radius** (in km)
  - **Geohash code** (Niemeyer, 2008)
  - **Multi-polygon** (following a definition given in the Geo-JSON standard)

[5] In technical terms it means that a similarity ratio of two compared chunks of text is above a defined threshold.

3.1.8 Links Library

Links library is a list of URLs provided by partners, saved in the database with the possibility to be continuously updated.

3.1.9 Curated Reading List

Curated Reading List is a result of merging topics from **Links Library** provided by the partners with the external dynamic information found in the Internet. The regular updates from the GDELT project (see below) will be used as a source of information from the external world.

3.1.10 GDELT

GDELT (The Global Database of Events, Language, and Tone) is a project by Google that "monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, themes, sources, emotions, counts, quotes, images and events driving our global society every second of every day, creating a free open platform for computing on the entire world" (Gdeltproject, 2019).

Information from GDELT is used for different purposes in PoliRural: as it already contains named entities and semantically categorized, it can be used to update Curated Reading Lists (3.1.9), in the process of Horizon Scanning[6] for identification of emerging issues, etc.

3.1.11 Local Database

Local Database is a general term that represents any storage for data used in the project. This can include database of indexed documents for Semantic Analysis, Links Library as records in a chosen database or documents obtained from Social Media, whether normalized or in their raw form.

In the context of this project the term **Local Database** does not include databases for additional services like User Management or Periodic Tasks Management, etc.

3.1.12 Point of Interest (POI)

Point of Interest (**POI**) is an object, feature or phenomenon, which could be represented in the scale and character of the model (usually in a map) by point symbol. The content of points of interest is done by the domain developing and using concrete POIs. For example, agriculture deals with different POIs than tourism or transportation. But several types of POIs can be shared across domains.

3.1.13 Smart Point of Interest (SPOI)

Smart Points of Interest (**SPOI**) is a specific database of POIs. This dataset and primarily the data model could serve as an exchange format for point-based objects (points of interest). The main benefit of the proposed solution consists in its high interoperability. The SPOI data is published as the 5-star Linked Open Data (LOD). It means uniform, but flexible and expandable data format,

[6] For a general explanation of Horizon Scanning refer to http://portal.healthworkforce.eu/what-is-horizon-scanning-and-why-is-it-useful/

the standardized system of identifiers (URI), transfer protocol (HTTP), implementation of existing semantic vocabularies (FOAF, GeoSPARQL, RDF, Dublin Core, etc.) or mechanism of querying.

The SPOI dataset had been developed in the SDI4Apps project[7]. The current version of SPOI includes more than 33 million of points from the whole world. SPOI data usually arise from existing open or free spatial data transformed into the uniform SPOI data model. There are many different data resources from massive global data adopted from OpenStreetMap, GeoNames.org or Natural Earth to local, but more detailed data provided by regions (Posumavi region, Czech Republic, Sicily, Italy, Zemgale, Latvia or Belluno, Italy), projects (Citadel on the Move, Open Transport Net), municipalities (open data from Prague, Czech Republic) or volunteers (research data from the University of West Bohemia, Czech Republic, data collected by SPOI map client). Data is accessible through SPARQL endpoint and map client, but it is possible to create a data dump on request.

3.2 Text Mining Pipeline

A generalized pipeline of Text Mining for Foresight Process is given in Figure 5. Each of the stages is briefly discussed in the following chapters.

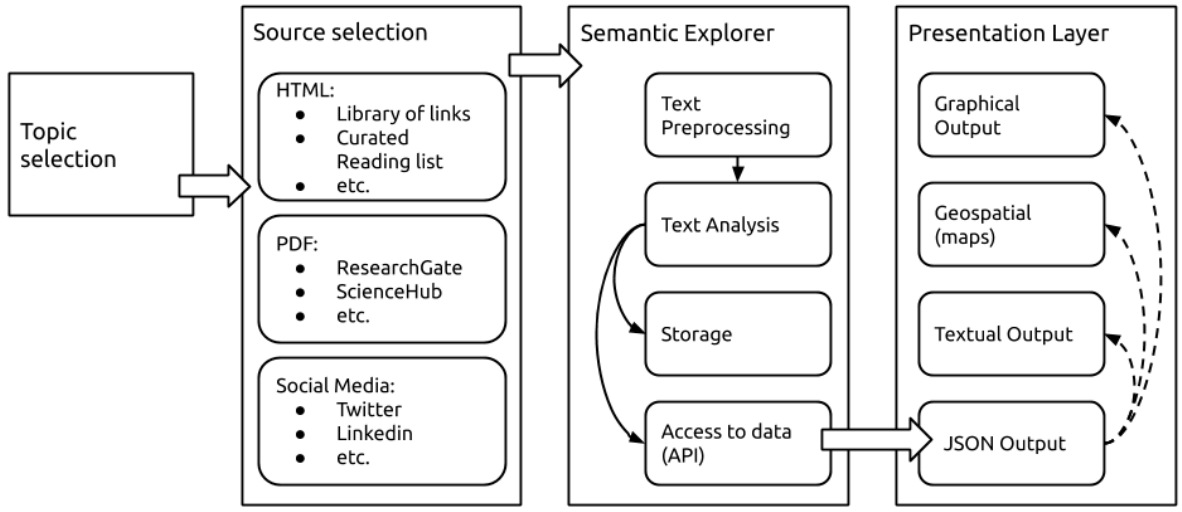

Figure 5 Text Mining Pipeline (generalized)

3.2.1 Topic Selection

Topic selection is a manual process of the Foresight exercise, conducted by the end-user. The result of this stage is a topic or a collection of topics, that define the successive stages of the Text Mining Pipeline. For example, a user wants to examine the dynamics of a set of selected performance indicators in the area of Landscape Development - therefore, the result of this stage is "Landscape Development" (a topic).

Together with the source, selected on the next stage, topics become inputs for Semantic Explorer.

[7] https://sdi4apps.eu/, 2019

3.2.2 Source Selection

Depending on what the end-user wants to achieve, a certain source of information should be selected. Examples of sources:

- Links to hypertext or PDF files (either direct links or selected from Library or Curated Reading Lists). There sources are used for extraction of Key performance indicators, (performed by Profiler) or Topic Modelling (see section 5.1.2.4 "Profiler").
- Uploaded documents[8]. This can be used for a variety of tasks, that are described in section 8.2 "System Components".
- Social Media (SM) posts. In this case an API is used, that returns a number of posts from SM according to filters defined by end-user. For example (conditions of filters are given in italic), "*tweets* from a certain *geospatial area* collected during *the last month* that contain selected *keywords* and classified as belonging to a certain *topic*).

In case of selection of links to RTF, PDF, DOC or other file formats of interest, they should first pass through the dedicated Crawler (see 4.4 "Crawler and Storage") in order to become available as analysable text.

3.2.3 Semantic Explorer

The Semantic Explorer is a system component which provides end-users with textual and quantified information, automatically extracted from the big corpus of texts.

The input for Semantic Explorer is a meaningful sentence (topic or topics) and text (or collection of texts) that are the results of the stages of Topic Selection (see section 3.2.1) and Source Selection (see section 3.2.2).

The Semantic Explorer performs all the tasks that prepare text for analysis (Preprocessing), and the analysis itself. It also stores processed text (as well as required meta information such as indexes and world/phrase vectors) in the defined storage and return the information to the end-user in a required form. All those procedures are described in detail in chapter 5 "Semantic Explorer".

3.2.4 Presentation Layer

The information obtained from Semantic Explorer is going to be presented as responses from a set of API endpoints in JSON format and its derivatives (for example, Geospatial information is going to be presented as GeoJSON). Those documents can therefore be interpreted later as markup textual output (HTML, MD, etc.), graphics (charts, diagrams, etc.) or maps (heatmap, multi-polygonal areas, etc.). This is a task of interpretation layer, and is supposed to be realized on the later stages of the project, after the agreement with partners.

[8] At the stage of creation of the prototype the only possible formats of uploaded documents will be PDF, SGML (such as HTML or XML) and plain text. On the later stages of the project more formats can be added (such as MS Word, LaTeX, Markdown documents, etc.), when developers will identify the most used formats in the Foresight Exercises and after agreement with end-users.

3.3 Streaming from Social Media

Streaming from Social Media is a process of registering data from platforms such as Twitter and LinkedIn in the Local Database to be available on the stage of Source Selection (3.2.2) of the Text Mining Pipeline.

A general pipeline of the Streaming from Social Media is represented in Figure 6.

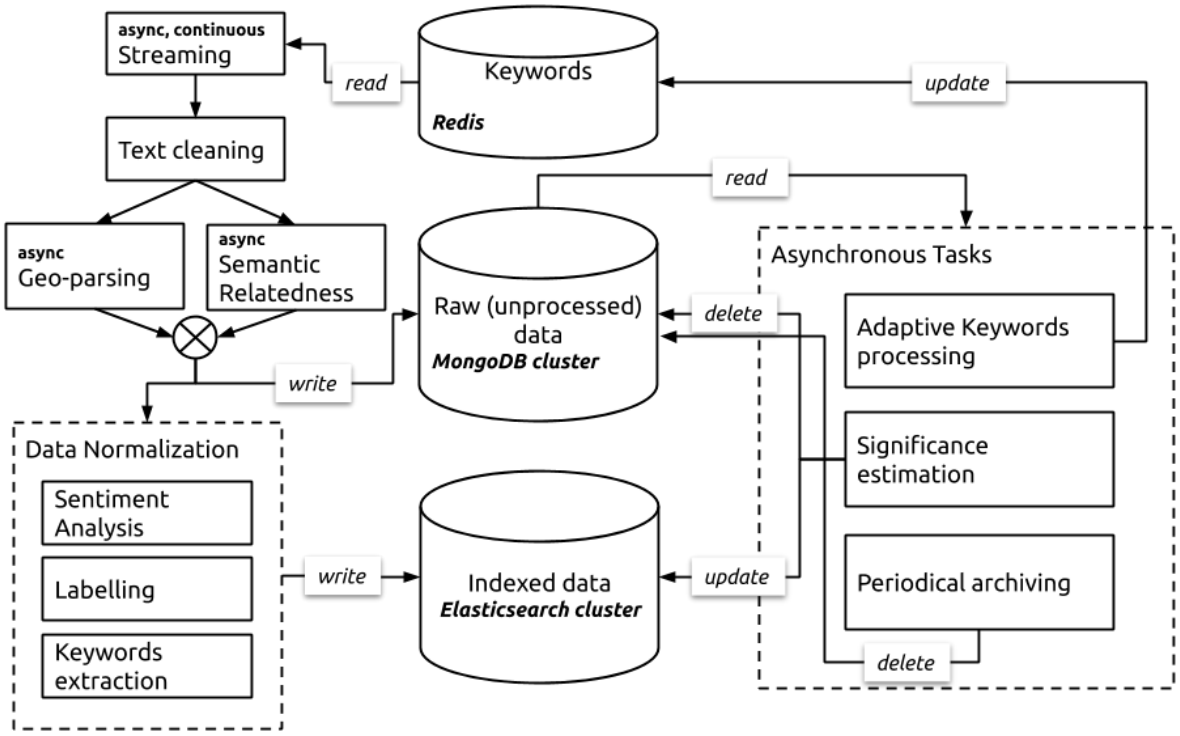

Figure 6 Streaming from Social Media Pipeline

Warning! Geoparsing and Semantic Relatedness Estimation are asynchronous processes and work in parallel. The amount of time required to process a post from Social Media heavily relies on the amount of text, number of named entities mentioned, topics touched on - and therefore unpredictable. Making those processes asynchronous with a considerable timeout eliminates bottlenecks in the queue of Social Media feed.

Warning! A post from Social Media can only be registered in the Local Database and sent for further processing only if both Geoparsing and Semantic Relatedness Estimation return positive results.

3.3.1 Text cleaning

During the text cleaning phase the following should be performed:

- Removal from the original text of all symbols, stop-words, etc. that are not relevant to text processing.
- Removal of personal sensitive information (names, phone-numbers, emails, etc.)

Techniques: NLP, lemmatization, NER.

3.3.2 Geoparsing

This stage is required to filter out posts that do not contain information about geolocation(s). Such posts cannot be used for the analysis of Pilot areas and therefore should be discarded.

Technique: API call to Geoparser (see section 5.1.2.2) of the Semantic Explorer will be used.

3.3.3 Semantic Relatedness Estimation

This stage is necessary for filtering out useless tweets or posts from SM platforms. The number of raw posts streaming from Social Media platforms are forecasted to come in enormous quantities, while only a fraction of them can serve as a valuable input for Semantic Explorer.

The measure of Semantic Relatedness quantifies the degree of relationship between a document (Social Media post) and the defined topics in the scope of the PoliRural project (such as "rural attractiveness", "rural landscapes", "rural policy" etc.). In other words, it shows what is the probability of a post to be "about something" that interests end-users.

If the probability is lower than an estimated threshold, the post is discarded entirely. Otherwise it's being registered in a Local Database for raw data, and passed to the next stage of Data Normalization.

Warning: for this stage it is critical to find a good semantic measure to show that a post correlates well with human judgment, and helps solving the task of ontology matching.

Technique: Topic matching with topics obtained via API call to the corresponding endpoint of Semantic Explorer (chapter 5).

3.3.4 Data Normalization

Data Normalization is a process of converting an original post from SM (tweet, post from linkedin, etc.) to the unified form for feeding to clients of API or/and live feed (end-users).

Normalization includes:

- Sentiment Analysis
- Labelling
- Keywords Extraction

3.3.4.1 Sentiment Analysis

The result of this stage is an indicator of polarity of opinion, expressed in numerical form on the scale of [-1..1], where:

- -1 is the most negative
- 0 is neutral
- 1 is maximum positive

Technique: NLP, Sentiment Analysis.

