D4.1 Regional Library for Needs Analysis

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:	1.2	
Date:	30 May 2021	
Responsible Partner:	NUVIT	
Contributing Partners:	KAJO	
Reviewers:	Petra Korkiakoski (HAMK) Pavel Kogut (21C Consultancy)	
Dissemination Level:	Public	X
	Confidential - only consortium members and European Commission Services	
Keywords:	Regional library, crawler, data sources, acquisition.	

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
0.1	07/10/2019	Karel Pánek	NUVIT	initial version
0.2	09/10/2019	Karel Pánek	NUVIT	improvements, upload for collaborative editing
0.3	18/10/2019	Karel Pánek	NUVIT	non-technical information, structure improvements, example of source specification
0.4	24/10/2019	Karel Pánek	NUVIT	details elaboration
0.5	19/11/2019	Petra Korkiakoski	HAMK	review
0.6	20/11/2019	Karel Pánek	NUVIT	update
1.0	25/11/2019	Karel Pánek	NUVIT	input extent table
1.1	31/03/2021	Karel Pánek	NUVIT	Updates to the deliverable according to the monitors' comments
1.2	26/05/2021	Karel Pánek	NUVIT	Updates to the deliverable according to the internal reviewers' comments.
1.2	30/05/2021	Miloš Ulman	NUVIT	Final check of formatting

Table of Contents

List of Tables

List of Figures

Executive Summary

PoliRural approach strongly benefits from diversity of input data in terms of their focus, source and evaluation. In order to deal with such a diversity, PoliRural introduces its own methods of automatic data collection from dynamic sources (such as social networks) and static sources (such as web pages).

The purpose of regional library is to automatically collect public data from static sources across the Internet, in order to make them (repeatedly) available for internal processing within Text Mining Technology (WP2) and for future reference within Innovation Hub & System Dynamics Technology (WP3).

In addition to oversight of this deliverable, NUVIT provides actual crawler technology.

This deliverable D4.1 focuses on process of Regional Library creation and consists of generic crawler technology implementation and integration, initial regional sources specification, and readiness for execution of continual acquisition of initial (and future) sources.

1. Position and role of Regional Library for Needs Analysis in PoliRural

1. Purpose

Public internet sources are relevant for the purpose of rural needs gathering and policy assessment. In order to acquire up-to-date information and assure reliable and efficient access to it within Text Mining Technology (WP2) and Innovation Hub (WP3), the process of automatic and periodic data acquisition is introduced. As part of central repository, the data obtained in this way are then made available in real-time, both manually and programmatically, across the project.

Regional libraries are collected as an input for text mining tools to analyse specific situation(s) in pilot regions and their needs.

Part of output is in the form, which will support integration with DiHs tools, and it will be possible to use the results of the analysis in a map context. Also, results of sentiment analysis can be visualised and it can help regions to understand problems and challenges better.

2. Role in Text Mining Technology

In the Text Mining Technology, the content of Regional Library will be used automatically, as input for various text-mining pipelines. As those processing pipelines (and their context) will evolve over time, it is important to support the reprocessing of input in consistent and effective way.

3. Role in Innovation Hub

In Innovation Hub, the content of Regional Library will be used manually 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 external static sources), and
- vindication and improvement (for particular processing verification).

The Innovation Hub is based on selected SDGs (Sustainable Development Goals) and will be able to create maps and visualisations related to topics such as sustainability, poverty, education, decent work, etc.

The potential of the future using of the Innovation Hub by regional stakeholders is ensured due to collaboration with universities.

2. Data sources for Regional Library

2.1. Overview

For the purposes of PoliRural, the Regional Library will mirror or register content from all static sources, that are available and relevant to rural needs gathering and policy assessment. Static sources consist of web published documents and include common online news articles, academic papers, blogs, discussion forums, etc.

Due to their nature, dynamic sources (such as social networks) are not stored within Regional Library and they are processed directly within Text-Mining Technology (WP2).

Data from both static and dynamic sources will become part of central repository as input for text processing pipelines.

