SHERPA
Rural Science-Society-Policy Interfaces

D4.1 Framework for identification, selection and evaluation of past and ongoing research projects

31 January 2020

SHERPA receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 862448

FRAMEWORK FOR IDENTIFICATION, SELECTION AND EVALUATION OF PAST AND ONGOING RESEARCH PROJECTS

Project name	SHERPA: Sustainable Hub to Engage into Rural Policies with Actors
Project ID	862448
H2020 Type of funding scheme	CSA Coordination and Support Action
H2020 Call ID & Topic	RUR-01-2018-2019–D / Rural society-science-policy hub
Website	www.rural-interfaces.eu
Document Type	Deliverable
File Name	Framework for identification, selection and evaluation of past and ongoing research projects
Status	Final version
Dissemination level	Internal
Authors	Hercules Panoutsopoulos, Borja Espejo Garcia, Spyros Fountas
Work Package Leader	Agricultural University of Athens (AUA)
Project Coordinator	ECORYS

This document was produced under the terms and conditions of Grant Agreement No. 862448 for the European Commission. It does not necessarily reflect the view of the European Union and in no way anticipates the Commission's future policy in this area.

Table of contents

Executive Summary

The overall objective of SHERPA is to collect research-based knowledge with the aim to contribute and offer support to the formulation of recommendations for future policies relevant to EU rural areas. It aims to create a science-society-policy interface which provides a hub for knowledge and policy. By this way, actors will be supported with regard to rural development and will be helped to identify opportunities for protecting rural areas from long-term threats of decline. Effective use of knowledge gained from investment in research, and empowerment of key actors and stakeholders in the development of public policy lie at the core of SHERPA.

Capitalising upon existing knowledge requires the employment of a sound methodological approach able to ensure a systematically exploitation of available research. Within this context, the aim of the present document is to provide a detailed presentation of the SHERPA methodology for taking stock of past and ongoing findings of rural-related research.

The SHERPA methodology consists of a total number of four phases with each of these phases comprising a number of appropriately designed steps. Research in the context of SHERPA involves a balanced combination of human labour and technology-supported, automated tasks. As a result, there is potential to take advantage of the experience of domain experts and, at the same time, exploit the power of cutting-edge technologies.

The first phase of the SHERPA methodology, named as "Identification of topics", relates to the identification and definition of rural-related topics, which are going to be used as drivers for project search and research content extraction. Analysis of rural-related topics into sub-topics and association of topics with appropriately defined keywords are part-and-parcel of the rural topics' identification phase.

The "Search-Retrieve-Pool" phase of the SHERPA methodology that comes next is about searching, finding, retrieving and pooling relevant projects and information/details about them. Iterative execution of the search step coupled with human-driven supervision guarantee the appropriateness and quality of the results obtained.

The third phase of the methodology (i.e. "Extraction of research content and synthesis of outcomes") lies at the core of SHERPA research and is concerned with automated content extraction processes.

Reporting of content extraction outcomes is *sine qua non* for research in the context of SHERPA. Therefore, the fourth phase of the methodology (i.e. "Reporting research-related results") deals with issues having to do with the appropriate formatting of input for interactions at the Multi-Actor Platform (MAP) level.

The SHERPA methodology has been built on the basis of a systemic approach. The input that is needed for the execution of each step in each phase of the methodology, as well as produced output, are explicitly defined. Processing that takes place in the context of each step is described across a number of dimensions, namely the "what", "where", "how", and "who" of each step. By this way, all work undertaken from the initiation of project search to finalisation of the SHERPA Discussion Paper is made transparent and clear.

1. Introduction

The goal of WP4 is to take stock, analyse and produce a web-based repository of outcomes of past and on-going research projects related to rural topics, focusing on those funded by the European Union. Task 4.1, in particular, relates to development of a methodology for systematic identification, retrieval and review of completed and on-going projects. The methodology described, includes a number of phases and constituent steps that cover a number of tasks ranging from the search and identification of projects of relevance to the topic being considered, through the reporting of research findings and results. Criteria for scanning existing databases and identifying relevant projects are listed, and report templates are provided. This document starts with an overview of the SHERPA methodology, which is expanded in detailed descriptions of each project phase and constituent steps.

2. SHERPA methodology for the retrieval and evaluation of past and ongoing research projects

2.1. Overview of SHERPA methodology

2.1.1. SHERPA methodology at a glance

Figure 1 below provides an overview of the methodology that is employed for the implementation of research in the context of the SHERPA project.