3.3.4.2 Labelling

Calculating the probability that the post relates to a certain topic. This stage is possible only after the initial extraction of the topics from Links Library or defined topics provided by partners.

Technique: ML, Multiclass Classification Task.

3.3.4.3 Keyword extraction

Each text should be converted to the list of keywords (bag of words) for the later (periodical) stage of Adaptive Keyword processing.

Technique: NLP, lemmatization.

3.3.5 Asynchronous Tasks

The main purpose of asynchronous tasks is the selection of the number of SM posts already saved in the Local Database for additional processing, and (if necessary) deletion or update, based on some common information.

3.3.5.1 Adaptive Keywords processing

Due to the linguistic elasticity and dynamic nature of policy making, public opinion and constant changes in internet jargon, there is a necessity in the continuous process of updating existing keywords.

Technique: described in the paper mentioned (Vitiugin and Castillo, 2019) and will be implemented as a periodic task. The periodicity of triggering the task depends on the number of social media posts, that successfully pass through Normalization and stored in the storage for Indexed Data.

3.3.5.2 Significance estimation

The task of this stage is to mark out posts that aren't representative (see section 3.1.6 "Direct quote") and to remove them from the Local Database. Examples are re-tweets on Twitter, posts from LinkedIn shared via Twitter, etc.

Technique: NLP, a ratio based on texts *multiplicity* and *centrality*[9]. A text that is compared to another text, for which this ratio is above a certain threshold, is being considered as a "direct quote" and should be removed from the storage. This should be written as a periodic task, because the estimation of representativeness can only be measured post-factum within a set of posts collected, stored in the Local Database, and segmented[10] for the purpose of comparison of their ratios.

3.3.5.3 Periodical archiving

The amount of data is expected to be massive even on the stage of prototype solution for English language only. After adding all languages it is expected to grow at the rate of several gigabytes a

[9] https://en.wikipedia.org/wiki/Outline_of_natural_language_processing#General_natural_language_processing_concepts

[10] For example, by the country of origin, language, etc.

month. Therefore, there should be a process of periodical archiving of old data from Raw Data storage.

3.4 Curated Reading List Management

A separate process (and a special case) of the Text Mining is a creation and recurrent update of **Curated Reading Lists** (see 2.1.9 "Curated Reading List"). To match topics, extracted from Links Library and information from GDELT, we will use Topic Manager (5.1.2.3) of Semantic Explorer system component. In this process Topic Extraction is responsible for extraction of topics from partners' data, while Topic Modelling is used for creation of the Shared Term Spaces across the information obtained from GDELT.

For each language there should be a separate Curated Reading List.

Recurrent update of a reading list is an inherently asynchronous process, which depends on the information obtained from the partners.

Curated Reading Lists are open in nature and should be available to all partners of the PoliRural project.

A generalized pipeline of the Curated Reading List update is given in Figure 7.

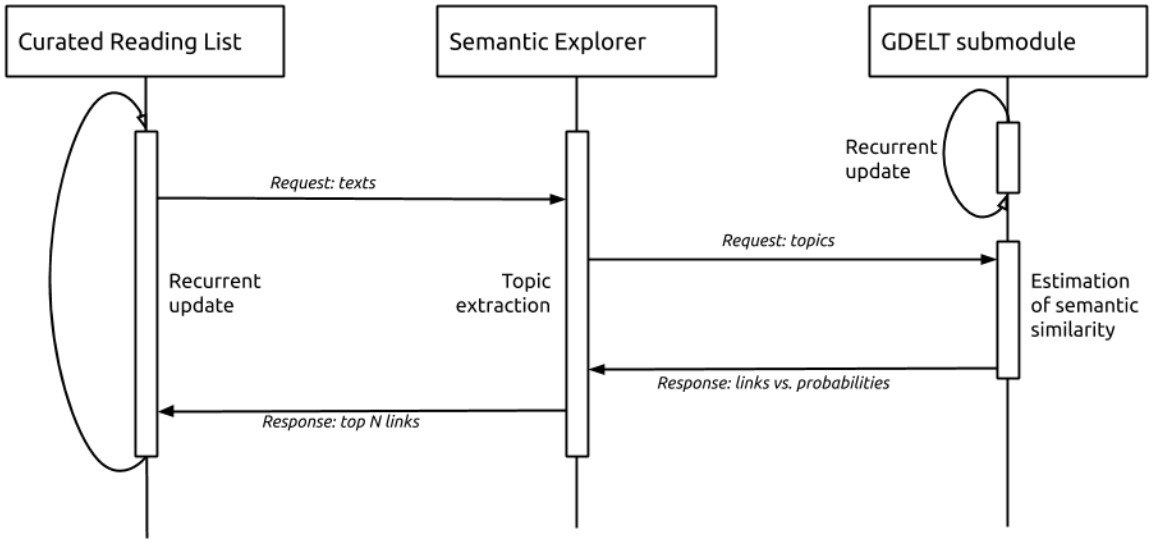

Figure 7 Asynchronous process of Curated Reading List update

4 Source Selection

In this section a short overview of all possible sources of information is presented as well as the type of information that should be extracted from each particular source type and techniques used to solve those tasks.

Types of sources that can be processed by Semantic Explorer:

- PDF document (static information)
- forum / discussion board (SGML, dynamic information)
- news and articles (SGML/PDF, mixed)
- posts in Social Media (JSON, dynamic information)

Each document should be accompanied by the following information:

- language (required)
- Meta-information: description, creator's username[11], date of creation, etc.

Each source is being "trapped" into a chosen pipeline of pre-processing, depending on its format, and whether its content is static or dynamic.

4.1 PDF Documents

PDF documents are scientific papers, reports, etc. They contain static information, and therefore will not require periodic procedures (such as indexing) to update it in the DB.

PDFs are used as inputs for the task of Topic Extraction. The list of topics extracted from the library of PDFs are used on the later stages for creation of Curated Reading Lists (see section 3.4 "Curated Reading List Management") and Text Analysis (see section 3.2 "Text Mining Pipeline").

In general case documents of this type will be represented by URLs in the Links Library. Any URL, provided as a param by the end-user in the request to Semantic Explorer's API, should automatically be stored in the links library, and the document, which it refers to, should be automatically normalized, indexed and saved in the database.

The end-user can also upload documents directly to the cloud, in which case the URL to the document stored in the cloud becomes the link in the Links Library.

There should be a procedure of checking certain resources (such as ResearchGate[12]) periodically for new documents on the topics of interest.

4.2 Forums and discussion boards

This is an inherently dynamic source, which provides inputs for tasks such as identification of emerging issues, measurement of Sentiment dynamics (regarding previously defined topics), etc.

[11] This is not necessary in case of Social Media, because of the streaming (i.e. information from Social Media is being updated continuously and automatically).

[12] https://www.researchgate.net/

The main techniques that are used to process information from forums and discussion boards are Topic Extraction (see 5.1.2.3 "Topic Manager") of Semantic Explorer and Sentiment Analysis of SM Data Normalization (3.3.4).

The documents of this type are represented in the Links Library in the database. However due to the dynamic nature of the content, the process of semantic analysis of such links should be triggered by the end-user manually, i.e. by issuing a request to the API (section 5.3 "Access to Data (API)"). The results of the stage of pre-processing of these documents (bag of words, TF-IDF, words and phrases vectors, etc. - see section 5.1.1 "Text Preprocessing") will not be stored in the database (or will be stored for a limited period of time).

4.3 News and Articles

Documents of this type can be either static web-pages with text (HTML) or dynamic content from RSS feeds (XML) or news websites (HTML).
The processing of the static web-pages will be similar to the processing of PDF files. Dynamic content is being fed into the queue of Semantic Explorer on a regular basis (periodic task), and processed similarly to Forums and discussion boards.

4.4 Crawler and Storage

Crawler provides a general method for continual collection of unstructured data from the internet. It is primarily intended for bulk acquisition of common web-pages and other web published documents.

Amongst core engine, Crawler consists of protocol modules and format modules. Through protocol modules, Crawler implements several standards, such as DNS (RFC 1035), HTTP (RFC 2616), and HTTPS (RFC 2818). Through format modules, Crawler implements own SGML/HTML parser and integrates 3rd party format converters, such as PDF (Adobe) or DOC (Microsoft). Crawler can be extended to support other communication protocols and / or file formats as well.

The basic format processing is required at the Crawler level to perform:

- Link discovery, which is necessary in real-time for an automatic recursive operation.
- Meta/data differentiation and conversion into plain-text (including charset and encoding normalisation), which is suitable for further processing and long-term storage and future reference.

The Crawler often stands at the beginning of a specific text-mining pipelines. Due to the volume of source data, it can also represent the last resort for non-trivial evaluation, such as content vs non-content (i.e. web advertisement) differentiation, selective text truncation (to save resources located further in the pipeline), etc.

Due to the bulk nature of its operation, the Crawler does not operate in request/response manner. Instead it receives tasks in the form of internet links or domains and then delivers content as being acquired to a selected destination continually.

The Crawler architecture focuses on versatility, sustainability and performance. The communication protocols are implemented in fully asynchronous way. Various mechanisms are in place to avoid waste of resources (such as unnecessary reconnections), remote server overloads, etc. Such approach allows to communicate with tens of thousands of servers concurrently through a single Crawler instance.

Crawler API consists of single method "Acquire", for registering task in terms of input (selected origin, such as link or domain) and output (selected destination, such as storage or processing API).

In terms of long-term operation, Crawler also directly relates to Storage (5.2). Storage preserves clean (plain) text substance and crawler meta-data of information acquired over time across regional and global sources of interest. The idea of Storage is introduced in order to assure reliable and efficient access to all acquired inputs in the future, to allow:

- inquiry and reference (for online users needs and convenience)
- input context resilience (i.e. manual insertion of relevant documents into the system or automatic mirroring of selected 3rd-party resources)
- vindication and improvement (for particular processing verification)
- processing iteration (as contexts and processing chains evolve in time and need to be reapplied)

Crawler uses Storage to archive variable sets of attributes, such as normalized URI of origin, content, title, time of acquisition, checksums, etc. For standard operation it requires methods providing file ("Insert" / "Remove" / "Browse") and file's attribute ("Get" / "Put" / "Rid") operation.

4.5 Social Media

Streaming from Social Media (SM) is a constant feed of data from the selected platforms into a Local Database. Gathering, normalizing and indexing posts from SM allows researchers to:

- Build collections around a particular event or topic.
- Update Points of Interest and create Smart Points of Interest.
- Estimate users' feedback on a certain topic or policy via Sentiment Analysis.
- Gather datasets tailored to specific research questions by selecting posts from specific groups and communities.

In the scope of this project the streaming from two popular Social Media platforms will be covered[13]: Twitter[14] and LinkedIn[15].

Posts from SM are being filtered by keywords (see section 3.1.1 "Keyword") and/or by certain geographical areas (see section 3.1.7 "Geospatial Location"). Before being registered in the Local Database, every single post from SM is being converted into a JSON document and fed into its

[13] This is valid for the stage of prototype. Upon the agreement with the partners more platforms can be covered by streaming on the later stages.

[14] https://twitter.com/

[15] https://www.linkedin.com

own pipeline (see 3.3 "Streaming from Social Media") for validation and normalization. Selecting data from Social Media as a source for Text Analysis means applying filters to SM posts registered in the database.

5 Semantic Explorer

Semantic Explorer is a general term, which serves as an umbrella for a number of services, each of which takes a text and language as its input, and returns a meaningful result that characterizes the subject(s) of the text in a defined way. For the end-user each service is an API (Application Programming Interface) endpoint (see section 5.3 "Access to Data (API)").

5.1 Semantic Explorer Pipeline

Semantic Explorer consists of the set of services, available as API endpoints. The outputs of one service can be an input for another, which suggests a concept of a pipeline. For example, the result of NER is used by the Geo-parser, which returns coordinates and geo-shapes by geo-names. Likewise, a Profiler uses results of both NER and Topic Extraction to define a context, in which it should look for meaningful characteristics of the subject (see Figure 8).