3. Initial Specification

The form of initial specification of relevant sources was defined and specific lists of internet links, domains, search phrases, topics and known needs were requested from partners who will be conducting regional pilots. Initial specifications from partners were collected at the following extent presented in Table 1.

Country	Area	Landscapes	Phrases	Topics	Needs	Links	Websites	Social	Total
BE	Flanders	10	9	12	5	27	10	0	73
CZ	Central Bohemia	2	26	19	5	36	8	0	96
ES	Galicia	1					1		2
FI	Häme	8	13	17	6	6	24	12	86
GR	Central Greece	5	31	35	32	25	38		166
IE	County Monaghan	9	57	71	33	99		26	295
IL	Easter Galilee	10	19	12	7	4	2		54
IT	Foggia	13	8	7		6	13		47
LV	Vidzeme	14	30	25	5	61	79	49	263
MK	Gevgelija – Strumica	2	38	20	39	17	19		135
PL	Mazowieckie	15	16	15	8	8	22	3	87
SK	Nitra	8	13	9	5	22	74	14	145
									1449

Table 1 Number of sources per pilot regions

Initial collection of inputs is sufficient for practical assessment of acquisition efficiency, actual content usability and other aspects, that will determine further development of the acquisition and text-mining technology, as well as development of the collection itself.

Inputs collected are sufficient for initial needs library scope to exceed expected minimum of 900+ records, presumably allowing to achieve significantly more than estimated 100+ opinions and 10+ emotions per each of 12 regions.

An example of input is available in Annex 1.

4. Basic architecture of components

4.1. Generic Crawler

The Crawler is a software tool, that systematically visits selected web pages in order to download their content for a specific purpose. In PoliRural, such purpose is to collect publicly available textual information that is relevant or supporting the needs gathering and policy assessment. Such collection maintained by the Crawler then enables the application of text-mining processes as defined in WP2.

Technically, the Crawler provides a generic method for continual acquisition of any unstructured data from the internet. It is primarily intended for bulk acquisition of standardly available webpages and other web published documents.

Crawler is integrated within Semantic Explorer as inherent functionality of text-mining technology. As opposed to static inputs collected at initial stage of the project (see 2.2 above), at later stage, expert users will be allowed to add new links for crawling through Innovation Hub.

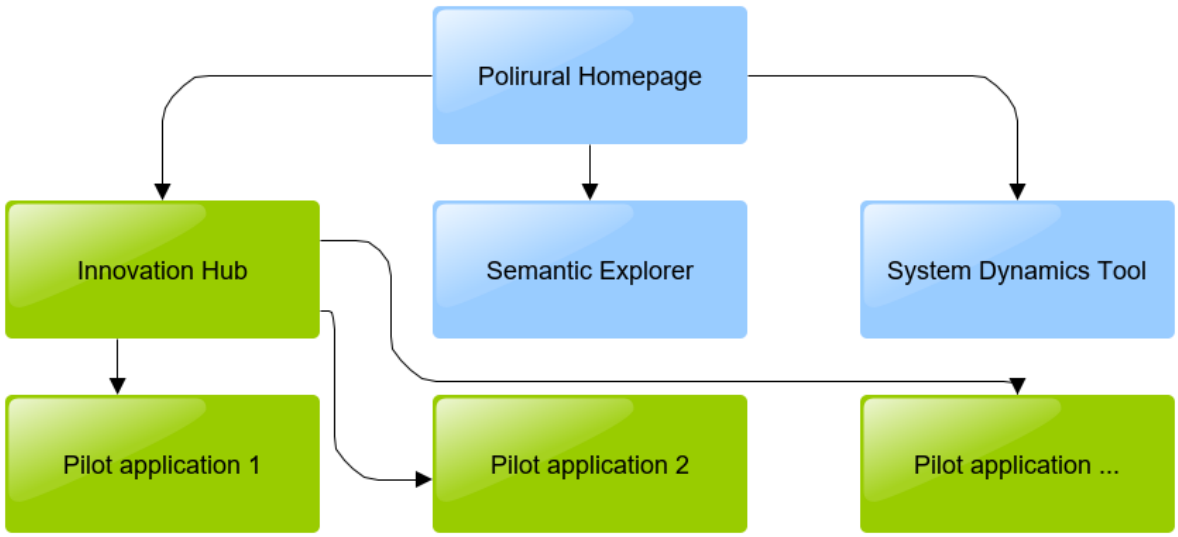

Figure 1 Architecture of components

4.1.1. Design

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-server instance of the Crawler.