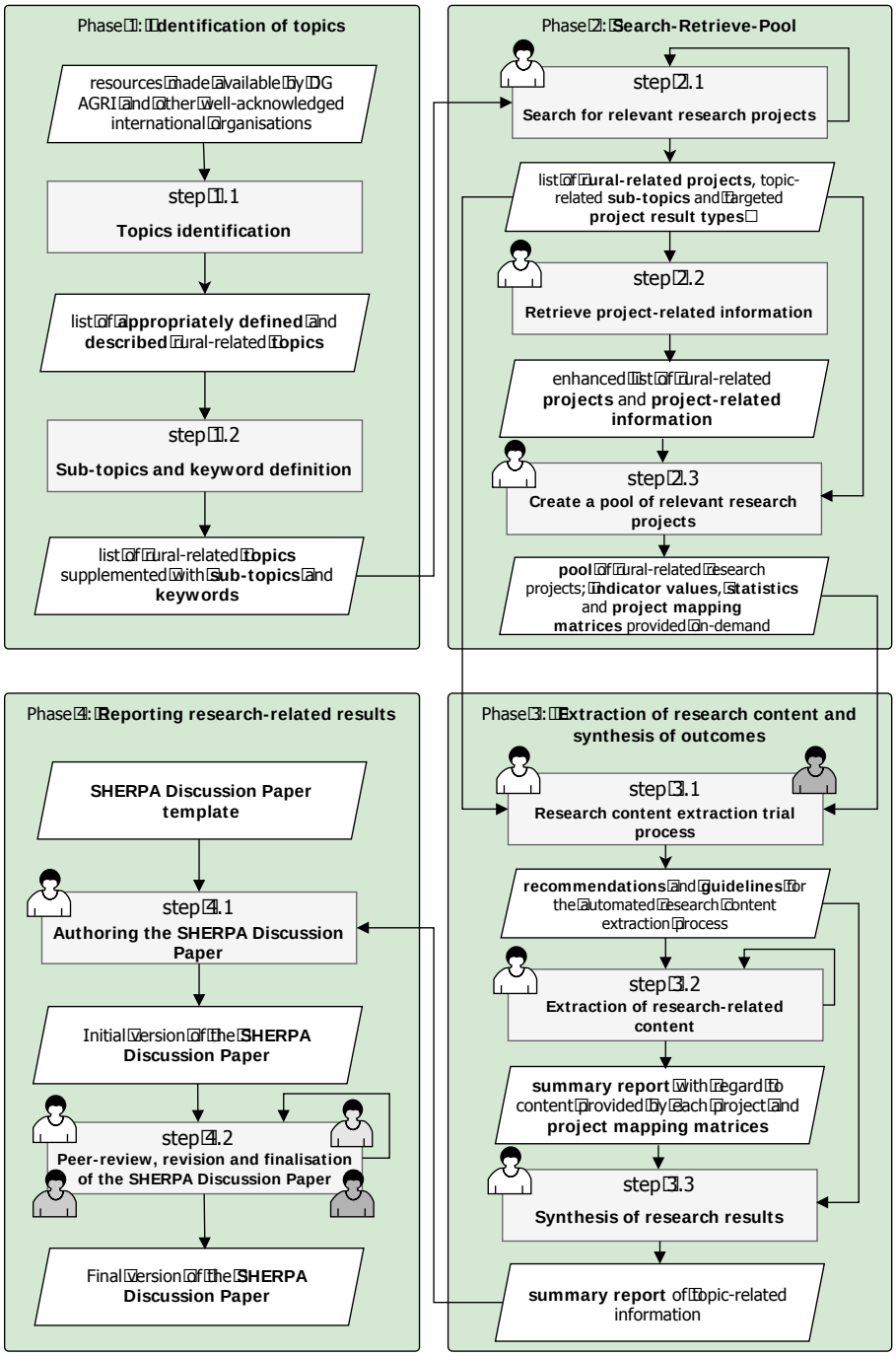
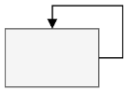
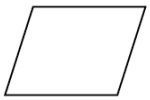
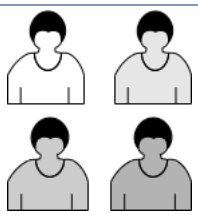


Figure 1: Overview of SHERPA methodology

The visual notations used in Figure 1 are explained in Table 1 below.

Table 1: Explanation of visual notations used in the description of SHERPA methodology provided in Figure 1

Visual notation	Explanation
	Denotation of a phase of the SHERPA methodology.
	Denotation of the execution of a step involved in a phase of the SHERPA methodology.
	Denotation of iterative execution of steps involved in a phase of the SHERPA methodology.
	Denotation of input to, or output from, steps involved in a phase of the SHERPA methodology.
	Denotation of connections between steps in the methodology, input to a step and output from a step.
	Denotation of the roles involved in the execution of a step in the SHERPA methodology.

2.1.2. Roles

This section presents the roles associated with the phases and steps of the methodology employed in the content of the SHERPA project, and outlines the tasks in which each role is involved. Phase- and task-related involvement of each role is also described in the following sections (**Sections 2.2-2.5**), in which the SHERPA methodology is analytically presented.

Scientific Editor

The Scientific Editor is member of a SHERPA consortium partner. The person assuming this role is different for each rural topic with the role assignment being based on an attempt to match the Scientific Editor's profile with particularities of the topic at hand. The Scientific Editor is responsible for the following tasks:

- Assessment of the appropriateness of results obtained from the automated project search and filtering of results.
- Refinement of search queries and involvement in further manual research with the aim of obtaining more results of relevance.
- Contribution to the creation and update of sub-topic lists and the types of project results to be reviewed for the relevance of their content.
- Supervision of tasks relating to retrieval of rural-related project information.
- Recommendations with regard to indicators and statistics to be extracted from project-related information.

- Leading role in the trial process of extracting content from project documents.
- Review and editing of automatically produced project summaries.
- Supervision of the process of extracting research content and reviewing of outcomes.
- Authoring of the SHERPA Discussion Paper and integration of changes and corrections based on feedback received.

Communication Editor

The **Communication Editor** is member of a SHERPA consortium partner and has the task of making comments and editorial suggestions related to **provided information** and **layout issues**. Feedback from the Communication Editor is used in making changes and revisions, and then the production of the final draft of the document.

Review Editor

The **Review Editor** is member of a SHERPA consortium partner and has the task of providing comments and feedback about the content of the SHERPA Discussion Paper and the degree to which it achieves the intended goals (i.e. to facilitate discussion and interactions in the context of Multi-Actor Platforms). Feedback from the Review Editor is used for making changes and revisions, and then the final draft of the document.

Support Staff

Support Staff are members of the Agricultural University of Athens (AUA) who are concerned with the following tasks:

- Search for project-related documents and extraction of content out of them as part of the trial process of research context extraction from project documents.
- Supports the Scientific Editor in the implementation of further manual research for rural-related projects.
- Supports the Scientific Editor in integrating changes in the SHERPA Discussion Paper (after review and comments).

2.2. Phase 1: Identification of topics

The starting point of the SHERPA methodology for research implementation in the context of the project is a phase dedicated to: (i) the **identification** of rural-related **topics**; (ii) the **subdivision** of each topic into **sub-topics**; and (iii) the association of each topic with a **list of keywords**. This phase lays the foundations for: (i) **searching for** and **identifying** projects (completed or in progress) associated with each topic; and (ii) **extracting** information from documents produced as project outputs. Based on the above, the phase of the "**Identification of topics**" of the SHERPA methodology consists of the following **two** (**2**) **steps**:

- **Step 1.1**: Topics identification; and
- **Step 1.2**: Sub-topics and keyword definition.

Each of these steps is described in detail in the following subsections with a focus on the "**what**", "**where**", "**how**" and "**who**" of tasks involved in each step. These are accompanied by references to the **inputs** required for the execution of each step, the **outputs** produced and the **challenges** that may need to be addressed as part of each step.

2.2.1. Step 1.1: Topics identification

WHAT

The "**Topics identification**" step relates to the creation of a list of rural-related topics together with topic definitions and descriptions that will drive the search for relevant projects funded by the European Union.