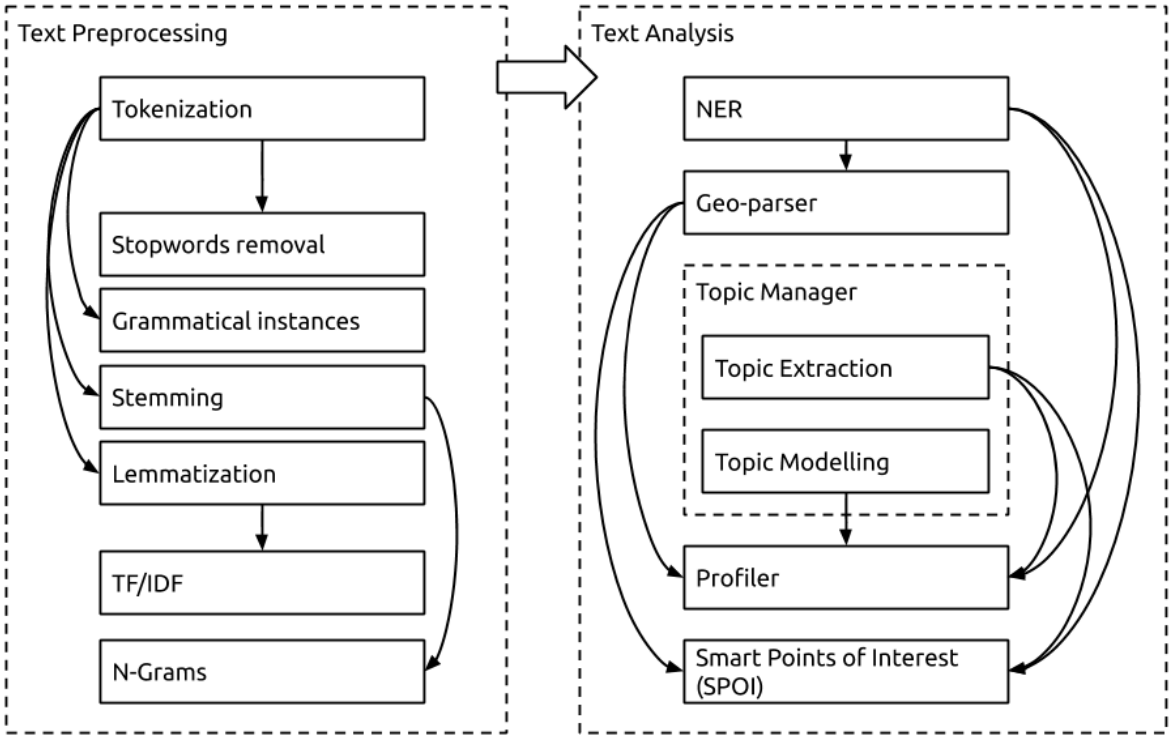

Figure 8 Semantic Explorer Pipeline

5.1.1 Text Preprocessing

Before the text can be processed, it requires to be structured and transformed into a machine-readable format. The first step is to divide the text into its single elements as words (tokenization) and represented as a vector.

To extract relevant terms, mainly two different approaches are distinguished: working with stop words or grammatical instances. In the current project both processes are required.

Extracting grammatical instances means to tag each word with a part of speech such as verb,

article or noun. From this, the relevant phrases or chains of words should be extracted. This is necessary on the later stage (see 5.1.2.4 "Profiler").

Stop words are used to remove irrelevant terms and function words (articles, conjunctions, pronouns, etc.). Stemming is cutting each word to its basic form. Lemmatization reduces word to root form based on dictionary. This is necessary for the creation of bag-of-words on the later stages of preprocessing, and will serve the purpose of topic matching on the stages of Topic Extraction and Topic Modelling (see 5.1.2.3 "Topic Manager").

Finally, independent of which strategy is used up to this point, the frequency of the terms (TF/IDF) is stored for further analysis[16] (Manning et al., 2009).

5.1.2 Text Analysis

The description of each Semantic Explorer sub-module is accompanied by input and output specifications, described by keywords (such as "summary" or "highlighted text") with comments necessary for development. The detailed info on the inputs can be found in section 5.3.1 "Unified Inputs for Semantic Explorer" and on outputs in section 5.3.2 "Semantic Explorer Textual Outputs".

5.1.2.1 Named Entity Recognition (NER)

NER serves as an extractor of all the named entities: people, celebrities, geographical names, organizations, notable events, products, timestamps, nationalities or religious or political groups, objects such as buildings, airports, highways, bridges, etc.

Inputs:

- **Text**
- **Language** (optional)

Output: **Highlighted text**

5.1.2.2 Geo-parser

Geo-parser takes a text as an input and returns a geopoint and shape for each geo-name found in the text. This includes administrative units of any scale (countries, regions, states, counties, cities, localities, neighbourhoods, villages, etc.) as well as topological formations such as forests, mountains or water bodies.

The geographical name is an open data taken from "Who's On First" gazetteer[17]. As the gazetteer have records for "all places in the world", a local index should be narrowed down to Europe in order to make queries faster and to simplify the search of places (for example to exclude ambiguity in determining if "Paris" is the capital of France or a town in Texas, USA).

The search for places will combine simple queries in the Elasticsearch index with context-dependent matching of natural features (Elise Acheson et al., 2018).

[16] Manning et al. "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora", 2009

[17] https://whosonfirst.org/

Inputs:

- **Text**
- **Language** (optional)

Output: **Geospatial Location**

5.1.2.3 Topic Manager

Topic manager is a system component, based on the idea of extraction short summary (topic) from an unstructured text as well as statistical estimation of possible subtopics (see 3.1.2 "Topic").

The component itself is logically divided into two independent parts for Topic Extraction and Topic Modelling. The main difference between them is that Topic Extraction produces readable summaries for a collection of texts, which Topic Modelling assigns a probability to the most frequent words to relate to a certain topic and returns to the end-user top N results.

Topic Extraction

The input for Topic Extraction is an unstructured text or collection of texts (more generally - words collected into documents). Given that each document is a mixture of a small number of topics, each word's presence in this document is attributable to one of the document's topics. Each word in a document is given a probability to be related to a certain topic, which makes it possible to construct a short summary preserving orthographic structures (readability).
The main technique used for Topic Extraction is Latent Dirichlet allocation[18].

Inputs:

- **Text** (or collection of texts)
- **Language** (optional)

Output: **Summary**

Topic Modelling

The main task of topic modeling is uncovering of hidden semantic structures in a text body. The result of topic modelling is a statistical model for discovering the abstract topics from an unstructured text or a collection of texts.

It is assumed that words that are close in meaning will occur in similar pieces of text (the distributional hypothesis). A document typically concerns multiple topics in different proportions; thus, in a document that is 10% about "landscape" and 90% about "climate", there would probably be about 9 times more tokens about landscape than climate. The topics produced by topic modeling techniques are clusters of similar words.

The main technique for Topic Modelling is Latent semantic analysis[19].

[18] https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

[19] https://en.wikipedia.org/wiki/Latent_semantic_analysis

Applications:

- searching for similar content (e.g. matching texts from library with policies)
- selection of qualitative indicators (for scenarios matching)
- uncovering emerging trends

Inputs:

- ***Text*** (or collection of texts)
- ***Language*** (optional)

Output: ***Summary***

5.1.2.4 Profiler

The profiler extracts meaningful information that highlights key indicators and/or characteristics of a given subject.

Inputs:

- ***Text*** (required)
- ***Geo-name*** (required) - subject area for the analysis[20].
- ***Language*** (optional)
- ***Keywords*** (optional) - serve to narrow down semantic analysis around a certain topic or topics. For example, the keyword "landscape development" will only trigger analysis of the features of landscape development, but will omit anything regarding topics such as "climate change" or "labour migrants". Leaving keywords blank should first trigger topics extraction (see 5.1.2.3 "Topic Manager"), and then will mark out semantic features for each detected topic.

Output: ***Highlighted text***

5.1.2.5 Smart Points of Interest (SPOI)

Text Mining will also be used for updating of SPOI. There are four main ways how to use results from TM:

- New points of interest - this usage is possible, but not presumable, because all mandatory attributes and properties of new points (including coordinates) must be covered by the text.
- Additional information for existing points of interest - SPOI data model contains many optional properties, which could be mined from a text and added to a point of interest as a content of property (for example description, address or phone number) or as a link (for example link to photos, videos or textual documents).
- New links among points of interest - if a text contain information on neighbourhood or topology, they can be added to the SPOI as object relations. It applies to links to representation of the equivalent or similar point in an external LOD database[21], as well.

[20] In general case this can be any string, not necessarily a name of a geospatial entity. This will most probably produce unexpected results, but can still be useful for extraction of trends.

[21] https://lod-cloud.net/

- Updating of SPOI Ontology - outputs from text mining could be implemented above all for specification of existing classes (types of objects) in SPOI Ontology. Through text mining alternative labels or textual description could be discovered.

Inputs:

- ***Text*** (required)
- ***Geo-name*** OR ***geo-area*** (required) - subject area to search for POIs within. If a geo-name given, SPOI sub-module should first obtain a geo-area via API-call to Geo-parser.
- ***Language*** (optional)
- ***Keywords*** (optional) - serve to narrow down semantic analysis around a certain topic or topics. For example, the keyword "landscape development" will only trigger analysis of the features of landscape development, but will omit anything regarding topics such as "climate change" or "labour migrants". Leaving keywords blank should first trigger topics extraction (see section 5.1.2.3 "Topic Manager"), and then will mark out semantic features for each detected topic.

Output: Mixed (list of ***Geospatial Locations*** AND ***Summary***)

5.2 Storage

Storage is the Local Database, where the results of both streaming and requests from end-users will be saved for long-term goals.

The act of registering a chunk of data in a database (regardless of its type and destination) depends on the type of data on the Source Selection stage (described in section 3.2.2):

- Posts from Social Media come via socket streaming and are stored in the Local Database as described in the chapter 3.3 "Streaming from Social Media".
- Textual documents, uploaded by users are to be stored in the Local Database only if end-user selects this option after the processing of the source. It is supposed that enormous amount of texts will be analysed. Only those that provide valuable information for the foresight process, are going to be stored in the Local Database for future use.
- Textual documents of the dynamic nature (comments, discussion boards), processed by Crawlers, will only be stored partially: in the Storage for Indexed Data, but not in Raw Data. The reason for this is that such sources are being constantly updated.

5.2.1 Social Media

There are two separate (and parallel) storage schemes used for data collected from Social Media streaming:

- Raw data
- Normalized and Indexed data

5.2.1.1 Raw data

Raw data is represented "as-is" (i.e. with preservation of all information that comes through our feed - every document contain all fields). Raw data is used for additional analysis on the later stages of the project (for example, to collect all information from a certain area regarding a

chosen event or a policy). Raw data be periodically archived and deleted from the main database (time-frame to be defined in the project settings, and will be fine-tuned regarding the amount of data).

Warning: the access to Raw Data will not be given to end-users, but to administrators only.

Storage scheme: MongoDB cluster.
Document type: unstructured. Each document will consist of the attributes and the values of the types, as received from the source via socket streaming. It is a responsibility of a developer to know the structure of data and DSL. The only common attributes are the datetime of registering and the source name ("twitter", "linkedin", etc.)

5.2.1.2 Normalized and Indexed data

Normalized and indexed data is a product of Normalization, as described in section 3.3.4 "Data Normalization" as well as section 5.1.1 "Text Preprocessing".

Warning: documents discarded at the stage of Normalization won't make it to Indexing and will only be available in their raw form.

Storage scheme: Elasticsearch cluster.

Document type: semi-structured after Normalization, ready to be searched by specified field-names and/or filtered as described in 5.3.5 "Queries in Domain Specific Language (DSL)".

5.2.2 Textual Documents

In the same manner as posts from Social Media, textual documents will also be stored in the Raw data storage as well as in the database for Normalized and Indexed data. However, due to the nature of the content and a difference in the stage of Source Selection, the proportion of texts stored in the Raw Data collection will be minuscule.

5.2.2.1 Raw data

In the Local Database only those textual documents will be stored in a raw format that satisfy the following requirements:

- the document contains static information (see 3.2.2 "Source Selection")
- end-user decides to save it in a raw form after the document is processed (if it doesn't contain valuable info, there is no reason to increase the amount of information stored).

Storage scheme: Elasticsearch cluster[22].

5.2.2.2 Normalized and Indexed data

This type of storage is used to save the result of semantic analysis of a textual document. This result is a dataset that consists of pieces of information of different types, depending on the results of modules in the Semantic Explorer Pipeline (5.1).

[22] Instead of MongoDB (as in the case of Social Media) Elasticsearch is chosen as a storage scheme for raw texts due to the fact that they should be available to end-users (in the current project MongoDB cluster is supposed to be a data source available for technicians only).

Common data - results of the Text Preprocessing stage:

- tokens (without stopwords)
- grammatical instances
- N-grams
- TF/IDF for each token and each N-gram

Depending on the service selected by the end-user, each text document will also contain *at least one* of the following attributes:

- Named entities
- Geo-spatial information for each geo-name in the list of Named entities
- Topics and subtopics in the form required for the creation of Similarity Cluster (3.1.4)
- Topics and subtopics in the form required for the creation of Shared Term Spaces (3.1.5)
- Links to profiles - the result of Profiler (5.1.2.4)
- Links to Smart Points of Interest (SPOI) (5.1.2.5)

5.2.3 Object Diagram

See Figure 9 for the combined scheme of the storage of indexed documents (both Posts from Social Media and Textual Documents). If required, more objects can be added on the stage of development.

In addition to that, three more storage schemes will be supported[23]:

- Raw data: ***MongoDB*** (two collections - one for Social Media and another for Texts)
- Keywords for filtering posts from Social Media: ***Redis***
- User profiles and admin data: ***PostgreSQL***

[23] The schemes of those are not included in this document, because they are either a part of a standard functionality of selected packages (and therefore do not require modelling) or too simple to be represented as a drawing (e.g., single collection or set of keys).