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 (typically a content repository) continually.

Crawler core implements real-time scheduling, modularity, scalability and application programming interface (API).

4.1.2. Considerateness

Similar to regular internet user, in order to get information, Crawler visits 3rd party systems without explicit approval. These systems however vary in capabilities, parameters and endurance. Due to its automatic nature (as opposed to regular internet user), the Crawler is capable of significant utilization of these systems' resources. Therefore, a high standard of considerateness and effectiveness is maintained by Crawler.

4.1.3. On the fly processing

Not only acquisition itself, but also certain processing of the content needs to be performed on-the-fly due to the high-scalability (potential high-throughput) and real-time nature of Crawler operation.

4.1.4. Breadth-First Search (BFS)

Crawler is capable of discovering new links within documents on its own. Initially, single links (as specified by partners) will be visited, but automatic recursive operation is available for later stages of the project, if required. Link discovery involves certain level of content format parsing (3.1.4.2).

4.1.5. Real-time Evaluation

Crawler evaluates certain metadata at internal records, protocol (3.1.4.1) and format (3.1.4.2) levels. Such evaluation needs to be done in real time as Crawler needs to decide on tasks prioritization on an ongoing basis. This allows the Crawler for example to adapt to the current load of 3rd party web server. Some web serves also provide metadata that indicate whether the document was changed or not, allowing the Crawler to avoid unnecessary re-downloading of document's content. Real-time metadata evaluation also supports optimization of long-term storage requirements.

4.1.6. Content Normalisation

3rd party web servers vary in technical parameters of documents, such as file formats or special character encodings. For historical reasons, the web pages written in Czech language for example, may still appear in nine different variants of special character encoding. Also, principally textual documents are often published in graphic oriented formats such as PDF. To cope with these variations the Crawler itself offers automatic document conversion into plain-

text, including character set and encoding unification into UNICODE / UTF-8 standard. Such unification significantly simplifies further processing by downstream tools.

4.1.7. Modules

In order to support project requirements in a long-term outlook, Crawler is implemented modularly. Currently, the Crawler implements certain communication protocols and file formats and can be extended to support other protocols and / or formats in future.

Protocols and formats that are already available in initial stage of PoliRural project are listed in this section.

4.1.8. Protocols

At the level of protocol modules, Crawler implements the most important standards of communication with world-wide web servers:

- Domain Name System - DNS (RFC 1035)
- Hypertext Transfer Protocol - HTTP/1.1 (RFC 2616)
- HTTP Over TLS (RFC 2818)
- File Transfer Protocol - FTP (RFC 2228)

These protocols are necessary and typically most commonly used to make data publicly available for anonymous access via the Internet.

4.1.9. Formats

At the level of format modules, Crawler also directly implements or directly integrates parsers for the most common file formats found on world-wide web and can product plain-text transformations continuously during crawling:

- directly implemented parsers:
    - SGML (HTML, HTML5, SVG, XML)
    - PDF (Adobe)
    - CSV (with control characters set detection)
- external parsers:
    - DOC (Microsoft)
    - DOCX (Microsoft)
    - RTF (Microsoft)
    - ODT (The Document Foundation)

These formats are necessary and typically most commonly used to make data publicly available for anonymous access via the Internet.

4.1.10. API

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

In standard scenario, any content acquired by Crawler is forwarded to a storage or processing pipeline. If storage is available, Crawler utilizes it also for metadata, such as normalized URI of origin, content, title, time of acquisition, checksums, etc.

For standard operation, Crawler requires storage API to support file management operations ("Insert" / "Remove" / "Browse") and file attribute operations ("Get" / "Put" / "Rid").