WHERE

Sources of the identification of rural topics are **documents**, **reports** and **policy papers** of the European Commission's Directorate General for Agriculture and Rural Development (DG AGRI) and recognised international organisations such as the OECD, FAO, World Bank and UNESCO. The knowledge and resources of **SHERPA partners** also provide sources of input for the task of identifying topics.

HOW

Identification and definition of rural-related topics take place as combination of a **top-down** and a **bottom-up** approach. In the case of the top-down approach, members of SHERPA partners are involved in processes of information and content search to be used as input for identification and appropriate definition of rural-related topics. Search for resources that can provide access to such kind of information is implemented in **websites** and **open access databases** of organisations such as DG AGRI, OECD, FAO, World Bank and UNESCO. DG AGRI's contribution to this top-down identification of rural topics is also envisioned.

The bottom-up approach involves rural topic – related recommendations coming from interactions that take place in the context of Multi-Actor Platforms (MAPs). As particularly mentioned in SHERPA's **Dynamic Action Plan**, Multi-Actor Platforms propose rural topics of their own interest to be considered for the final topic list that SHERPA research will focus upon.

The final rural-related topic list is produced from the **matching** of topics identified and outputted from both the top-down and bottom-up approaches.

WHO

SHERPA partners conduct research to identify resources with the capacity to provide inputs to the definition of rural-related topics. Recommendations and input with regard to rural-related topics to be investigated are obtained from the **European Commission** (**DG AGRI**) and **Multi-Actor Platforms** (**MAPs**).

INPUT

Inputs for the identification and definition of rural-related topics are:

- **Information** and **content** retrieved from DG AGRI, OECD, FAO, World Bank and UNESCO (e.g. documents, reports, policy papers), and from using knowledge and resources of SHERPA partners.
- **Recommendations** provided by Multi-Actor Platforms (MAPs).

OUTPUT

The output of the "**Topics identification**" step is a list of appropriately defined and described rural topics which are ready for use for the needs of relevant project search.

CHALLENGES THAT NEED TO BE ADDRESSED

Identification and definition of rural-related topics requires a detailed mapping of: (i) the information and content needed as inputs for this task; and (ii) the sources from which information and content can be retrieved. To this end, there needs to be careful identification of: (i) the types of documents to be reviewed for relevant information and content; and (ii) the organisations and knowledge bases that can provide the required resources. In addition to the above, attention needs to be paid to the process of matching of topic lists that have been created as part of both the top-down and bottom-up approaches.

2.2.2. Step 1.2: Sub-topics and keyword definition

WHAT

The second step of the "**Identification of topics**" phase relates to subdivide the set of identified rural-related topics into **sub-topics**, and associating them with **lists of keywords**. This breakdown of rural topics into smaller elements will provide insights to the topics at a higher level of detail, aiming to be useful for the phase of project searches. The lists of sub-topics, assigned to each topic, are refined in the next phase ("**Search-Retrieve-Pool**" phase; see **Section 2.3**), in which information from identified and retrieved projects is considered.

WHERE

Input for the analysis of rural-related topics into sub-topics and their association with keywords is sought from **documents**, **reports** and **policy papers** produced by organisations such as DG AGRI, OECD, FAO, World Bank and UNESCO. In addition, the **knowledge of SHERPA partners** can also be exploited for the needs of execution of the tasks involved in this step.

HOW

The goals of this step are achieved by a thorough review of the literature retrieved from the above-mentioned sources. A specific focus is posed on **relevant terminology** and existing **vocabulary terms** appropriate for the purpose of the tasks.

WHO

The subdivision of rural topics into sub-topics and the production of keyword lists is through the involvement of **SHERPA partners**, **Multi-Actor Platforms** (**MAPs**) and the **European Commission**.

INPUT

The list of topics produced in the previous step (i.e. "**Topics identification**") is an input to the step of "**Sub-topics and keyword definition**".

OUTPUT

The outputs from the "**Sub-topics and keyword definition**" step is a list of rural-related topics, each of which is supplemented with associated sub-topics and keywords.

CHALLENGES THAT NEED TO BE ADDRESSED

Sub-dividing topics into sub-topics and associating them with keywords requires a careful mapping of: (i) literature (types of documents to be reviewed for relevant content); and (ii) sources from which relevant literature can be retrieved.

2.3. Phase 2: Search – Retrieve – Pool

The second phase of the SHERPA methodology is a pipeline of steps that starts with the **search** for research projects about rural topics, continues with the **selection** of projects that are relevant to the search scope, and **retrieval** of project-related information, and concludes with the **pooling** of relevant results. The execution of project searches, retrieval of project-related information, and pooling of results is based on automated tools. Whether supported by technologically advanced search engines or taking place through browsing paper-based material, seeking for information is a common activity that involves specific patterns of behaviour and actions. Such patterns have been extensively investigated and documented in the context of existing models such as the Information Search Process model (Kuhlthau, 1991) and the David Ellis model (Ellis et al., 1993; Ellis & Merete, 1997). What is evident from the review of available literature is that seeking

for information is an iterative process with its termination condition depending on the number of results and their quality and appropriateness with regard to the scope of search. The execution of further search iterations is decided on the basis of appropriately defined thresholds and result evaluation criteria. As Rosenfeld *et al.* (2015) pointed out, the search and browsing of results and involvement in search iterations are building blocks of behaviour of the information seeking process. The revision of the set of search terms in each iteration is a necessary condition for the optimisation of results. The success of the process of seeking for information is directly related to the set of keywords and search terms employed, and refinements implemented from one iteration to another.

The "**Search-Retrieve-Pool**" phase of the SHERPA methodology consists of **three** (**3**) **steps**, namely:

- **Step 2.1**: Search for relevant research projects;
- **Step 2.2**: Retrieve project-related information; and
- **Step 2.3**: Create a pool of relevant research projects.

Each of these steps is described in detail in the following subsections with a focus on the "**what**", "**where**", "**how**" and "**who**" of tasks involved in each step. In addition to that, there are references about the **input** required for the execution of each step, the **output** produced, and the **challenges** to be addressed as part of each step.