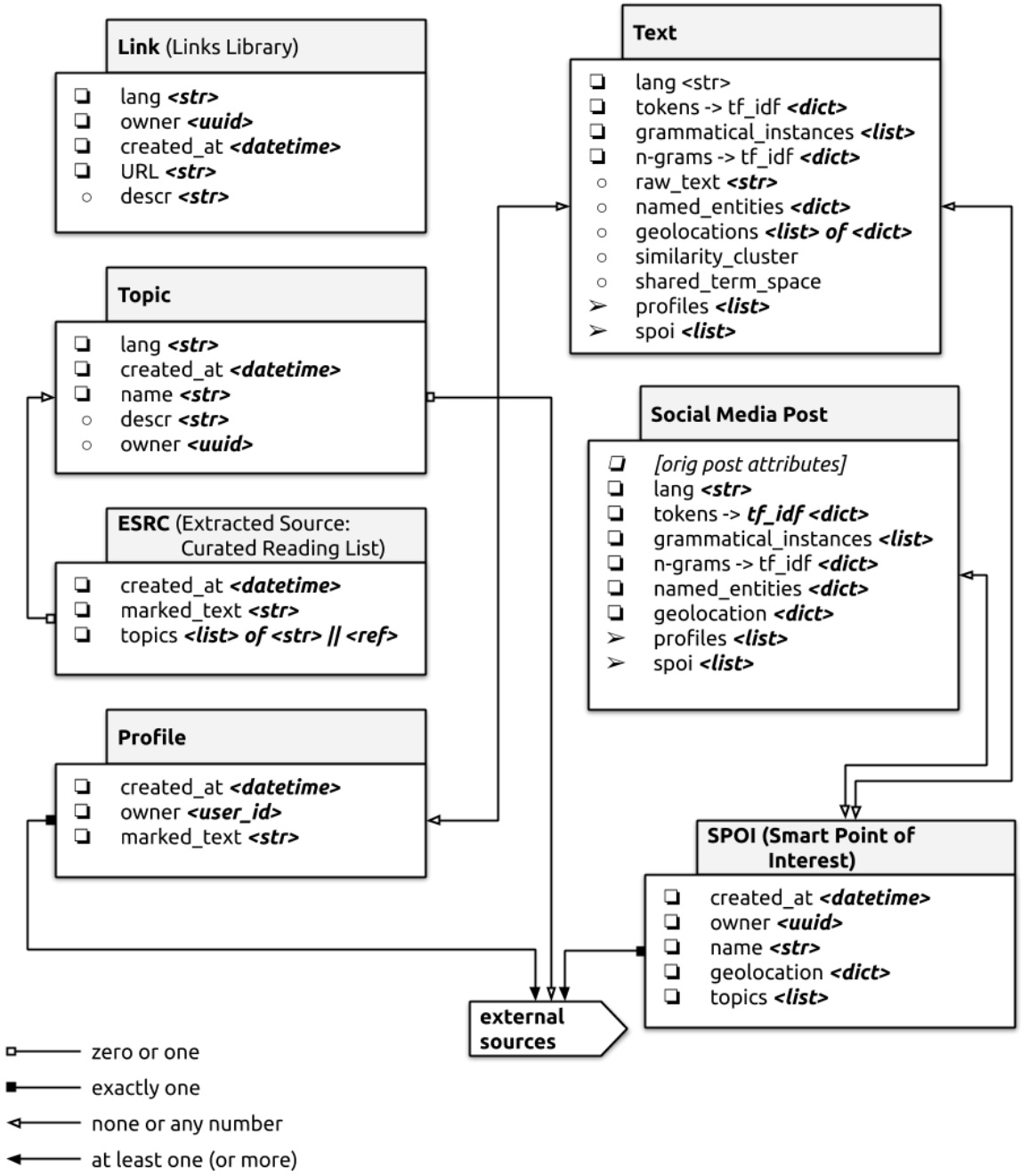

Figure 9 Indexed Storage Scheme

5.3 Access to Data (API)

Access to data should be provided by RESTful API[24]. API endpoints should be available via an HTTP requests. They may either simply return data (queried by GET request) or perform some operations on the data and then return results (POST, PUT and PATCH requests). DELETE requests will not be available to end-users, but to technical staff only.

When the end-user queries some data from the web-framework, the stage of Source Selection (3.2.2) in the Text Mining Pipeline means choosing a proper endpoint and a request method.

[24] https://restfulapi.net/

5.3.1 Unified Inputs for Semantic Explorer

Input is a list parameters of the POST or GET request to API.

Possible inputs:

- ***Text*** - one of the options below:
  - real text sent by user via API request (POST)
  - URL or collection of URLs from either Links Library (3.1.8) or Curated Reading List (3.1.9) - in this case text is extracted from online documents by a dedicated crawler (see section 4.4 "Crawlers and Storage").
- ***Language*** (optional) - if not provided, should be automatically recognized by a dedicated service.
- ***Geo-name*** (optional) - for example, "Flanders, Belgium" or, alternatively, geospatial area, defined as described in 3.1.7 "Geospatial Location".
- ***Keywords*** (optional) - list of keywords, that affects semantic analysis of a given text.

5.3.2 Semantic Explorer Textual Outputs

Type of output depends on a user's request and endpoint. Below are the possible types of output.

5.3.2.1 Summary

Used for *topic(s)* and/or *subtopic(s)*.
Format: JSON
Example:

```
[
    {
        "topic": "landscape development",
        "subtopics": [
            "landscape quality",
            "landscape dynamics",
            "landscape maintenance",
            "landscape architecture",
            "historical landscapes and heritage"
        ]
    },
    {
        "topic": "business model",
        "subtopics": [
            "new farming business model",
            "rural planning and modeling"
        ]
    }
]
```

5.3.2.2 Highlighted text

Format: JSON
Example:

```
{
```

```
            "type": "sgml",
            "language": "en",
            "uri":
        "https://en.wikipedia.org/wiki/Agriculture_in_Flanders",
            "highlights": "
                <GEO>
                    The Flemish countryside
                </GEO>
                is a
                <DEF>
                    highly urbanized
                </DEF> area.
                It is not only
                <DEF>
                    characterized by a high population density
                </DEF>
                and a
                <DEF>
                    highly fragmented landscape
                </DEF>
                , quite often the relationship between the countryside
        and the urban environment is also very strong. The countryside
        and the urban environment are
                <DEF>
                    <DYN>
                        becoming increasingly
                    </DYN>
                    intertwined
                </DEF>
                in a geographical, practical and cultural way. The
        present communication and transport facilities
                <DEF>
                    favour economic, social and cultural interaction
                </DEF>
                ..."
            }
        }
```

where <GEO> is a **geo-name**, <DEF> is a **definition**, <DYN> is characteristics of **dynamics**.[25]

For a better understanding of how this output can be represented to the end-user in a more readable form, see 6.2 "Textual Output".

5.3.2.3 Geospatial Location

Format: GeoJSON

Example (note that the key "features" can include more than one place - this is the case, if Geo-parser finds more than one geo-name in a given text):

[25] The tags here are given as examples. The exact list of tag names and their semantics will be defined after the model training, and can vary for language models.

```
{
    "type": "FeatureCollection",
    "crs": {
        "type": "name",
        "properties": {
            "name": "EPSG:4326"
        }
    },
    "features": [
        {
            "properties": {
                "placetype": "locality",
                "name": "Bytča",
                "belongsto": [
                    "Europe",
                    "Slovakia",
                    "Žilinský kraj",
                    "Bytca"
                ],
                "region": "Bytca",
                "country": "Slovakia",
                "place": "Bytča",
                "admin_region": "Zilna",
                "admin_region_id": "g2008_1.11242"
            },
            "geometry": {
                "type": "point"[26],
                "coordinates": [
                    18.561567,
                    49.227148
                ]
            }
        }
    ]
}
```


5.3.2.4 Mixed output

A response from API endpoint can combine several types of output: for example, a request to the endpoint for extracting topics can be accompanied by a place name - in this case in addition to topics and subtopics a response will contain geospatial location.

5.3.3 API endpoints

In this section all available API points are listed along with their HTTP request methods.

[26] NB: geo-point is given here for simplicity, but other types of geo-spatial info are possible, such as geohash, polygon and multipolygon.

5.3.3.1 List of Posts from Social Media

URI: `/api/sm/`

Methods:

- **POST** - register a post. This request is mainly being sent by the Normalizer after receiving and normalizing a post from SM platform.
- **GET** - filter posts by parameters, such as geo-location, date of creation, keywords, topics, platform (such as "twitter" or "linkedin"), etc.
- **PATCH** - this method is used after the analysis by a new topic, when there is a need to update a queryset with a measure of the relevance to a new topic. Should only be called internally after the POST request to Topic Extraction (see below), when applied to queryset from Social Media (i.e. the necessity for update should be hidden from the end-user)
- **DELETE** - internal request from periodic procedures (for example, in the case of deleting posts with Direct Quotes). Not available for end-users.

5.3.3.2 Social Media Content by ID

URI: `/api/sm/<post_id>/`

Methods:

- **GET** - simply return the content of the post
- **POST** - N/A
- **PATCH** - the same as PATCH in `/api/sm/`, but applied to a single post
- **DELETE** - internal request, N/A for end-users

5.3.3.3 List Text Documents from Links Library

URI: `/api/text/`

Methods:

- **GET** - returns list of a textual document by providing filters (keywords, URLs, dates of registering in the DB, usernames of creators, languages, etc.). This will not return content or semantic features of selected documents, but only metadata (such as ids and URLs in the Links Library).
- **POST** - register in the Local Database a new document and the results of its Preprocessing
- **PATCH** - N/A

5.3.3.4 Textual Document Content by ID

URI: `/api/text/<document_id>/`

Methods:

- **GET** - return document content (if available). Depending on the nature of information of the document the following will be returned:
    a. static - document meta-data, content (raw text) as well as all the results of Text Analysis performed by the moment of request
    b. dynamic - only meta-data
    c. mixed - only meta-data and the results of Text Analysis (but no content)
- **POST** - N/A
- **PATCH** - update document's properties, given in the request body
- **DELETE** - deletes a document from Links Library and all its data (available only for a user that created it)

5.3.3.5 Curated Reading List (CRL)

URI: `/api/crl/`

Methods:

- **GET** - return list of links from the Curated Reading List by providing filters (keywords, URLs, dates of registering in the DB, topics, languages, etc.).
- **POST** - N/A
- **PATCH** - N/A
- **DELETE** - N/A

5.3.3.6 Text Preprocessing

URI: `/api/preproc/`

Methods:

- **GET** - N/A
- **POST** - performs Preprocessing, calls `/api/text/` internally to store the document, returns the results of Preprocessing along with document meta-data
- **PATCH** - equal to restart Preprocessing and update saved document with new results
- **DELETE** - N/A

5.3.3.7 Extract Named Entities from Text

URI: `/api/ner/`

Methods:

- **GET** - N/A
- **POST** - performs NER (note that it calls Text Preprocessing (5.3.3.6) endpoint internally), then POSTs to `/api/text/` internally to store the document, finally returns the results of Preprocessing along with document meta-data
- **PATCH** - equal to restart NER and update saved document with new results
- **DELETE** - N/A

5.3.3.8 Geo-parse Text

URI: `/api/geoparse/`

Methods: similar to Extract Named Entities from Text (5.3.3.7), but in the POST pipeline it also calls Geoparser and returns its results as geo-spatial data (see 5.3.2.3 "Geospatial Location").

5.3.3.9 Extract Topics from Text (Similarity Cluster)

URI: `/api/simclust/`

Methods: similar to Extract Named Entities from Text, but in the POST pipeline it also calls Topic Extraction and returns its results as Summary (see Summary).

5.3.3.10 Topics Modelling endpoint (Shared Term Space, STS)

URI: `/api/sts/`

Methods: similar to Extract Topics from Text, but in the POST pipeline it also calls Topic Modelling and returns its results as Summary (5.3.2.1).

5.3.3.11 Profiler endpoint

URI: `/api/profiler/`

Methods: similar to Extract Topics from Text (5.3.3.9), but in the POST pipeline it first Geo-parses

Text (as described in 5.3.3.8) and then Extracts Topics from Text (5.3.3.9). Returns results as Highlighted Text (see 5.3.2.2 "Highlighted Text").

5.3.3.12 SPOI endpoint

URI: `/api/spoi/`

Methods:

- **GET** - returns list of SPOIs discovered in texts and saved in the Local Database.
- **POST** - similar to Profiler endpoint (5.3.3.11), but returns GeoJSON (see 5.3.2.3 "Geospatial Location").
- **PATCH** - N/A
- **DELETE** - internal request, N/A for end-users

5.3.3.13 SPOI by ID

URI: `/api/spoi/<point_id>/`

Methods:

- **GET** - returns content for a chosen SPOI in GeoJSON (see 5.3.2.3 "Geospatial Location") along with sources (IDs of the documents or/and request to SM data).
- **POST** - similar to Profiler endpoint (5.3.3.11), but returns GeoJSON (see 5.3.2.3 "Geospatial Location").
- **PATCH** - first performs **GET** to collect sources, and then re-writes the SPOI's information (for example, some documents with dynamic information does not contain data about current SPOI - such documents should be removed from the list of sources)
- **DELETE** - deletes a SPOI and all its data by <point_id> (available only for a user that created it)

5.3.4 Parameters of API calls

The set and format of the parameters depends on the type of information that the end-user requests and the performed action (extract or affect data). Below is a brief description of each type of request and corresponding parameters.

5.3.4.1 GET requests

General format for filtering data through API endpoint:

```
/api/resource_name/?param1_name=param1_value&param2_name=param2_value&...&paramN_name=paramN_value
```

Parameters serve as filters and allow for the use of modifiers. Each modifier can be applied to a field of a certain type:

- `exact` <universal> - equality
- `iexact` <string> - equality, case insensitive
- `exists` <universal>
- `startswith` <string>
- `istartswith` <string> - case insensitive "startswith"
- `endswith` <string>
- `iendswith` <string> - case insensitive "endswith"
- `contains` <string>

- `icontains` <string> case insensitive "contains"
- `match` <string> - matching pattern (can include *, e.g. racoon*)
- `in` <list> - inclusion
- `nin` <list> - "not in"
- `lt` <number>, <date>, <timestamp> - "less than"
- lte <number>, <date>, <timestamp> - "less than or equal"
- `gt` <number>, <date>, <timestamp> - "greater than"
- `gte` <number>, <date>, <timestamp> - "greater than or equal"
- ne <number>, <date>, <timestamp> - "not equal"

Format: `paramname__modifier=value`
Examples:

```
...&keyword__startswith=rural
...&description__contains=Controls
...&created_at__lte=2019-09-05T10:00:00
```

The full list of available filters and their modifiers should be available in the schema of each endpoint.