4.2. Repository

Repository is implemented within text-mining technology infrastructure and available to Crawler through trivial (get attribute / set attribute) API.

4.3. Discussion

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.

5. Process Readiness

Acquisition (as well as processing) of Regional Library content is by nature an iterative task with evolving success rate.

5.1. Present state

Crawler was successfully tested on random internet sites, i. e.:

- 10k documents of financnisprava.cz
- 400k documents of cs.wikipedia.org

At the time of writing this document, the process of automatic acquisition of static sources into Regional Library is available and further testing and development is carried out.

5.2. Future state

As soon as storage service becomes available (when WP2 technology reaches its first integration), Regional Library construction will be started and maintained on a regular basis.

Complete control over acquisition technology provides agility, required for future improvement, extension and fine-tuning. Initial iterations will be therefore followed by an assessment and planning of improvements at technology level. Presumably, source specifications will be updated as well, based on practical outputs or needs of text-mining processes.

6. Annex 1 - Example of Regional Inputs for Needs Library

Name of Pilot and organization:

Central Bohemia, Czech Republic; NUVIT

Names of regional landscapes (For example: Flemish coast (Vlaamse Kust), Bruges woodland and wetland (Brugse Ommeland), Region of the Lys (Leiestreek), ecc…) (locations):

Střední Čechy (Central Bohemia), Středočeský kraj (Central Bohemian Region)

Key search words (for example: Flanders + Rural, landscape, farming, attractiveness, land use change, mobility, rural planning etc…) (keywords):

Central Bohemia/Central Bohemian Region + agriculture, rural development, regional development, regional disparities, organic farming, farm, land use, attractiveness, local food, food production, newcomers, migration, soil degradation, soil erosion, countryside, rural areas, inner periphery, settlement structure, cultural landscape, housing, pollution, sustainable development, sustainable agriculture, tourism

Relevant words according to stakeholders (For example: landscape development, landscape quality, new farming business model, rural policy, etc…) (topics):

labour market in the agricultural sector, population ageing, demographic change, population change, organic farming, local food production, landscape consumption, drought, climate change, short food supply chain, circular economy, food security, renewable resources, education, low-carbon economy, bio economy, critical raw materials, housing costs, agritourism

Crucial needs (for example: Quantification of rural attractiveness and different pressures on landscape, landscape planning, rural development planning, scenarios matching qualitative and quantitative indicators, etc….) (requirements):

changing of demographic scenario, population change and its driving factors in rural areas, boosting of inner peripheral attractiveness, environmentally sustainable farming

Links to policy documents from all levels (EU, national, regional, local):

Aktualizace programu rozvoje územního obvodu Středočeského kraje na období 2018-2024 s výhledem do 2030 (https://s-ic.cz/wp-content/uploads/2018/01/Program-rozvoje-kraje_podklady-pro-aktualizaci-strategie-v2.pdf)

Politika územního rozvoje ČR, aktualizace č. 1 (2015)

(https://www.databaze-strategie.cz/cz/mmr/strategie/politika-uzemniho-rozvoje-cr-ve-zneni-aktualizace-c-1-2015)

Strategie regionálního rozvoje ČR 2014-2020

(https://www.databaze-strategie.cz/cz/mmr/strategie/strategie-regionalniho-rozvoje-cr-2014-2020)

Akční plán ČR pro rozvoj ekologického zemědělství 2016-2020 (https://www.databaze-strategie.cz/cz/mze/strategie/akcni-plan-cr-pro-rozvoj-ekologickeho-zemedelstvi-v-letech-2016-2020)

Akční plán pro biomasu v ČR 2012-2020 (https://www.databaze-strategie.cz/cz/mze/strategie/akcni-plan-pro-biomasu-v-cr-2012-2020)

Koncepce na ochranu před následky sucha pro území České republiky (2017) (https://www.databaze-strategie.cz/cz/mze/strategie/koncepce-na-ochranu-pred-nasledky-sucha-pro-uzemi-ceske-republiky)