2.3.1. Step 2.1: Search for relevant research projects

WHAT

The first step of the "**Search-Retrieve-Pool**" phase is concerned with the systematic search for research projects funded by the European Union and dealing with topics that have the "**rural**" concept at their core. Search results lead to the identification and pooling of relevant projects to be further scanned for the needs of extracting rural-related content and information. According to Rosenfeld *et al.* (2015), there are different types of information seeking targeting at different information needs. These types are: (i) the **known-item** type; (ii) the **exploratory** type; (iii) the **exhaustive** type; and (iv) the **revisiting-already-identified-pieces-of-information** type. Figure 2 below illustrates these information seeking types together with the information needs they serve.

In the case of the **known-item** information search type, there are very specific information needs to be satisfied and, consequently, a specific result that corresponds to that need. The identification and retrieval of the item of information required indicates termination of the search process. **Exploratory** search is the opposite of the known-item search. In this case, no prior knowledge of the search results needs to be obtained, and the use of search iterations is the norm. The quantity and quality of results obtained are assessed at the end of each iteration and the execution of each iteration, takes place on the basis of a revised and refined set of search terms. Termination of the process takes place when the number and appropriateness or quality of the results retrieved are considered to be satisfactory. As far as the **exhaustive** type is concerned, the agent performing the search operation aims to find and retrieve all of the results that may relate to a topic and relevant search terms. Finally, in the case of the **revisiting-already-identified-pieces-of-information** type, the focus is on methods that support the revisiting and reuse of previously identified information sources (i.e. information sources obtained as search results at earlier stages without being clear whether it is useful or not at the time of their retrieval).

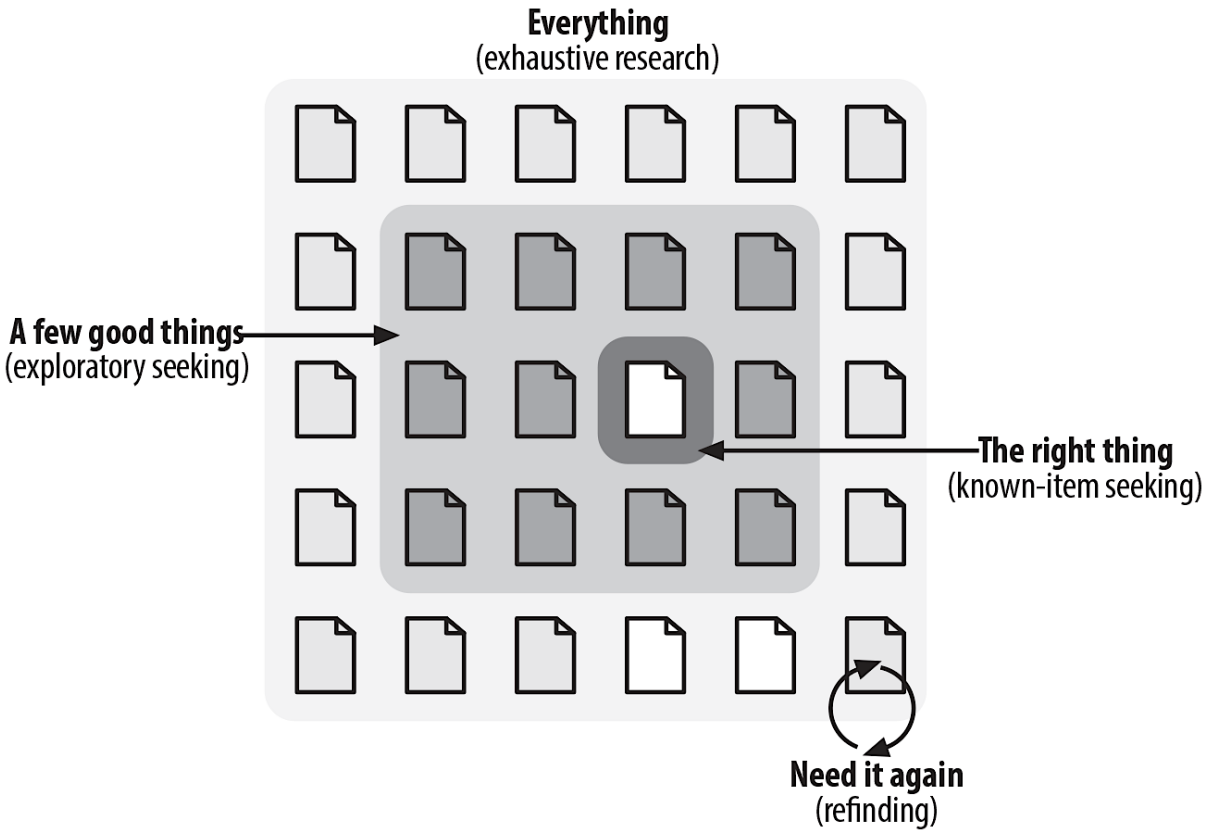


Figure 2: Types of information seeking and targeted information needs (Rosenfeld et al., 2015)

The project-related search executed in the context of SHERPA comprises the **exploratory** and **exhaustive** search types recognising that: (i) all relevant, rural-related projects cannot be known in advance; and (ii) the necessity of finding as many relevant rural-related projects as possible. Characteristics of the **revisiting-already-identified-pieces-of-information** search type also comply with particularities of the SHERPA-related research which are that projects found in one search iteration may prove to be useful at a later stage.

The "**Search**" step also involves refinement of the sub-division of each rural topic into a number of **sub-topics** as implemented previously (in the "**Identification of topics**" phase). The list of sub-topics assigned to each rural topic emerges by an initial analysis of project-related information and project objectives. After each iteration of project searching, the list of topic-related sub-topics is revised by taking into account the new findings. A list of **types of project results** is then proposed. These refer to types of information and content that have been produced as project outputs (e.g. data/indicators, policy interventions, recommendations, etc.).

WHERE

Projects that deal with rural-related topics are searched in databases of project funding programmes. As of January 2020, these programmes are:

- FP6 (https://cordis.europa.eu/projects/en),
- FP7 (https://cordis.europa.eu/projects/en),
- H2020 (https://cordis.europa.eu/projects/en),
- ESPON (https://apps.espon.eu/db2/),
- LIFE (http://ec.europa.eu/environment/life/project/Projects/index.cfm),

- ERA-NET (https://ec.europa.eu/research/fp7/index_en.cfm?pg=eranet-projects), and
- INTERREG (https://www.interregeurope.eu/library/).

In addition, **web-based sources** able to be searched for and found by human agents through use of **general-purpose search engines** are also considered.