Common parameters:

- `format` - available formats: xml, json, yaml
- `username` - not a param, but a part of authentication token (together with "api_key"). NB: this can also be sent as Authorization header.
- `api_key` - a part of authentication token (together with "username"). NB: this can also be sent as Authorization header.
- `limit` - limits the number of objects returned. Applicable only in case of detailed reports. Default: 36. Example: ...&limit=100
- `offset` - number of records to skip from the beginning. Together with "limit" is used to divide data to pages (paginators). Default: 20. Example: ...&offset=40
- `order_by` - sorting by multiple fields
- `search` - searching a term in a text

5.3.4.2 POST / PATCH / DELETE requests

This type of requests requires filling up `body` parameter of the API call. Below is an example of using this kind of request via `curl` command:

```
curl -X POST \
  https://endpoint.uri/ \
  -H 'Content-Type: application/json' \
  -d '{
      "place": "Voerstreek",
      "lang": "en"
}'
```

5.3.5 Queries in Domain Specific Language (DSL)

A domain-specific language (DSL) is a computer language specialized to a particular application domain. This is in contrast to a general-purpose language (GPL), which is broadly applicable across domains.

Queries over HTTP should be translated to DSL specific to the requested storage. The end-user should have a unified set of params and a freedom in combining them regardless of the access point requested (Social Media posts, or the Links Library, or a Curated Reading List). Therefore, a sub-component for the translation from user-defined params to DSL should be a part of all components that provide access to Local Databases.

6 Presentation Layer

Presentation and interpretation of the results should take the form of web-pages available for registered users.

Warning: The Presentation Layer (front-end) will NOT be a part of the Semantic Explorer prototype. The only output planned is a set of API endpoints in the JSON format. Graphical and markup outputs are mentioned here only for a better understanding of how the results of Text Mining solution can be interpreted. The actual front-end will come on the later stages of the project, and will be based on the specifications co-designed with the end-users.

6.1 Graphical Output

Depending on the request from the end-user, different types of graphical outputs are possible.

6.1.1 Similarity clusters

Similarity clusters (see 3.1.4 "Similarity Cluster") are graph structures that describe similarity of user-defined "root concept" to the concepts, extracted from a given set of documents. To build a diagram of this class the end-user should define a set of documents, from which terms are to be extracted, as well as the depth for the graph, i.e. number of layers (each layer is being built around a node, which in this case contains a "root term" for a sub-node).

Used for topic modeling (see Figure 10).

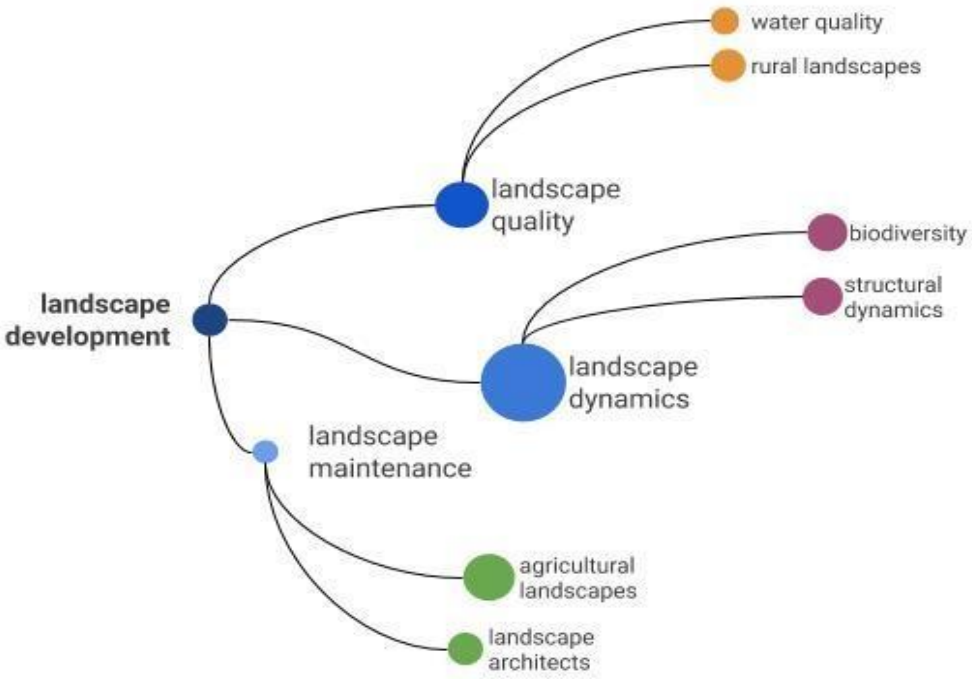

Figure 10 Example of Similarity clusters[27]

[27] https://en.wikipedia.org/wiki/JSON

4.1.1 Evaluation Diagram

Evaluation Diagram (see Figure 11) explains "public voice" on a certain term. This type of graphics uses a certain term for analysis and scans Social Media content for a given time period to obtain polarity (on a scale from -1 to 1, where -1 is extreme negative, 0 is neutral, and 1 is extreme positive)[28], as well as user's opinion in a meaningful sentence (see 3.3.4.1 "Sentiment Analysis").

Used for better representation of Polarity analysis of the opinions of stakeholders regarding certain Policy.

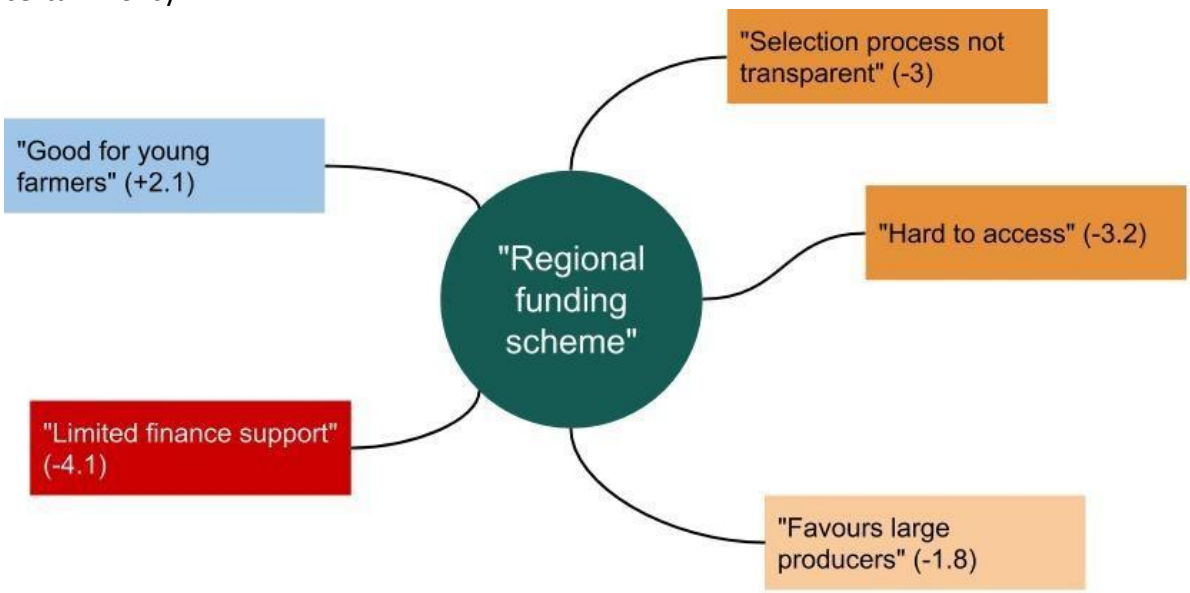

Figure 11 Example of Evaluation Diagram

6.2 Textual Output

Highlighted text can better be understood if represented in the following form:

"The Flemish countryside is a highly urbanized area. It is not only characterized by a high population density and a highly fragmented landscape, quite often the relationship between the countryside and the urban environment is also very strong. The countryside and the urban environment are becoming increasingly intertwined in a geographical, practical and cultural way. The present communication and transport facilities favour economic, social and cultural interaction..."

[28] In this kind of diagrams negative responses will prevail, since in social networks negative opinions about any subject dominates over positive. Therefore this kind of diagrams, when applied to data obtained from Twitter, Facebook and the like, will rather indicate problems, than successes on any given subject.

6.3 Geospatial Representation (Maps)

All API endpoints that return data from Social Media, can be represented on the maps in the different forms, the most used of which will be **Geospatial Heatmaps**.

Heatmap (see Figure 12) is a representation of data in which values are represented by colours painted over a chosen area on the geographical map[29]. A simple heatmap provides an immediate visual summary of information, e.g. number of tweets in a certain geohash. More elaborate heatmaps allow the viewer to understand complex datasets, such as a standard deviation of opinion polarity over a certain region as calculated by certain topic (for example, chosen Policies).

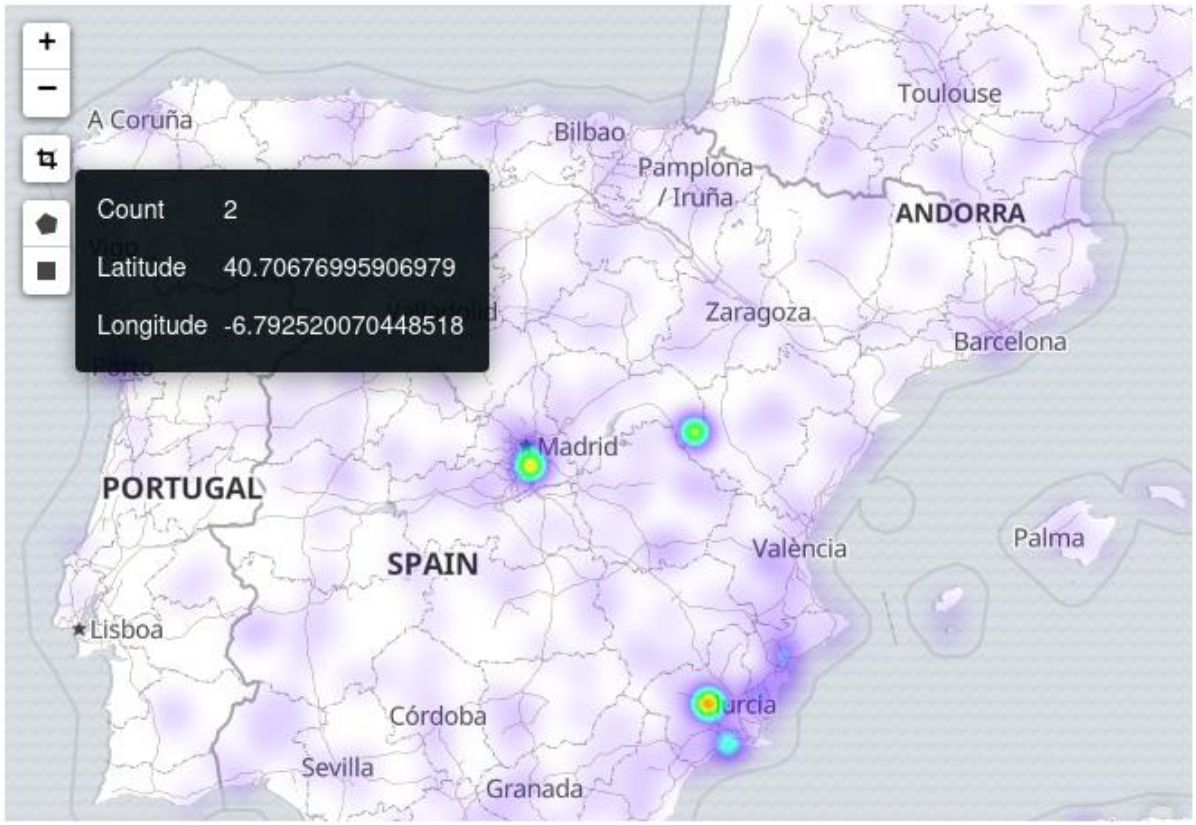

Figure 12 Example of Heatmap

[29] Please, note, that the term "heatmap" refers to a more general way of representing values as colours, painted on a 2-D surface, not necessarily a geographical map. But for the purposes of the current project we use the term "heatmap" in the geographical context.

7 Infrastructure

7.1 Infrastructure Planning Guidelines

The guidelines for the planning of infrastructure of the Text Mining solution (Figure 13) lay down the general principles for composing the System Components diagram and the Deployment diagram found below.

