D4.3 Regional Library for Policy Evaluation

Project	PoliRural	
Project title:	Future Oriented Collaborative Policy Development for Rural Areas and People	
Grant	818496	
Website:	www.polirural.eu	
Contact:	info@polirural.eu	
Version:	1.6	
Date:	31 May 2021	
Responsible:	NUVIT	
Contributing:	KAJO, CZU	
Reviewers:	Petra Korkiakoski, HAMK Agnese Krievina, AREI	
Dissemination Level:	Public	X
	Confidential - only consortium members and European Commission Services	
Keywords:	Regional Library, text mining, Semex.io, pilots	

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
1.0	24/10/2019	Karel Pánek	NUVIT	initial version
1.1	13/05/2020	Karel Pánek	NUVIT	review
1.2	22/05/2020	Karel Pánek	NUVIT	updated figures
1.3	25/05/2020	Petra Korkiakoski	HAMK	Internal review and comments
1.4	25/05/2020	Karel Pánek	NUVIT	review
1.5	25/05/2020	Karel Pánek	NUVIT	formatting corrections
1.6	31/05/2020	Miloš Ulman	CZU	Conclusion added
1.7	17/03/2021	Karel Pánek	NUVIT	Updated according to the monitors' comments
1.8	31/03/2021	Karel Pánek	NUVIT	Updated according to the internal reviewers' comments
1.8	30/05/2021	Miloš Ulman	CULS	Final check of formatting

Table of Contents

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 a regional library is to automatically collect public data from static sources across the Internet. Such data is collected into a database within Text Mining Technology (WP2) for further processing. Together with other inputs and text-mining outputs the data is available to users as a part of Innovation Hub (WP3), through a dedicated online interface at https://semex.io/.

Following the deliverable D4.1 Regional Library for Needs Analysis, this deliverable D4.3 Regional Library for Policy Evaluation focuses on Regional Library expansion in terms of topics and proportionality. In addition to oversight of this deliverable, crawler, classification and parsing technologies are provided to support future expansion of Regional Library in semi-automatic or automatic way.

1 Current Expansion

Following deliverable D4.1, this deliverable D4.3 focuses on Regional Library expansion in terms of topics and proportionality. Source specifications were updated, based on practical experience with initial versions of text-mining processes.

Based on the collection of the first set of resources, disproportions were identified in terms of the content and scope of coverage of individual areas of interest across the territories of the pilot solutions. It has been found that for some territories, the sources of authentic external information are severely limited (for homogeneous resources, see below). In addition, in terms of technology development and validation, English resources were naturally preferred in the initial stage. Therefore, the partners implementing the pilot projects were recommended to expand the initial resource survey in the categories of policy / strategy, need and need-policy gaps and also emerging topics.

The sources identified in this way were subsequently added to the Regional Library. In the next stages, iterative expansion is expected, depending on the experience and needs of field experts with the outputs from the processing of this data. Engagement of field experts ensures relevance and validity of added sources.

For the initial stage of the project (focused on the actual creation and start of technology verification), it was decided to use only individual documents that are identified - relevant to field experts. In the next stage, scaling (in terms of technology and content) is assumed, which will allow field experts to incorporate other sources (e.g., homogeneous - systematic datasets or collateral textual information). For these cases, crawler technology is suitable, which does not focus only on individual links, but on in-depth acquisition. The actual implementation of the PDF parser is a secondary tool, providing better control over aspects of the conversion reliability of this format into plain text, which is necessary for its further processing.

1.1 Topics

In order to thematically expand the database, individual pilots identified additional resources focusing on policies (and strategies), as well as gaps between needs and policies. Additionally, university partners were asked to provide materials related to study programmes listed on newly emerging areas of topics.

As of 2020-05, the specifications from partners were expanded to the following totals:

Country	Area	Partners involved	Sources
BE	Flanders	VITO	154
CZ	Central Bohemia	NUVIT	445
ES	Segóbriga	TRAGSA Social Innolabs	150
FI	Häme	Hame University of Applied Sciences (HAMK) Smart & Lean Oy	580
GR	Central Greece	Agri University of Athens GAIA Central Greece Region Neuropublic	100
IE	Monaghan	MAC MID CLG	538
IL	Galilee	MIGAL	40
IT	Apulia	InnovAgritech Gal Murgia Piu - PPP Confagricoltura	41
LV	Vidzeme	Vidzeme Planning Region LRF AREI BOSC	226
MK	Gevgelija – Strumica	AgFutura Green Growth Platform	90
PL	Mazowieckie	ERDN	13
SK	Slovakia Region	Slovak University of Agriculture Slovak Rural Parliament / VIPA City of Nitra Agroinstitut	118
Total			**2495**

Inputs collected are sufficient for needs analysis and policy evaluation library scopes to meet expected success criterion of 1800 records as required by the grant agreement. The library will be further developed during the project with regard to the practical needs and experience of users with automatically processed results.

1.2 Proportionality

In order to improve proportionality, selected pilots were requested to identify additional resources focusing on regional needs and policies. Another roughly 280 sources were collected and are in the process of review and addition to total mentioned above at the time of writing.

1.3 Access

Regional Library contents are available as part of text-mining solution and to all partners through a dedicated online user interface https://semex.io/.

2 Future Expansion

2.1 Issues

Apparently, an objective lack of publicly available sources exists in cases of some non-English languages. In the context of Regional Library expansion, this problem could only be addressed in part. Further identification of less accessible (i.e. unofficial) sources could be achieved by cross-use of existing data across all pilots.

2.2 Opportunities

Future expansion of Regional Library in semi-automatic or automatic way will be discussed for future project stages. General access to initial version of CRAWLER technology was published for future project needs. It is available online, both programmatically (online and offline APIs) and manually through web browser (as shown below).