HOW

The search for projects is closely associated with **query building**. Identification of projects that deal with specific rural topics requires the use of search queries built upon appropriately selected, topic-related keywords. Search queries need to consist of search terms indicative of the feedback required with regard to research projects and their relation to rural topics. Thus, query building is based on lists of keywords closely linked to the rural topics under investigation. From a technical point of view, the submission of queries and execution of the search task is based on the use of a search automation tool (i.e. a crawler).

Search for research projects of relevance is also complemented by:

- Engagement of the Scientific Editor and Support Staff to "**manual**" project search through use of general-purpose search engines.
- Utilisation of **partnership networks of SHERPA partners** with the aim to identify projects of relevance that may have not been properly sought for.

The project search task also has a temporal dimension. The search is not confined to completed projects but also considers ongoing projects. For this purpose, the iterative execution of the search step (at an appropriate frequency) is recommended for finding newly funded projects, which qualify for the SHERPA research. The exploitation of **knowledge sources** and **information** available to DG AGRI, SHERPA partners and partner networks are likely to provide additional channels for identifying ongoing projects of relevance. Given the importance of monitoring to identify new projects, and networking with them, more details on this issue are provided in **Section 3**.

Criteria for the identification and selection of relevant projects are:

- The frequency in which a project appears as a result of the execution of different search queries.
- The number of search terms found in the textual description of a project that has been returned as a result of a search query.
- The frequency of appearance of a project in the list of top-N results (where N = 5, 10, 20, 30).
- Contextual information (assessed as part of a review conducted by the Scientific Editor).

The list of sub-topics assigned to each rural-related topic and the list of project result types are revised and updated after each search iteration by using more information. When searches for projects relating to a rural topic are complete, a list of topic-related sub-topics is made available. The list of project result types is updated after completion of: (i) the execution of each iteration of project searches in the context of the investigation of specific topics; and (ii) investigation related to a rural topic. A complete list of project result types is available upon completion of the "**Search**" step.

WHO

Project search is implemented in an automated tool, developed and maintained by the Agricultural University of Athens (AUA), with the support of the **Scientific Editor** (see **sub-section 2.1.2** for the description of the Scientific Editor's role). The Scientific Editor is responsible for the assessment of the results obtained, and their filtering. Based upon the quality of results, the Scientific Editor decides whether more search iterations will be required. In this case, the Scientific Editor needs to proceed to a refinement of the search queries to be used. They may also conduct manual research in order to develop the list of results with regard

to rural-related projects. The Scientific Editor creates and updates lists of topic-related sub-topics and project result types in light of newly obtained search findings.

The **Support Staff** provides assistance to the Scientific Editor with regard to execution of "manual" search for projects of relevance.

INPUT

The step related to project search is executed by using inputs from: (i) a list of **rural-related topics** and topic-related **sub-topics**; (ii) sets of topic-related **keywords**. Inputs to the "**Search**" step are produced in the "**Identification of Topics**" phase (described in **Section 2.2**).

OUTPUT

The output of the "**Search**" step is a final list of projects considered by the Scientific Editor to be relevant to the rural-related topic under investigation. This list provides inputs to the "**Retrieve**" step that follows. An output from the "**Search**" step is lists of topic-related sub-topics and a list of project result types.

CHALLENGES THAT NEED TO BE ADDRESSED

The effective use of keywords has to do with the way in which they are combined in the development of the query. Search queries can be built by combining keywords through the use of logical operators. **Logical operators** or **logical connectives** are **symbols** or **words** used in **mathematical logic**[1] to "*connect two or more sentences (of either a formal or a natural language) in a grammatically valid way, such that the value of the compound sentence produced depends only on that of the original sentences and on the meaning of the connective.*"[2]

The most frequently used logical operators are summarised in Table 2, together with a short explanation of their meaning.

Table 2: List of logical operators that are most frequently used, their mathematical notations and brief explanations of their meaning

Logical Operator	Mathematical Notation	Explanation
AND	"∧"	**conjunction** **A** ∧ **B** is **true** only if A is true **and** B is true[3]
OR	"∨"	**disjunction** **A** ∨ **B** is **true** if A is true, **or** if B is true, **or** if both A and B are true[4]
NOT	"¬"	**negation** **Negation of a proposition P** (¬ **P**) is the proposition whose proofs are the **refutations of P**[5]
XOR	"⊻"	**exclusive disjunction** Logical operation (**P** ⊻ **Q**) that outputs true only when inputs differ (**one is true, the other is false**)[6]

[1] According to Wikipedia (https://en.wikipedia.org/wiki/Mathematical_logic), **mathematical logic** is "*a subfield of mathematics exploring the applications of formal logic to mathematics. The unifying themes in mathematical logic include the study of the expressive power of formal systems and the deductive power of formal proof systems.*"

[2] Definition of the term "**logical operator**" retrieved from Wikipedia (https://en.wikipedia.org/wiki/Logical_connective).

[3] https://en.wikipedia.org/wiki/Logical_conjunction

[4] https://en.wikipedia.org/wiki/Logical_disjunction

[5] https://en.wikipedia.org/wiki/Negation

[6] https://en.wikipedia.org/wiki/Exclusive_or

In the case of a list of keywords comprising 6 keywords, the number of search queries that can be built upon use of a single logical operator combining two operands (e.g. the "AND" and "OR" operators) is 30 (i.e. 6 different keyword options for the first operand x 5 different options for the second operand, since there is no point in using the same operand twice). Based on the example above, there is a large number of search queries that can be built by combining keywords through the use of logical operators. The automation of project searches removes the burden of human labour. The use of appropriate, predefined search queries, can make the search process more effective.

The execution of queries developed through use of specific keywords may not yield the expected results, because of the absence of keywords employed in the text that has been searched. However, the availability of synonyms per keyword or search term can enhance the search process by using them as alternative search terms and so help obtain better results.

2.3.2. Step 2.2: Retrieve project-related information

WHAT

The second step of the "**Search-Retrieve-Pool**" methodology phase relates to **retrieval of information** about projects. The final project list, per rural topic, augmented with project-related information is the input to the "**Pooling**" step that comes next.