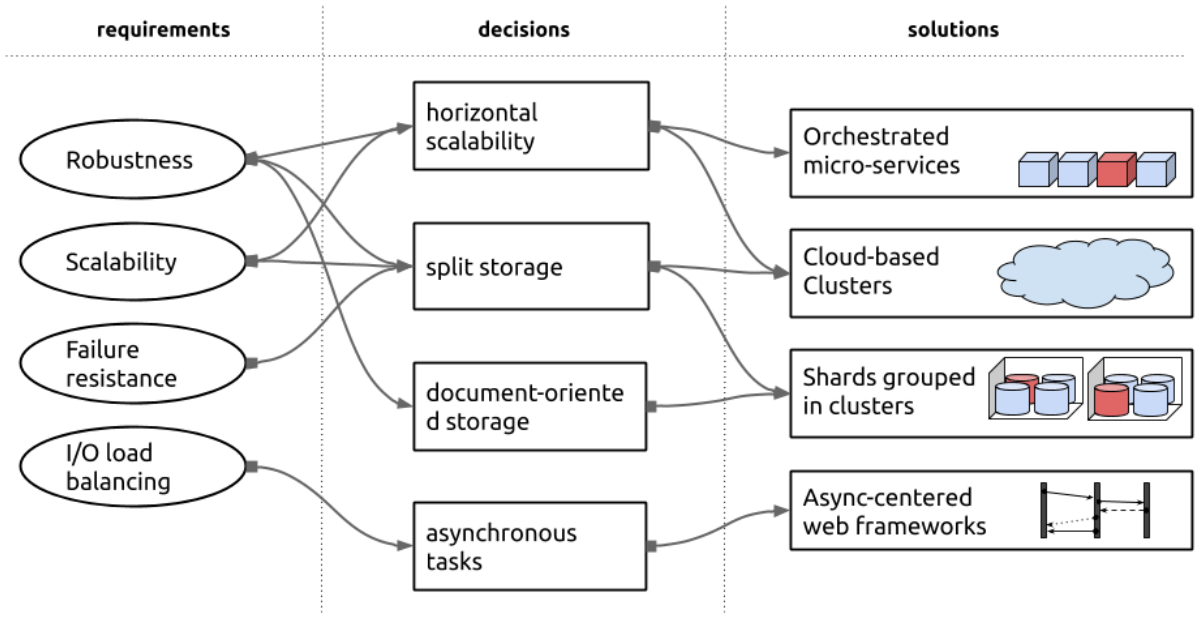

Figure 13 Infrastructure Planning Guidelines

7.2 System Components

A generalized diagram of the system components of Text Mining solution is presented in Figure 14.

The diagram serves as a guide for the process of deployment, where several containers can either be deployed on the same Virtual Server (VS) or separate (for horizontal scaling). The diagram shows both provided and required interfaces between external components (request-response). Internal connections between components of the same service can be updated on demand and are not shown in the diagram. Examples of internal requests are "read", "write", "update", "delete", etc.

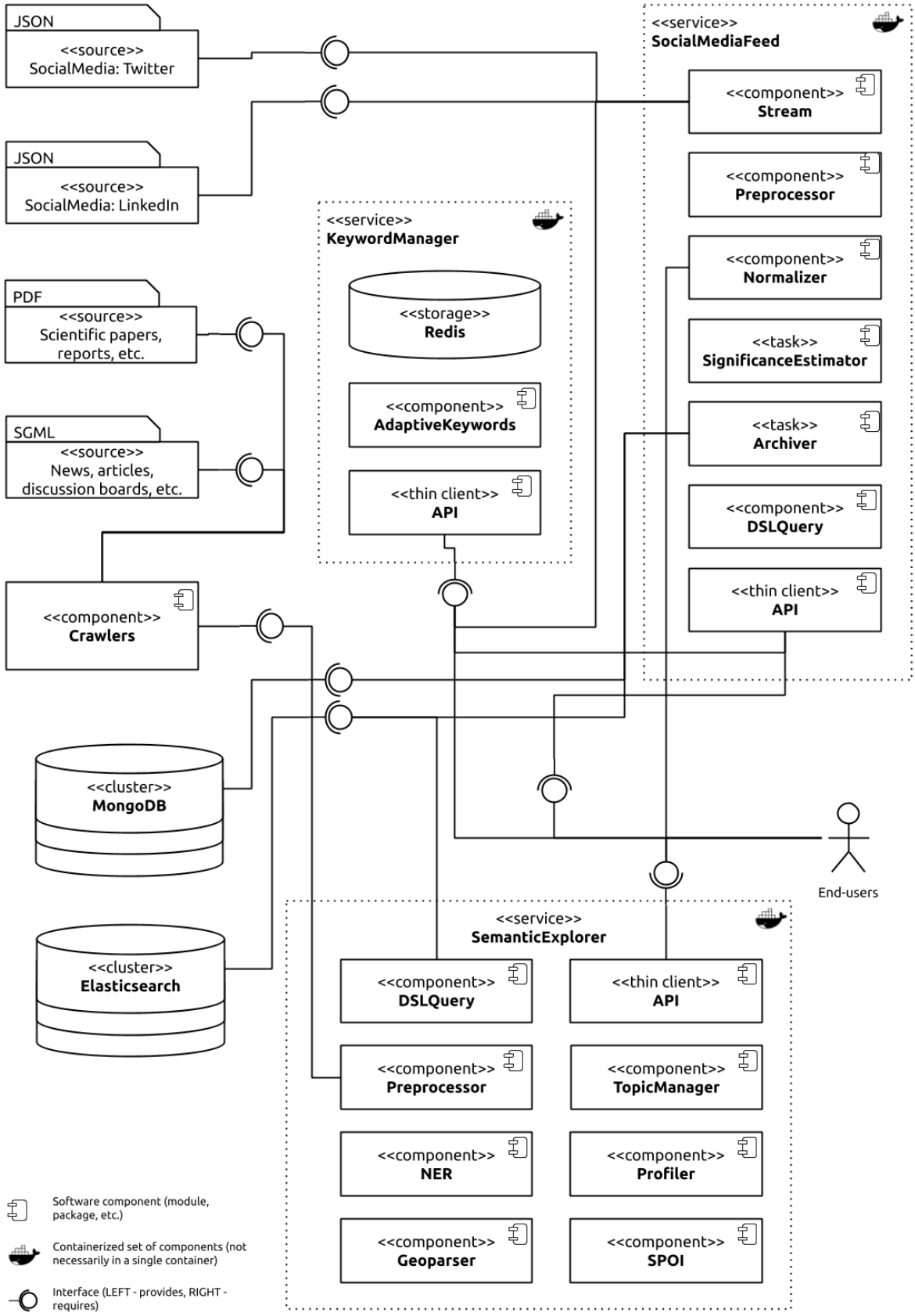

Figure 14 System Components diagram of the Text Mining solution

7.3 Deployment

7.3.1 Technological stack

For the realization of the services and facilities presented in the System Components diagram (section 7.2, Figure 13) a set of available open source solutions will be used (see Figure 15).

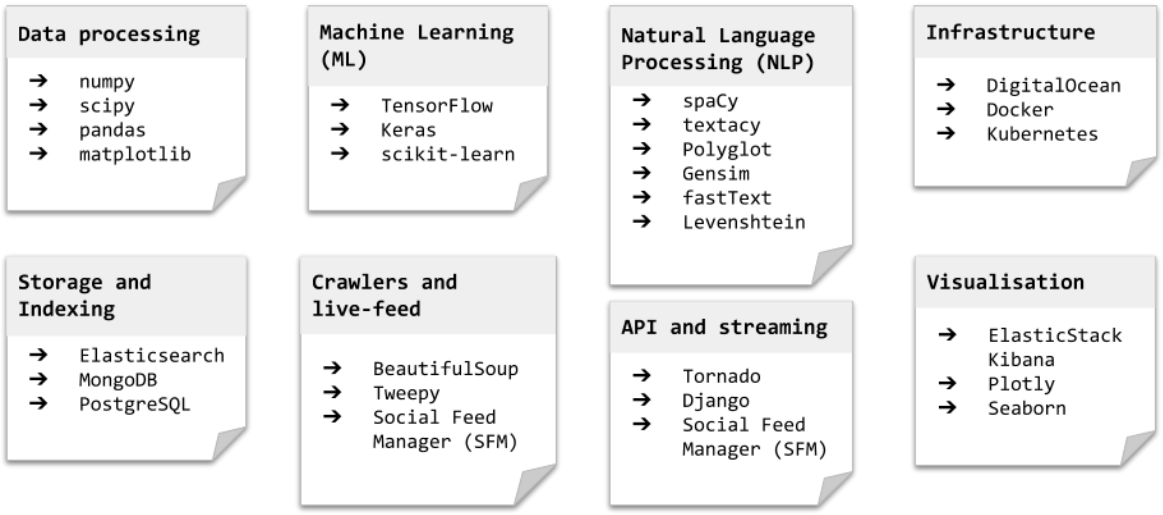

Figure 15 Technological stack

7.3.2 System Architecture and Deployment

The final stage of the development of the Text Mining solution is its deployment and making it available through the web. The platform chosen for the deployment is DigitalOcean[30] for being ready for deployment of applications with distributed architecture. It allows for easy horizontal scalability, while being considerably inexpensive as compared to its competitors on the market.

The requirement of creating a set of microservices defines architecture as a set of headless services, communicating with each other via API. The communication between services is asynchronous, based on the idea of the message queue with allowed timeouts, whereas each of those services should be available to each other and to the end-user. For example, a user can request for named entities in a text, for which NER (of Semantic Explorer) is responsible. At the same time another service - Profiler - requires results of NER to be able to return its own results to the end-user. Therefore, services should be separated ("containerized") and made available both in the operational time, as well as on demand. This requires "orchestration" software, for which Kubernetes (k8s) has been chosen[31].

At the same time taking into consideration heavy use of databases and trained language models, a storage should be allocated that survives Pod and Node restarts - Stateful sets.

The diagram (Figure 16) displays the product's infrastructure deployed on the chosen platform

[30] https://www.digitalocean.com/

[31] https://www.digitalocean.com/products/kubernetes/

(DigitalOcean) as a set of k8s nodes and stateful sets.

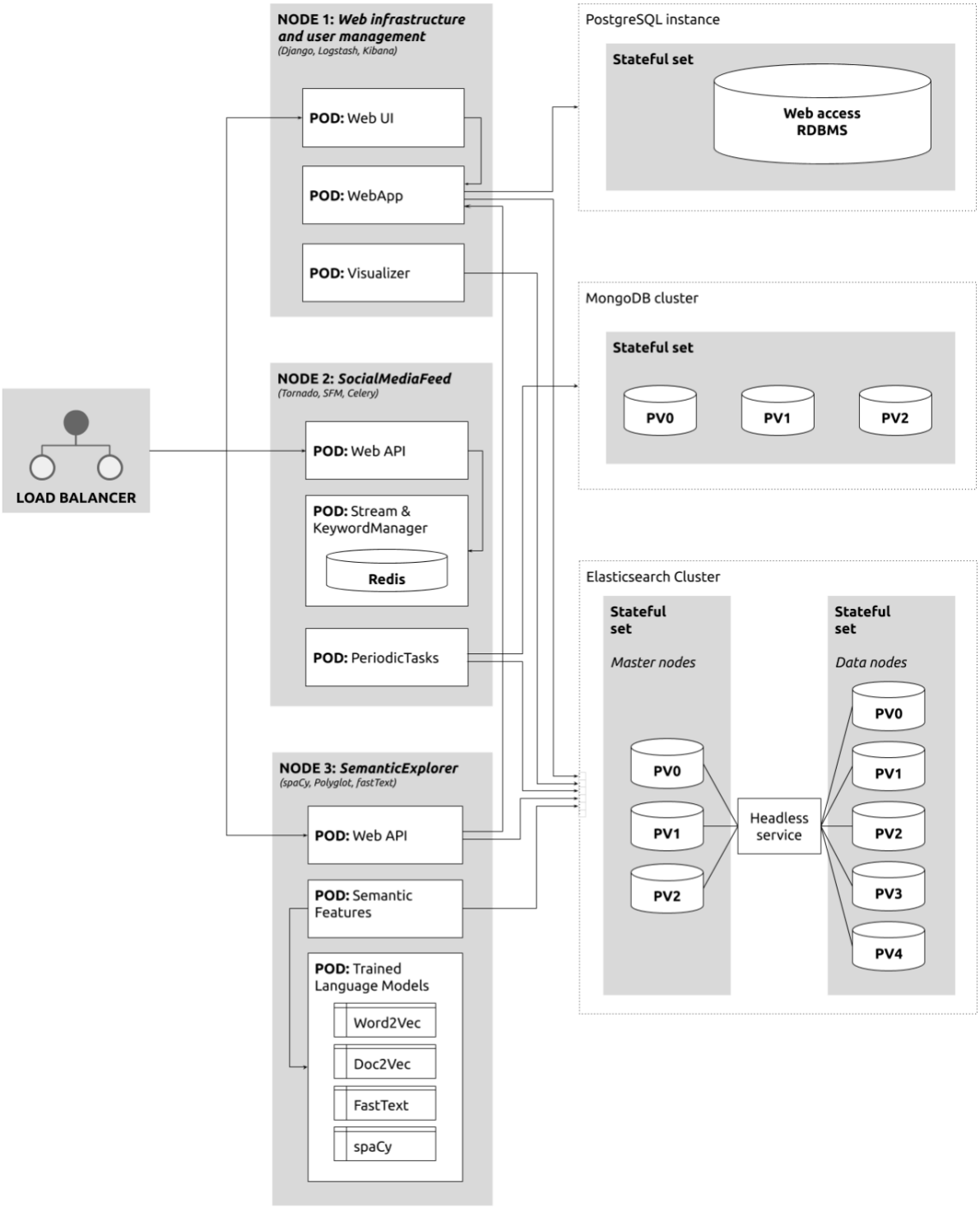

Figure 16 Infrastructure on the platform of DigitalOcean

7.3.3 Development Guidelines

API endpoints should follow the RESTful API and Event Scheme Guidelines[32] (Fielding, 2000).

[32] See "A model set of guidelines for RESTful APIs and Events" https://github.com/zalando/restful-api-guidelines

Everything should be covered by tests:

- Functional tests
- Unit-tests

The two environments should be configured and run in parallel:

- **Testing environment** deployed on an arbitrary platform with Jenkins[33] as an automation server.
- **Production environment** deployed on DigitalOcean as presented in figure 16).

7.3.4 Development Pipeline

All the code is stored in the dedicated repositories on Github[34]. During the development process *only two* branches of each repository should be maintained on Github: master and dev. Feature branches are allowed in the local repos only!