Národní akční plán ke snížení používání pesticidů v ČR (2012) (https://www.databaze-strategie.cz/cz/mze/strategie/narodni-akcni-plan-ke-snizeni-pouzivani-pesticidu-v-ceske-republice)

Strategie bezpečnosti potravin a výživy 2014-2020 (https://www.databaze-strategie.cz/cz/mze/strategie/strategie-bezpecnosti-potravin-a-vyzivy-2014-2020)

Státní politika životního prostředí ČR 2012-2020 [akt. 2016] (https://www.databaze-strategie.cz/cz/mzp/strategie/statni-politika-zivotniho-prostredi-cr-2012-2020-akt-2016)

Strategický rámec Česká republika 2030 (2017) (https://www.databaze-strategie.cz/cz/mzp/strategie/strategicky-ramec-ceska-republika-2030)

Střednědobá strategie (do roku 2020) zlepšení kvality ovzduší v ČR (2015) (https://www.databaze-strategie.cz/cz/mzp/strategie/strednedoba-strategie-do-roku-2020-zlepseni-kvality-ovzdusi-v-ceske-republice)

Strategie přizpůsobení se změně klimatu v podmínkách ČR (2015) (https://www.databaze-strategie.cz/cz/mzp/strategie/strategie-prizpusobeni-se-zmene-klimatu-v-podminkach-ceske-republiky)

Strategie ochrany biologické rozmanitosti ČR 2016-2025 (https://www.databaze-strategie.cz/cz/mzp/strategie/strategie-ochrany-biologicke-rozmanitosti)

Státní program environmentálního vzdělávání, výchovy a osvěty a environmentálního poradenství 2016-2025 (https://www.databaze-strategie.cz/cz/mzp/strategie/statni-program-environmentalniho-vzdelavani-vychovy-a-osvety-ceske-republiky-a-akcni-plan)

Program předcházení vzniku odpadů ČR (2014) (https://www.databaze-strategie.cz/cz/mzp/strategie/program-predchazeni-vzniku-odpadu-cr)

Politika ochrany klimatu v ČR (2017) (https://www.databaze-strategie.cz/cz/mzp/strategie/politika-ochrany-klimatu-v-cr)

Národní program snižování emisí ČR (2015) (https://www.databaze-strategie.cz/cz/mzp/strategie/narodni-program-snizovani-emisi-ceske-republiky-2)

Koncepce podpory místní Agendy 21 v ČR do roku 2020 (2012) (https://www.databaze-strategie.cz/cz/mzp/strategie/koncepce-podpory-mistni-agendy-21-do-2020)

Digitální Česko: Koncepce Digitální ekonomika a společnost (2018) (https://www.databaze-strategie.cz/cz/mpo/strategie/digitalni-ekonomika-a-spolecnost)

Koncepce podpory malých a středních podnikatelů 2014-2020 (https://www.databaze-strategie.cz/cz/mpo/strategie/koncepce-podpory-malych-a-strednich-podnikatelu-2014-2020)

Národní akční plán ČR pro energii z obnovitelných zdrojů 2010-2020 [akt. 2015] (https://www.databaze-strategie.cz/cz/mpo/strategie/narodni-akcni-plan-cr-pro-energii-z-obnovitelnych-zdroju-2010-2020-ii)

Státní energetická koncepce České republiky (2015) (https://www.databaze-strategie.cz/cz/mpo/strategie/statni-energeticka-koncepce-ceske-republiky-2015)

Státní politika v elektronických komunikacích - Digitální Česko v. 2.0 (https://www.databaze-strategie.cz/cz/mpo/strategie/statni-politika-v-elektronickych-komunikacich-digitalni-cesko-v-2-0-cesta-k-digitalni-ekonomice-statni-politika-v-elektronickych-komunikacich-digitalni-cesko-v-2-0-cesta-k-digitalni-ekonomice)

Strategie digitálního vzdělávání ČR do roku 2020 (https://www.databaze-strategie.cz/cz/msmt/strategie/strategie-digitalniho-vzdelavani)