WHERE

Information about rural-related research projects is obtained from databases of research project programmes that have been queried for finding relevant projects. Project websites are used as sources of project-related information as well. Figure 3 presents an example of project-related information made available from Cordis database (https://cordis.europa.eu/). A first layer of provided information is, among others, about project acronym and full name, project ID, project programme/framework, funding, project coordinator and partner details, and funding given to each partner. Information about project objectives and achieved results, deliverables and project-related reports, as well as publications made in the context of the project constitute a second layer of provided information.

HOW

The retrieval of project-related information is implemented by use of a web scraping tool, developed and maintained by the Agricultural University of Athens (AUA). Web scraping is the collection of data from web-based sources and as Mitchell (2018) points out, it can be automatically implemented through execution of a software application that "*queries a web server, requests data (usually in the form of HTML and other files that compose web pages), and then parses that data to extract needed information*."

WHO

The type of information needed to be retrieved and stored per research project is identified by the **Scientific Editor**. Project-related information can be exploited for the creation of a pool of projects, and project-related information, relating to the list of rural-related topics that have been identified in the first phase. The Scientific Editor has the responsibility of supervising the tasks that are involved in the "**Retrieve**" step.

INPUT

The input required for execution of the "**Retrieve**" step is the final listing of projects from the "**Search**" step which are relevant to the rural-related topic.

OUTPUT

The output of the "**Retrieve**" step is an enhanced list of projects, per rural-related topic, containing project-related information. Together, the enhanced lists of projects make up a pool of rural-related research projects that is stored in the SHERPA repository.

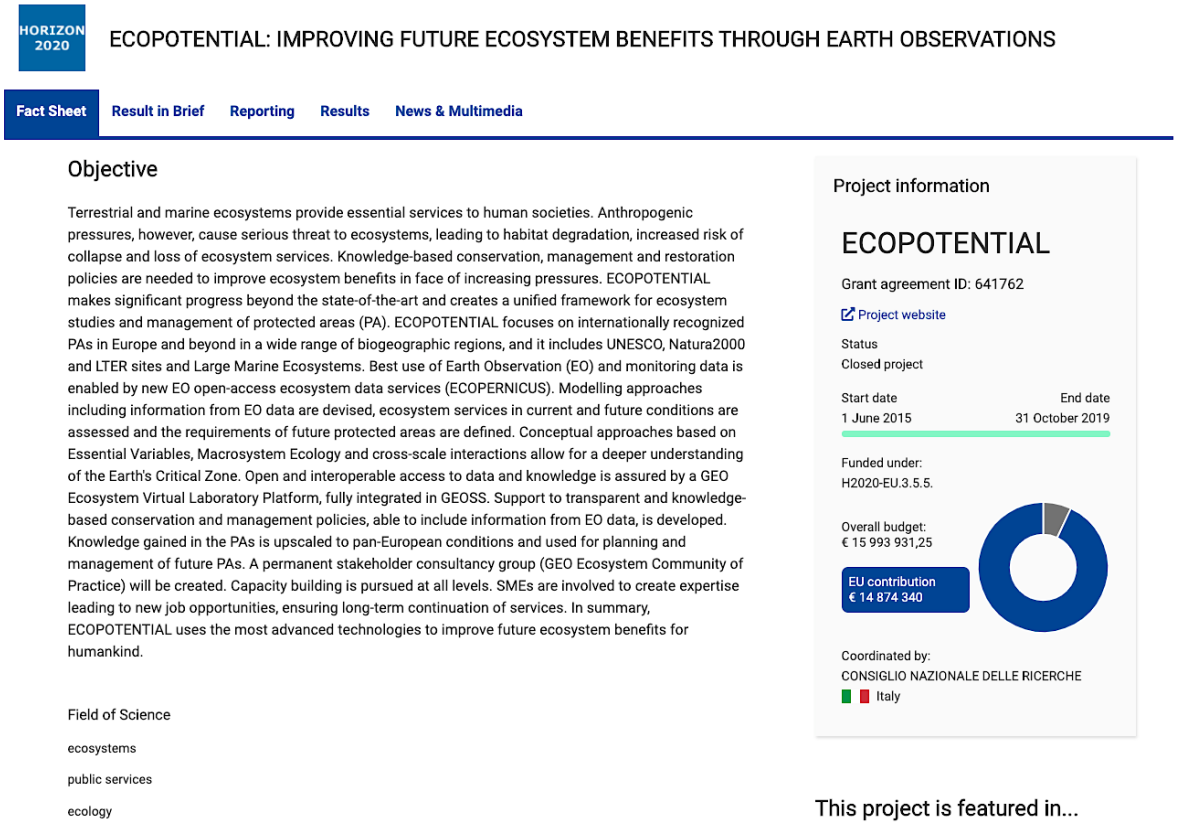


Figure 3: Information provided by the Cordis database about the ECOPOTENTIAL project

CHALLENGES THAT NEED TO BE ADDRESSED

Databases in which projects can be searched, and from which project-related information and project products can be retrieved, differ in structure and do not provide information in a consistent way. As a result, there is no one-size-fits-all solution for the use of scraping tool solutions that can be used for fetching project-related information. Thus, there is need for the development of custom-made solutions tailored to the particular needs of each database interrogated.

Heterogeneity in the provision of project-related information, by different project databases, has implications for the type of information that can be collected. To supplement the available project list per rural topic, with project-related information, there is a need for a careful identification and definition of properties with regard to the information to be identified and collected. In a further step, this list of properties serves the purpose of informing the creation of a database model necessary for storing project-related information in the SHERPA repository.

2.3.3. Step 2.3: Create a pool of relevant research projects

WHAT

Following the "**Retrieve**" step, in the "**Search-Retrieve-Pool**" phase, is the "**Pool**" step. This step is linked to the creation of a pool of information about rural-related projects and project-related information associated

with rural topics. Stored information is exploited for on-demand extraction of indicator values and statistics able to provide useful insights into patterns and trends.

Example **indicators** are:

- Frequency of references to rural-related topics.
- Number of completed and ongoing research projects.
- Number of research projects per topic across different time periods.
- Number of projects per sub-topic.
- Most frequent project result type.
- Minimum, maximum and average funding per rural-related topic.

Indicative examples of **supporting statistics** are:

- Distribution of projects across research project programmes per rural-related topic.
- Distribution of projects across topic-related sub-topics.
- Project budget per research project programme in the context of each topic.
- Distribution of coordinating organisations across European countries.