The development pipeline starts with a commit to the branch dev. After each commit there should be an automatic deployment to the Testing environment, where Jenkins runs tests. If at least one test fails, Jenkins will send emails to the author of the commit and lead developer.

Once in a while[35] changes in dev are merged into master branch, and deployed in a Production environment. In case of a "hot fix" this can also be done on demand. The deployment in the Production environment is possible only if all tests pass!

[33] https://jenkins.io/

[34] https://github.com/

[35] In the course of the development the dev team and partners of the Project should agree on a sensible period of time for updating the Production environment.

8 Challenges

8.1 Multilingual inputs

The prototype solution should be built around documents in English and on the later stages of the project adapted to regional contexts and languages. This is a challenging task, taking into consideration linguistic differences (such as word order, use of numbers and quantities, date, time and currency formats, amount of declensions, complexities of conjugations, etc.) as well as more social and cultural characteristics (e.g. compliance with cultural dynamics, number of ways to express opinion polarity, etc.).

Therefore, there is a certain risk of producing a discrepancy in the accuracy of the results obtained for different languages. There should also be added purely statistical nature of the semantic models that will be trained to perform semantic analysis - i.e. there is always a margin of error in producing results of the semantic analysis in even the same language, once a new sentence or a phrase is added to already processed documents.

There are two ways we can tackle this problem:

- gathering more labelled data with the help of experts and / or by organizing a citizen science campaign[36].
- use automated translations to train semantic models (this requires *real big* labelled datasets in original language to transfer all the shades of meanings to a new language).

8.2 Semantic Discrepancies

Regardless of the results of model training, there will always be discrepancies in the human understanding of certain concepts and a machine estimation of those concepts, being expression of certain topics. For example, a Profile of a pilot area can include KPIs that the end-user does not consider as important, while the Semantic Explorer highlights them, because those indicators were identified after the topic selection.

This problem is the one that cannot be solved mechanistically, but only empirically - by the iterative process of choosing better topics and keywords.

In its turn it lays down an additional load on the servers (iterative process mean continuous requests for performing some operations), which leads to the challenge described in section 9.4 "Monitoring and Maintenance".

8.3 Comparison of Opinion Polarity

While comparing answers of surveys with the opinion polarity of in data from Social Media is possible, it contains a certain risk of being inaccurate, because the expression of dissatisfaction in Social Media is much more common than satisfaction. Statistically (and empirically) it is much

[36] See "Citizen Science Platform" https://digitalearthlab.jrc.ec.europa.eu/activities/citizen-science-platform-%E2%80%93-tool-extend-evidence-base-policy/57787

more probable to hit a positive post in SM, that describes personal life (family, vacation, and the like) rather than social (policies, access to funds, etc.). In SM data posts with negative opinion about social services and governmental actions are much more common. This is not the case when a person is being asked directly. Therefore, there can be a discrepancy of the average opinion as estimated by Sentiment Analysis in comparison with the results of surveys.

8.4 Monitoring and Maintenance

It is necessary to mention that the creation of the Text Mining solution does not end with the delivery of the finalized product. It is an iterative process not only in terms of its development, but also in its usage. Information that is to be analysed, processed and stored is expected to be massive in amounts and complexity. Therefore, the goal is to create a solution that satisfies the following requirements:

- should be agnostic to the amounts of information to keep timeouts as low as possible when providing access to the results of data processing.
- should be easily (horizontally) scalable for ingesting information from different sources and in different formats, while providing universal access to it.

This requires not only careful infrastructure planning (which is presented in the section 8 "Infrastructure"), but also understanding of the fact that the solution will need constant monitoring and maintenance. Once being deployed in the production, the product should not only help in solving problems that lead to its creation, but also to keep up with the dynamic nature of the information in the web and an increasing complexity of its processing.

9 Conclusions

The extraction of meaningful information from texts available online remains one of the most challenging tasks, as well as one of the most requested. It is of increased importance to policy making, impact assessment and scientific research. The creation of the Text Mining solution within the PoliRural project will allow experts to be aware of local challenges and needs, while policy makers will benefit from it by having more accurate and timely information at almost every stage of the decision-making process.

In particular Text Mining solution is being created to assist in the following tasks:

- *Horizon scanning* serves a purpose of identifying relevant trends that may impact the region. The result of it is a list of topics crucial in the analysis of a certain pilot. The components Topic Manager (see 5.1.2.3 "Topic Manager") address this task.
- *Issues and needs analysis* should capture well-known or recognised needs that are already mentioned in policy documents, as well as anticipate future local needs based on the dynamics of regional development. Profiler sub-component of Semantic Explorer (see 5.1.2.4 "Profiler") is designed to solve this task.
- *Pilot profiling (for SDM)* aims at understanding the normal course of action and identifying trends or behaviours over time when modelling with local agents or communities. The outputs from the Profiler sub-component of Semantic Explorer (see 5.1.2.4 "Profiler") are expected to help in this process by creating pilot profiles.
- *Being up-to-date* normally involves reading a lot of sources, being subscribed to many groups in social media and following discussion boards in order to being able to spot hot topics and trends. A submodule for creation and updating Curated Reading Lists helps to stay up-to-date by extracting relevant and timely information from the constantly updated sources (see 3.4 "Curated Reading List Management").

It is necessary to mention that the creation of the Text Mining solution does not end with the delivery of the finalized product. It is an iterative process not only in terms of its development, but also in its usage. Information that is to be analysed, processed and stored is expected to be massive in amounts and complexity. Therefore, the goal is to create a solution that satisfies the following requirements:

- should be agnostic to the amounts of information to keep timeouts as low as possible when providing access to the results of data processing.
- should be easily (horizontally) scalable for ingesting information from different sources and in different formats, while providing universal access to it.

This requires not only careful infrastructure planning (which is presented in the section 8 "Infrastructure"), but also understanding of the fact that the solution will need constant monitoring and maintenance. Once being deployed in the production, the product should not only help in solving problems that lead to its creation, but also to keep up with the dynamic nature of the information in the web and an increasing complexity of its processing.

The system will be provided as SaaS (Software as a Service) and will be available online as a web application for the partners of PoliRural project. The input from experts will help to guide future developments and a scheme for maintenance. The value of the system should be demonstrated through a number of use cases, and this type of semantic analysis of the texts available online is intended to be of practical use to policy analysts and policy makers.

10 Annex I - Initial Text-Mining Use-Case Scenarios (CKA)

Text-Mining or TM can be seen as a form of "smart search." It has the potential to reduce help researchers involved in preparing inputs to the various activities of the Foresight workshop. In this sense it is a tool in the hands of researchers preparing background material for use by Foresight practitioners.

Use Case 1: TM Support to Initial Exploratory Work

The first thing to address in a Foresight initiative is the range of challenges to be considered and eventually ranked in terms of urgency or importance. The Foresight team should spend some time exploring this issue. The result of this work is to make sure that the group does not overlook important issues that may need attention, but which are currently being ignored.

This work is often loosely referred to as "issues analysis." The main tools at this stage are:

- I**nformal Consultations** with local stakeholders intended to capture what they see as the important issues to address in the future. These may refer to un-addressed internal issues, perceived gaps in local policy or flaws in the system and the way policy is implemented. This process is rather unstructured and chaotic. It provides for the capturing of insights in an ad-hoc or opportunistic way, with no agenda and no constraints, just an open mind. Nevertheless, someone must take on board the task of putting a structure on the findings. Grouping them and characterizing them in a useful way. This provides some scope for creativity. It is important that all inputs and the voices of all stakeholders are reflected in this.

- **Literature Surveys** that look more generally at what is happening in the world. The main technique is to look at opportunities and threats that are currently unfolding elsewhere, on the basis that some of these may one day become relevant locally.

At this stage it appears that it is in carrying out the literature survey that TM has the greatest potential to support the FS process in this early exploratory stage of the work. There is a global literature on trends of different kinds. These trends can be seen as enablers or drivers of change. Many research groups and institutions have compiled lists of trends, mega-trends, macro-trends and micro-trends. It is useful to scan these to see if the factors they refer to might be of relevance to the FS activity at hand. These are often accompanied by visions and scenarios explaining how these will change the world. They are usually high level in nature and not specific to any given place. Further work will need to be done to make sense of them, but for now they provide food for thought and starting point for strategic conversations about the future.

Making lists is one of the basic tasks in a Foresight activity. This initial exploratory work should result in an initial list of "issues" or "challenges" that may have an impact on the regional in the coming years, and which may require a policy response. They could be classified under headings such as "opportunities" and "threats." This material needs to be tamed by the foresight team. It needs to be structured in such a way that it can be used to support initial group-work with the

stakeholders.

As part of this process, and to support a discussion of the items on the list, it is useful to provide readings list to people involved in the FS activity. There is so much to read on any of the items on the list that they should be carefully curated to ensure that they are informative, timely and pertinent.

TM could provide support in this initial exploratory phase of work, by helping the researchers trawl through a vast literature to identify:

- **Lists of issues**
- **Lists of books and articles to justify and qualify those issues**
- **Lists of sources of further insight such as authors, newsletters and blogs etc.**

This material should be made available to the stakeholders, who should have a chance to comment or qualify or otherwise express their views on the content.

At this stage it could be useful to organize a workshop with key stakeholders, whose main purpose is to get them talking to each other and sharing their experiences.

Use Case 2: TM Support to Sense Making Activities

Having identified lots of issues and challenges, there is a need to make sense of these from a regional perspective, in order to determine their importance and urgency, and evaluate one's capacity to act to address an issue by “changing the game” by “breaking the trend” or by “mitigating its effects.”

An important step in making sense of the external issues is to localize them in order to understand their significance for the region and refine their definition.

Issues such as “climate change” or even “adverse weather events” are vague and can be hard to relate to the evolution of the region, especially if they are illustrated with reference to fires in the arctic circle, starving polar bears or piles of plastic in the mid-pacific garbage patch.

These concepts need to be “localized” so that local stakeholders can fully understand their significance for the region. The best way to do this is to draw inspiration from the global concepts already identified and search for their manifestation locally. Ideally this will provide insight into if, how and to what extent these global phenomena are manifest in the region, and how they are likely to evolve in future. Perhaps the main way in which climate change is manifest locally, is in terms of flooding or crop failure. The task now is to gather examples and data on this and take advantage of easy access to the people who have experienced these issues, so that they can talk about the impact it has had on their business and on their life.

TM can be of assistance in this task. The process will be more or less the same as for the

exploratory phase but is more geographically specific. The sources will most likely be in a local language. Other than this, the task is as before,

- **Identify local list of issues**
- **List local books and articles to justify and qualify those issues**
- **Lists of local sources of further insight such as individuals, journals and blogs etc.**

Another important aspect of the work of "sense making" is to get a feel for the dynamics of change. The factors that will drive or enable change with respect to the issues that have been previously identified.

The main tool in this case is called "Drivers' Analysis" the analysis of various trends and how they interact with each other and how they will drive or enable the evolution of the issues and challenges that have been identified.

This employs a range of tools which represent variations on PEST analysis. Like all such tools, they seem easy to use at first sight, but can be done well or badly depending on the skill of those managing the activity.

PEST, PESTLE, PESTLED, SLEPD, STEEP, STEEPLE, STEEPV are simple mnemonic variations intended to divide up the vast and complex world of possibilities in a way that is more or less exhaustive and guarantees broad coverage or most or all major issues that might need to be considered. PEST corresponds to a list of headings for factors that are 'P' for Political, 'E' for Economic, 'S' for Social or 'T' for Technological. PESTLE is an extension of this to include Legal and Environmental factors. In the same spirit STEEP refers to Socio-cultural, technological, environmental, economic and political factors. STEEPLE is an extension of this with ethical and legal factors. STEEPV is another extension that take account of new and emerging values. The 'D' in SLEPD refers to demographics. Whatever the preference they are merely headings under which one can organize lists of issues to be researched with a view to understanding what is happening and asking timely questions about the relevance of these from a local point of view.

The drivers' analysis is often done using a brainstorming approach. In my view this does not give good results, and that it is much better to first use desk research to:

- Provide an initial list of drivers and enablers.
- Provide a curated reading list of articles, papers or books to illustrate how they are interrelated.

There is potential for TM to provide assistance in the development of a list of drivers. More specifically it could help create:

- **Lists of drivers and enablers under the headings suggested by the various mnemonic tools such as PEST, PESTLE, PESTLED, SLEPD, STEEP, STEEPLE and STEEPV.**

- **Lists of drivers and enablers under headings such as trends, mega-trends, macro-trends, micro-trends, supertanker-trends.**

- **Books and articles to justify and qualify those issues.**

- **Sources of further insight such as individuals, journals and blogs etc.**

This material needs to be curated, edited and organized. It should be circulated in advance of a group-work meeting structured to

- Review, extend and discuss the drivers and enablers (factors).
- Explore the relationships between these and the challenges.

Having systematically develop these insights, other tools may be applied to go further in the sense making process. The various lists generated can become quite long and at some point, not too early in the process, it is wise to start to limit the focus by

- Ranking the various issues (opportunities and threats) in terms of impact and urgency.

- Ranking the drivers and enablers in terms of their local relevance and significance.

There are many ways for doing this. These include:

- One-dimensional ranking by “priority”

- Two-dimensional ranking using a “matrix” approach where independent criteria are represented along 2 axes. For example, one could refer to impact and another to urgency

- Another approach is to use more sophisticated multi-criterion decision support. Systems exist that mix qualitative and quantitative criteria, that allow the users to decide artefacts such as scales, weights, thresholds and “utility.” Such systems often support various kinds of analysis such as sensitivity or robustness of the ranked outcomes.

Usually a simple method is quite adequate, and the literature is full of group work methods that can be used interactively with stakeholders.