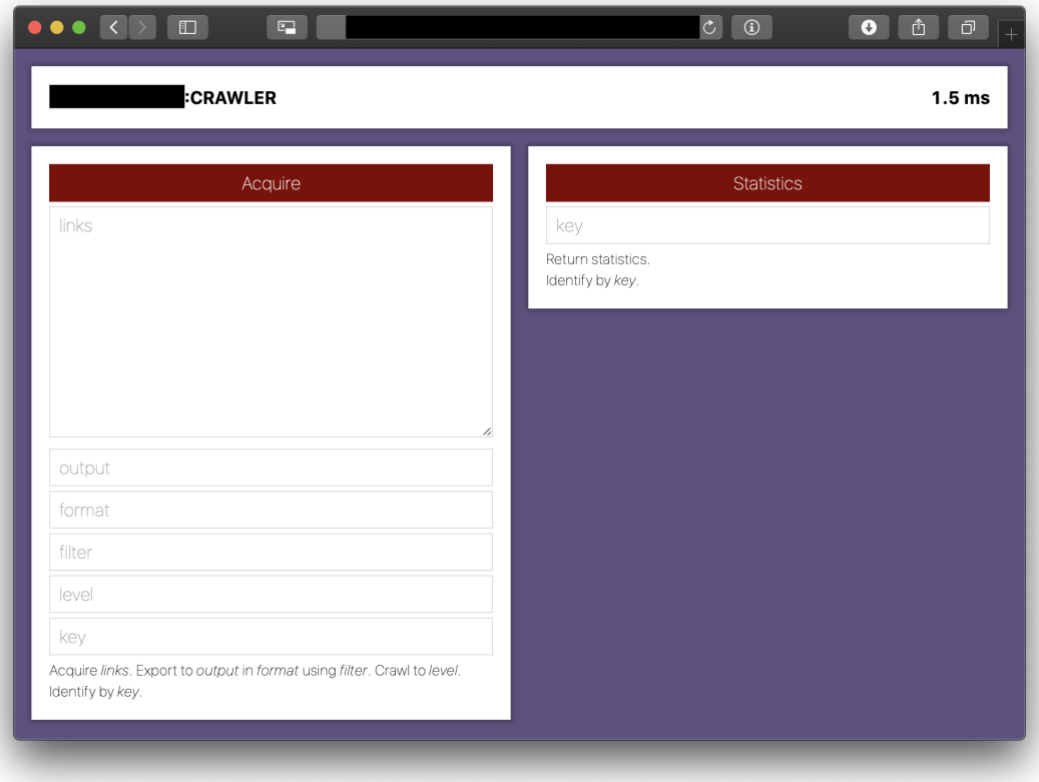

Image: screenshot of manual CRAWLER interface (server identification redacted for security reasons)

NUVIT's own PDF parsing technology was implemented for the CRAWLER in view of the expectation of the future need for greater control over processing of documents of extraordinary length, non-English character sets and tabular data. NUVIT's own classification

technology was implemented for the CRAWLER, utilization of which also will be discussed for next project stages.

3 Conclusion

The Regional Library is a vital source of information for the PoliRural text mining tool (see D 2.3 for details) that will be used primarily by researchers in foresight exercises. The library value depends on the number and quality of publicly accessible sources in the languages that are of interest in the PoliRural.

As the information extraction from non-English sources, particularly from PDFs containing complex tables, is technically challenging, the further development of the crawling technology will be highly beneficial for the performance of the text mining tool.

Annex 1 – Responses to the monitors' comments

Comment made by the monitors	Explanation
Overall: this builds on D4.1, expanding topics and specifications and is a short document that describes the development of the regional library for policy evaluation; systematic guidance on the contribution? It identifies the challenge of extracting information from non-English sources, especially pdfs. Developing the crawling technology will be important for assisting with this process. Original due date for this deliverable was end Feb, but it was extended to end May 2020 It is not clear if all partners entered the exante and interim evaluation for each national EU programmes such as EU Rural Development Programme and EU Regional Development; ESF etc. Since a systematic guidance on the coverage of all policy areas in charge for the regional development in rural areas, this affects D4.3 as well. This Deliverable will take into account the adjustments made by the consortium in the wake of the first interim review. For that reason, adjustments in D4.3 will provide guidance for the work of the pilot teams, which they will appreciate.	Based on the collection of the first set of resources, disproportions were identified in terms of the content and scope of coverage of individual areas of interest across the territories of the pilot solutions. It has been found that for some territories, the sources of authentic external information are severely limited (for homogeneous resources, see below). In addition, in terms of technology development and validation, English resources were naturally preferred in the initial stage. Therefore, the partners implementing the pilot projects were recommended to expand the initial resource survey in the categories of policy / strategy, need and need-policy gaps and also emerging topics. The sources identified in this way were subsequently added to the Regional Library. In the next stages, iterative expansion is expected, depending on the experience and needs of field experts with the outputs from the processing of this data. Engagement of field experts ensures relevance and validity of added sources. For the initial stage of the project (focused on the actual creation and start of technology verification), it was decided to use only individual documents that are identified - relevant to field experts. In the next stage, scaling (in terms of technology and content) is assumed, which will allow field experts to incorporate other sources (e.g. homogeneous - systematic datasets or

	collateral textual information). For these cases, crawler technology is suitable, which does not focus only on individual links, but on in-depth acquisition. The actual implementation of the PDF parser is a secondary tool, providing better control over aspects of the conversion reliability of this format into plain text, which is necessary for its further processing. Page no. 8.