The list of sub-topics assigned to each rural topic together with lists of identified project result types, provided as input from the "**Search**" step, are used for a more in-depth project-related analysis leading to creation of **mapping schemes** according to which projects can be mapped with regard to: (i) **topic** and **sub-topics** they relate to; (ii) **types of results** that they produce.

WHERE

The pool of rural-related research projects and project-related information is stored in the SHERPA repository. Indicator values, statistics and project mapping matrices are extracted on-demand and delivered to the user through appropriate visualisations. The user is able to retrieve results in the form of **reports** able to be downloaded and stored locally in his/her digital device.

HOW

Information about rural-related research projects is stored in SHERPA repository with the help of values assigned to project-related properties identified and defined in the "**Retrieve**" step. This set of properties is used for creation of the database model required for storing the pool of rural-related projects into the SHERPA repository.

The extraction of indicator values and statistics is achieved with the help of appropriately developed code scripts executed using stored data. A number of fit-for-purpose libraries which are made available by all widely adopted programming languages (e.g. for establishing connection with the database in which data is stored) are used for this task.

WHO

The storage of the pool of rural-related projects into the SHERPA repository is a task that is implemented by the Agricultural University of Athens (AUA). The Agricultural University of Athens has the responsibility for the development of the code scripts that are executed upon user request for calculation of indicator values and statistics extraction, as well as creation of project mapping matrices. Statistics and indicator values that can be provided on-demand are defined after recommendations by the **Scientific Editor**.

INPUT

The enhanced lists of projects (i.e. lists of identified projects supplemented with information for each project), which are produced per rural-related topic and constitute the output of the "**Retrieve**" step, are the input for the creation of the pool (inventory) of rural-related research projects. The "**Pool**" step also receives the lists of topic-related sub-topics and project result types as input from the "**Search**" step.

OUTPUT

The "**Pool**" step outputs a pool of rural-related research projects stored in the SHERPA repository. It also outputs indicator values, statistics and project mapping matrices, which are made available on-demand and displayed through appropriate visualisations. These results can be downloaded and stored locally in the form of a report.

Indicator values, statistics and project mapping matrices are extracted through the execution of appropriately developed code scripts that use stored data (i.e. information from the pool of rural-related projects) as input.

CHALLENGES THAT NEED TO BE ADDRESSED

Coming up with a set of properties to be used for annotating project-related information requires a thorough review of the way in which project summaries are offered as the output of search queries of different project programme databases. This review needs to focus upon answering questions such as:

- What type of information is made available by provided project summaries?
- How is project-related information structured and displayed?
- Are there any similarities in information that is being provided?
- What are the differences with regard to project-related information and the way it is delivered?

2.4. Phase 3: Extraction of research content and synthesis of outcomes

"**Extraction of research content and synthesis of outcomes**" is the third phase of the methodology of SHERPA and is core to the research undertaken. It actually relates to the **identification of project outputs** (i.e. documents) of potential use for the needs of SHERPA-related research, and the **extraction of content** from identified outputs and the **synthesis of results retrieved** for use in reporting. More specifically, the "**Extraction of research content and synthesis of results**" phase consists of **three** (**3**) **steps**, namely:

- **Step 3.1**: Research content extraction trial process;
- **Step 3.2**: Extraction of research-related content; and
- **Step 3.3**: Synthesis of research results.

Each of these steps is described in detail in the following subsections with a focus on the "**what**", "**where**", "**how**" and "**who**" of the tasks involved in each step. In addition, details are included of the **inputs** required for the execution of each step and the **output** produced. **Challenges** to be addressed as part of each step are also highlighted.

2.4.1. Step 3.1: Research content extraction trial process

WHAT

The "**Extraction of research content and synthesis of outcomes**" phase starts with a step that aims to **simulate the process of extracting research content** from the list of identified projects and relevant document types (deliverables, reports, publications, etc.). This takes into account the types of results which have been proposed for consideration (e.g. data and indicators, policy interventions, policy

recommendations). The primary goals of this step are: (i) the **verification** of the types and list of project results required; and (ii) the creation of a set of **recommendations and guidelines** for the automated content extraction step.

WHERE

The content extraction trial process draws upon a limited number of topics (for example, one or two rural-related topics) from the topic list produced in the previous step. The search for relevant content is made in **project-related documents** (i.e. deliverables, reports, publications) retrieved from the available **project programme databases** and other **web-based sources** (see **Sub-section 2.3.1.2**).

HOW

The **search** for project-related documents and **reviewing** of research results with the aim of **extracting** useful, topic-related content, is a manual task. The trial process includes two tasks. The first task relates to the **search and identification** of project-related documents including content that may be relevant to the chosen topic. The search is made of existing databases, and other web-based sources, by use of keywords associated with the proposed types of research results sought (e.g. data and indicators, policy interventions, recommendations, etc.). The second task is about **searching for** and **extracting content from** project documents which appear to contain information of relevance to the rural topic being tackled, and targeted project result types. The search for content uses the list of sub-topics and keywords which are associated with each rural-related topic in the first phase. In this way, conclusions are drawn with respect to whether topic-related keywords which are already available, will be sufficient or if **more specialised keywords are needed**, for each topic, in order to find (and extract) relevant content.

WHO

The search for project-related documents and extraction of content out of them, as part of the trial process, are tasks performed by the **Scientific Editor** in close collaboration with the **Support Staff**.

INPUT

The input required for execution of the "**Research content extraction trial process**" step are: (i) the list of rural-related topics together with associated sub-topics and keyword lists (provided as output of the "**Identification of topics**" phase); (ii) the list of projects that are linked to each rural-related topic (made available as outputs of the "**Search-Retrieve-Pool**" phase); and (iii) the list of results of the types of project outputs (also provided as output of the "**Search-Retrieve-Pool**" phase).

OUTPUT

The output of this step is a **set of recommendations and guidelines** for the automated content (search and) extraction process that follows. These recommendations have to do with: (i) whether available keyword lists are efficient to search for relevant project-related documents and relevant content into them, or lists with more detailed or specific keywords are required; (ii) the types of project-related documents that appear to be more useful for content extraction; and (iii) the structure of project-related documents and more specifically, clues about document sections that appear to be "richer" in relevant content. The set of recommendations and guidelines that is produced as output of this step also provides a project **mapping scheme** by taking into account the list of potential project result types. This scheme is utilised for the creation of project mapping matrices (e.g. mapping of projects with regard to the type of data or indicators they have produced as outcome).