Finally, it is also possible to carry out a “SWOT” analysis, with a focus on the factors that provide a foundation for the future, namely the “strengths” and “weaknesses.” This may draw upon detailed knowledge of the region, its drivers and enablers such as its economic performance and demography.

These steps all build up to the development of a “vision” for the future of the region, in which the

main challenges have been addressed and main opportunities have been harnessed.

Many Foresight activities stop at this point. But without a roadmap or a plan for action, no recommendations will be provided, no policy response can be expected and nothing further is likely to happen.

Use Case 3: TM Support to Solution Seeking Activities

When there is a deeper understanding of where region intends to be at some time in the future, say in 10 or 15 years' time based on its policy response to the various opportunities and threats, there is now a need to develop a plan of action, based partly on creative thinking but mainly on a consideration of what has worked elsewhere.

This requires an exploration of how these issues have been addressed elsewhere, or of how they are currently being addressed elsewhere. There has been a tendency for local government to use their websites to communicate about policy options and programs in place.

There is a body of research literature available where researchers describe policies and their impact. There is also documentation available from the public websites of relevant agencies, often only in local language. In many cases all one can find is a reference to a program and perhaps a reference to a department, secretariat to which one can apply for further information.

It is hard work to trawl these and identify documents or fragments that may be relevant. So, this is an area where TM could in principle prove very helpful.

This may require an iterative approach in order to acquire the vocabulary needed to inform the TM system. There may be a need to experiment with the best tactics and methodology to apply, but after "human intervention" the desired outcome of a TM campaign might consist of the following elements.

For each of the main challenges the TM system could support the work of the researcher by helping to establish initial lists of

- **Policies, programs or initiatives that have been designed to address those challenges**
- **Agencies, institutions or people who have been involved in such activities**
- **Studies or articles whose express or effective purpose is to evaluate the impact or effectiveness of those policies, programs or initiatives**
- **Agencies, institutions or people who have been involved in such activities**

The development of such lists may require cooperation with key sources of information such as the RSA (Regional Studies Association) which publishes scholarly work on regional policy related issues, via a number of independent and highly ranked journals.

This still leaves a lot of work to be done, but the use of TM could help to reduce the cognitive load on the researcher of the initial search task.

Use Case 4: TM Support to SDM

Another important task in the project is the use of System Dynamic Modelling. The initial working hypothesis is that this can play an important role with respect to drivers' analysis.

It could be of use for both researchers that support the work of the FS process, and for the facilitators who might use it to help animate group work sessions with stakeholders.

The use of TM in conjunction with SDM could help the SDM team by facilitating the search for

- **Dynamical rules that can be used to drive the simulations**
- **Data sets that can be used as inputs**
- **Literature that qualifies and illustrates the use of the models and accompanying data sets**

Despite growing interest in its use, SDM is not yet a mainstream tool, so there is a need to develop interest groups or communities of practice and find a home for relevant data, rulesets and related literature.

It will take some time to cultivate these possibilities, but several come to mind.

- The CAP payment agencies as a source of detailed up to date and historical geo-spatial data on farms, farming activities and related household incomes
- COPERNICUS as a source of geo-spatial data sets on land-use in both urban and rural spaces
- EUROSTAT as a partner to help develop related norms and standards
- The open data initiative as a forum for making relevant data sets available to the policy community

TM might also find a use navigating these very large and very rich data bases for the purpose of supporting policy process that employ SDM in decision support and stakeholder engagement.

11 Annex II - Example of data received from Pilot Flanders (VITO)

These data were sent by VITO following the first workshop for the Pilot 'Flanders' concerning rural attractiveness and the local needs. This provides a clear example of the interactions between the workshop-based exercises in the Pilots and text mining.

Location

List of all provinces and communities:

https://www.vlaanderen.be/gemeenten-en-provincies/overzicht

Names of Regional Landscapes (not all in English) and their websites:

https://www.regionalelandschappen.be/ (in Dutch)

1. Flemish coast (Vlaamse Kust):

   www.belgiancoast.co.uk / www.dekust.be

2. Bruges woodland and wetland (Brugse Ommeland)

   www.brugseommeland.be/en

3. Region of the Lys (Leiestreek)

   https://langsdeleie.be/en/the-lys-region-nature-with-artistic-genes/Meetjesland

   https://www.toerismemeetjesland.be/en

4. Rupel Region (Rupelstreek)

   www.toerismerupelstreek.be (in Dutch)

5. Dijle and green corridor (Dijleland en Groene Gordel)

   www.rld.be (in Dutch)

6. Antwerp Campine Region (Antwerpse kempen)

   www.antwerpsekempen.be (in Dutch)

7. Limburg Campine Region (Limburgse kempen)

   www.toerismelimburg.be/en

8. Region of the Meuze (Maasland)

   www.toerismelimburg.be/en

9. The Voer Region (Voerstreek)

   https://www.voerstreek.be/?lang=en

10. Haspengouw

    www.toerismelimburg.be/en

Key search words

SEARCH WORDS: Flanders + Rural, landscape, regional landscape, farming, agriculture, attractiveness, land use change, mobility, rural planning

RELEVANT WORDS ACCORDING TO STAKEHOLDERS: landscape development, landscape quality, new farming business model, climate change, climate adaptation, climate mitigation, climate resilience, soil, land and water management, rural policy, rural policy measures, landscape dynamics, Leader, biodiversity, ecosystem services, landscape maintenance, landscape architecture, historical landscapes and heritage,

CRUCIAL NEEDS: Quantification of rural attractiveness and different pressures on landscape, climate resilient landscapes, sustainable land and water management, landscape planning, rural development planning, scenarios matching qualitative and quantitative indicators.

Websites in English

https://www.vlm.be/en

https://lv.vlaanderen.be/nl/voorlichting-info/publicaties/studies/report-summaries

https://enrd.ec.europa.eu/sites/enrd/files/be-fl_qnt_summary_v1_0.pdf

https://enrd.ec.europa.eu/country/belgium_en

https://ec.europa.eu/agriculture/sites/agriculture/files/rural-development-2014-2020/country-files/be/factsheet-flanders_en.pdf

https://enrd.ec.europa.eu/sites/enrd/files/nrn_profile_be-f.pdf

https://ruraalnetwerk.be/welcome

http://buur.be/news-item/report-adaptive-public-transport-flemish-rural-areas-delivered/

https://www.vlaanderen.be/publicaties/rural-development-programme-flanders-2014-2020-rdp-iii

https://ruraalnetwerk.be/sites/default/files/publicatie_files/Brochure%20Best%20Practices_Vlaams%20Ruraal%20Netwerk_goed%20voor%20druk.pdf

http://www.flanderstoday.eu/living/village-hubs-could-set-rural-life-motion

https://pure.ilvo.be/portal/nl/publications/perception-of-rural-landscapes-in-flanders-looking-beyond-aesthetics(c9c5da1c-7164-418f-a5f3-9efc2df51c48).html

https://lv.vlaanderen.be/sites/default/files/attachments/prospects_and_challenges_for_agricultural_diversification_in_a_peri-urban_region_flanders_-_belgium.pdf

https://books.google.be/books?id=7FZZfapeWJQC&pg=PA55&lpg=PA55&dq=%22Flanders%22+rural+landscape&source=bl&ots=D-J75oOxvL&sig=ACfU3U3tzdw5FO7WxjpP6dVmg3eOIOwq2g&hl=en&sa=X&ved=2ahUKEwiJhIuNp5vjAhUYQkEAHeOOD444ChDoATAHegQICRAB#v=onepage&q=%22Flanders%22%20rural%20landscape&f=false

https://books.google.be/books?id=RruCbckkX4YC&pg=PA299&lpg=PA299&dq=%22Flanders%22+rural+landscape&source=bl&ots=zAxGxzc6_A&sig=ACfU3U0jV5zWQ4-jD8mHPDSz8AT4ccpG0Q&hl=en&sa=X&ved=2ahUKEwiJhIuNp5vjAhUYQkEAHeOOD444ChDoATAIegQICBAB#v=onepage&q=%22Flanders%22%20rural%20landscape&f=false

https://www.vrt.be/vrtnws/en/2019/03/06/take-a-walk-through-the-landscape-that-inspired-breugel/

https://www.interregeurope.eu/innocastle/news/news-article/5130/first-stakeholder-meeting-in-flanders/

https://pure.ilvo.be/portal/nl/publications/changing-land-use-in-the-countryside-stakeholders-perception-of-the-ongoing-rural-planning-processes-in-flanders(d50fec1e-902f-45d1-a3cc-250ddd6d9467).html

https://rega.kuleuven.be/if/farming-in-flanders

https://www.flandersinvestmentandtrade.com/invest/en/sectors/agribusiness

http://www.flanderstoday.eu/living/searching-simple-life

https://en.wikipedia.org/wiki/Agriculture_in_Flanders

https://twitter.com/boerenbond?lang=en

https://www.oecd.org/regional/regional-policy/land-use-Belgium.pdf

https://northsearegion.eu/share-north/news/new-concept-in-flanders-mobihubs/

https://www.autodelen.net/blog/shared-mobility-the-rosetta-stone-for-rural-areas

12 References

Acheson, Elise, Volpi, Michele, Purves Ross S. (2018). Machine learning for cross-gazetteer matching of natural features", Department of Geography, University of Zurich, Switzerland, https://www.tandfonline.com/doi/full/10.1080/13658816.2019.1599123

Charvat, K., Kafka, S., Splicha, M., (2007). Uniform Resource Management as Tool for Content Awareness of Information and Knowledge Inside Communities, Naturnet - Redime NEWSLETTER No. 6 December 2007

Cuhls, K. (2003). From Forecasting to Foresight Processes—New Participative Foresight Activities in Germany, Journal of Forecasting J. Forecast. 22, 93–111 (2003) Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/for.848

Daugstad, Karoline & rønningen, Katrina & Skar, Birgitte. (2006). Agriculture as an upholder of cultural heritage? Conceptualizations and value judgements - A Norwegian perspective in international context. Journal of Rural Studies. 22. 67-81. 10.1016/j.jrurstud.2005.06.002.

European Commission (2019). The future is rural: the social objectives of the next CAP. Last consulted on 4 September 2019
https://ec.europa.eu/info/news/future-rural-social-objectives-next-cap-2019-feb-15_en

European Union (2014). The rapid growth of EU organic farming. Consulted 2019, September 2nd https://ec.europa.eu/info/sites/info/files/food-farming-fisheries/farming/documents/market-brief-rapid-growth-eu-organic-farming_en.pdf

Fielding, R., T., (2000). Architectural Styles and the Design of Network-based Software Architectures, Doctoral dissertation. Last consulted on 14 September 2019 at: https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Giraldi, J., (2017). Text Mining for assessing and monitoring environmental risks. Concept paper, http://publications.europa.eu/portal2012-portlet/html/downloadHandler.jsp?format=pdf&identifier=9eb7c32b-371a-11e7-a08e-01aa75ed71a1&language=en&part=&productionSystem=cellar

Hradec, J., Ostlaender, N., Macmillan, C., Acs, S., Listorti, G., Tomas, R., Arnes Novau, X., 2019. Semantic Text Analysis Tool: SeTA, EUR 29708 EN, Publications Office of the European Union, Luxembourg, 2019, ISBN 978-92-76-01518-5, doi:10.2760/577814, JRC116152. Last consulted in September 2019 at: http://publications.jrc.ec.europa.eu/repository/bitstream/JRC116152/kjna29708enn_1.pdf

Kayser, V., (2016). Extending the Knowledge Base of Foresight: the Contribution of Text Mining, https://d-nb.info/1156272084/34

Lausch, A., Schmidt, A., and Tischendorf, L., (2015). "Data mining and linked open data – New perspectives for data analysis in environmental research." Ecological Modelling 295 (2015): 5-17 https://www.researchgate.net/publication/266260663_Data_mining_and_linked_open_data_-

New perspectives for data analysis in environmental research

Manning, C. D., Ramage, D., Hall, D. and Nallapati, R. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 248–256, Singapore, 6-7 August 2009. c 2009 ACL and AFNLP

Niemeyer, G., (2008). Available at: https://blog.labix.org/2008/02/26/geohashorg-is-public

Paulheim, H., (2013). Exploiting Linked Open Data as Background Knowledge in Data Mining, DMoLD 2013, 1082
http://ceur-ws.org/Vol-1082/extendedAbstract.pdf

Vitiugin, F. and Castillo, C. (2019). Comparison of Social Media in English and Russian During Emergencies and Mass Convergence Events, http://chato.cl/papers/vityugin_castillo_2019_russian_english_language_differences_twitter_emergency.pdf

List of websites cited:

https://www.gdeltproject.org/, last consulted on 14 September 2019

https://sdi4apps.eu/, last consulted on 14 September 2019

https://www.digitalocean.com/, last consulted on 14 September 2019

13 Annex III – Responses to the monitors' comments

Comments made by EU monitors	Explanation
The description of the policy development process on p. 11 is very over-simplified…	Added a phrase on P11 to explain that the procedure has been overly streamlined to allow text mining developers to have a clear idea of the policy process
This also demonstrates the importance of having meaningful input from regional stakeholders. (P11)	Added a short explanation about the importance of regional Pilots in the composition of semex.io library
p. 16/17 is replete with jargon and process. The narrative a few pages earlier tries to defend the methodology and one wonders if the software designers understand that the technology is not necessarily the best way to go about this? Which policymaker is really going to read a curated reading list? (p.16)	Added an explanation on P17 in the introduction to the chapter on Foresight clarifying that for TM developers it is important to have clear how the FS procedure works.
There is an error in this sentence: To mention the most important, generational renewal…	Phrase changed on page 8