Dlouhodobý záměr vzdělávání a rozvoje vzdělávací soustavy ČR 2015-2020 (https://www.databaze-strategie.cz/cz/msmt/strategie/dlouhodoby-zamer-vzdelavani-a-rozvoje-vzdelavaci-soustavy-ceske-republiky-na-obdobi-2015-2020?typ=struktura)

Strategie politiky zaměstnanosti do roku 2020 (2015) (https://www.databaze-strategie.cz/cz/mpsv/strategie/strategie-politiky-zamestnanosti-do-roku-2020)

Evropa 2020 / Europe 2020 (2010) (https://www.databaze-strategie.cz/cz/eu/strategie/evropa-2020)

Inovace pro udržitelný růst: Biohospodářství pro Evropu 2012-2020 (https://www.databaze-strategie.cz/cz/eu/strategie/inovace-pro-udrzitelny-rust-biohospodarstvi-pro-evropu-2012-2020)

Rámec politiky EU v oblasti klimatu a energetiky 2020-2030 (2014) (https://www.databaze-strategie.cz/cz/eu/strategie/ramec-politiky-v-oblasti-klimatu-a-energetiky-v-obdobi-2020-2030-2014)

Směrnice EU o podpoře využívání energie z obnovitelných zdrojů (2009) (https://www.databaze-strategie.cz/cz/eu/strategie/smernice-eu-o-podpore-vyuzivani-energie-z-obnovitelnych-zdroju-2009)

Společná zemědělská politika EU [akt. 2013] (https://www.databaze-strategie.cz/cz/eu/strategie/spolecna-zemedelska-politika-eu-2013)

Strategie EU pro přizpůsobení se změně klimatu (2013) (https://www.databaze-strategie.cz/cz/eu/strategie/strategie-eu-pro-prizpusobeni-se-zmene-klimatu-2013)

Uzavření cyklu – akční plán EU pro oběhové hospodářství (2015) (https://www.databaze-strategie.cz/cz/eu/strategie/uzavreni-cyklu-akcni-plan-eu-pro-obehove-hospodarstvi)

SDĚLENÍ KOMISE EVROPSKÉMU PARLAMENTU, RADĚ, EVROPSKÉMU HOSPODÁŘSKÉMU A SOCIÁLNÍMU VÝBORU A VÝBORU REGIONŮ o přezkumu seznamu kritických surovin pro EU a o provádění iniciativy v oblasti surovin (https://eur-lex.europa.eu/legal-content/CS/TXT/?uri=CELEX%3A52014DC0297)

Links to Pilots' related websites:

Main websites:

https://www.czso.cz/csu/czso/11-zemedelstvi-8ksclawere

https://www.kr-stredocesky.cz

Other websites:

http://eagri.cz/public/web/mze/

www.msmt.cz

www.mzp.cz

www.mmr.cz

http://stredocesky.nsmascr.cz

http://www.kis-stredocesky.cz/

List of social media used and accounts/groups, if available:

N/A

7. Annex 2 – Responses to the monitors' comments

Comment made by the monitors	Explanation
Initial specification does not look very systematic but it was sufficient for the first phase of technical developments and testing in WP2 and WP3. This report is technical in parts and explains how the regional library works, including the use of Crawlers. A major weakness of this deliverable is that less attention has been directed on social capital, community, participation and bottom up approaches to rural development, there is a lot of focus on landscape and agriculture. The role of the Innovation Hub on P. 5 is not clear A number of critical questions are outstanding in relation to the regional library: Who is this for? How will it be used? Will it have life beyond the moment that it is no longer updated by the team?	Regional libraries are collected as an input for text mining tools to analyse specific situation(s) in pilot regions and their needs. Part of output is in the form, which will support integration with DiHs tools, and it will be possible to use the results of the analysis in a map context. Also, results of sentiment analysis can be visualised and it can help regions to understand problems and challenges better. Page no. 5. The Innovation Hub is based on selected SDGs (Sustainable Development Goals) and will be able to create maps and visualisations related to topics such as sustainability, poverty, education, decent work, etc. Page no. 5. The potential of the future use of the Innovation Hub by regional stakeholders is ensured due to collaboration with universities. Page no. 5.