CHALLENGES THAT NEED TO BE ADDRESSED

The topic(s) that are used to guide the trial process of content extraction need(s) to be selected appropriately so as to facilitate insights into the issues of: (i) search for project-related documents; and (ii) extraction of

content from identified documents. In addition to that, recommendations and guidelines need to be provided in a form that facilitates uptake from the "**Extraction of research-related content**" step which follows.

2.4.2. Step 3.2: Extraction of research-related content

WHAT

During the past few years, there has been an exponential growth in production of text made available in a range of digital forms and formats. Information contained in text is valuable for communication and collaboration activities within and across various domains and disciplines, which are consequently heavily dependent upon the extraction of value from collections of unstructured textual data. Making meaning out of large text volumes requires the use of a wide range of analysis methods. **Text Mining**, also named **Knowledge Discovery from Text,** provides valuable solutions through methods and tools for textual data processing and analysis that utilise techniques from Information Retrieval, Information Extraction and Natural Language Processing, in conjunction with Knowledge Discovery, Data Mining and Statistics models (Hotho, Nürnberger & Paaß, 2005, p. 22).

"**Extraction of research-related content**" is a step that involves automated extraction of content, out of project-related documents, needed to be used for providing input to interactions at the Multi-Actor Platform (MAP) level. Extracted content helps to obtain insights into the rural topic at hand by also taking into account of the range of potential project results (e.g. data and indicators, policy interventions, policy recommendations). Automated extraction of content takes place through use of tools developed upon fit-for-purpose models and algorithms. Recommendations and guidelines made available by the "**Research content extraction trial process**" step help to provide a focus with regard to sources of content that need to be considered.

In the context of the "**Extraction of research-related content**" step, there is also a focus on the creation of matrices presenting mapping of projects with regard to identified project result types.

WHERE

An automated search for documents from relevant research projects, with the aim of identifying and extracting useful content takes place in **project programme databases** that have been initially searched with the aim of finding topic-related projects. Project-related information that have been retrieved and stored in the previous phase (see **Sub-sections 2.3.2** and **2.3.3**), help to easily identify sources (i.e. documents) of useful topic-related content. Extracted and retrieved content is stored in the SHERPA repository by use of an appropriately developed database model. This model is based on the **template** for project-related content reporting shown in Figure 4.

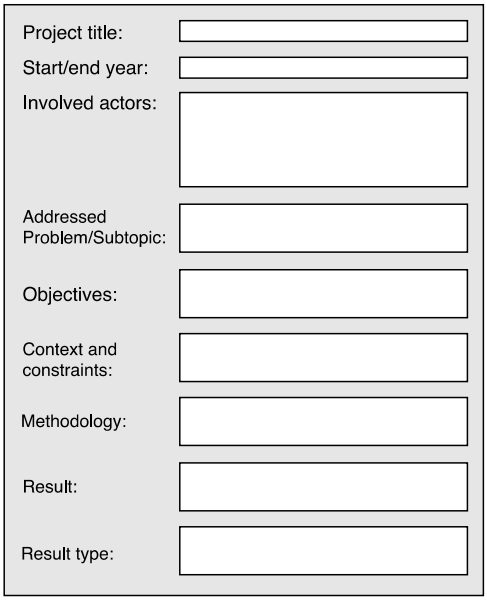


Figure 4: Project-related content report template

HOW

The goals of the "**Extraction of research-related content**" step are to: (i) **identify** and **retrieve** project-related documents (e.g. deliverables, reports, publications) that may contain useful content associated with the rural topic at hand and targeted project result types; (ii) **identify** content of interest in these documents; and (iii) **extract** content out of project documents.

The above stated objectives are achieved through the execution of a technology-supported process involving the following two tasks (as in the case of the research content extraction trial process; see **sub-section 2.4.1** above):

- **Task #1**: Search for project-related documents (e.g. deliverables, reports, publications) that may provide content of interest.

  Search for content sources is performed in an **automated way** through **use of keywords** indicative of the needed kinds of project documents (with regard to the sought types of results). The **criteria** for selecting project documents to be searched for relevant content, relate to:

  - the **number** of keywords found in a document;
  - the **frequency** at which keywords appear in the document;
  - **document sections** at which keywords have been found;
  - **whether** keywords found in a document appear to be close to each other or not; and
  - available **recommendations** about:
    - **types of project-related documents** that appear to be more useful for content extraction; and
    - **document sections** that appear to be "richer" in relevant content.

- **Task #2**: Search for and extraction of content out of identified (in the previous task) project-related documents.

  The task is executed by using: (i) appropriate keyword lists (keyword lists that have been made available together with topic lists in phase 1, or enhanced keyword lists produced after

recommendations outputted by the "**Research content extraction trial process**" step); and (ii) **text mining** methods/tools.

The outputs of these tasks are reviewed by the Scientific Editor and, if needed, more than one iterations are executed. Iterations take place till the identified project-related content sources and content extracted from them are considered to be satisfactory (termination condition of the iterations).

As far as **text mining methods/tools** are concerned, these can be distinguished into two broad categories: (i) **text pre-processing methods/tools**; and (ii) **content extraction/mining methods/tools**.

Text pre-processing methods and tools help to **clean text** and **prepare** it **for analysis.** Content extraction methods and tools are used for the purpose of **gaining insights** into the actual **information** and **content** available through the text. Detailed presentations and descriptions of text mining methods and tools is beyond the scope of this document, however brief descriptions of **indicative methods** in the above categories are provided in Tables 3 and 4.

Table 3: Indicative text pre-processing methods

Text pre-processing method	Method description
Word tokenisation	Word tokenisation is a process through which text sentences are split up in word tokens. This process outputs the list of words that text is made up from in order to use them in further text analysis tasks.
Sentence tokenisation	Sentence tokenisation is the task of splitting a text into its constituent sentences (Ma, 2018).
Stop words removal	Stop words identification and extraction is a core text cleaning task that aims to removal of frequently appearing words, found in text, which have no importance to further text analysis. A list of English stop words that is used by NLTK[7] is available at https://gist. github.com/sebleier/554280.
Stemming	Stemming is a heuristic process used to reduce words into their word stems, which are usually shorter versions of the original word token. For instance, application of a word stemming algorithm will reduce the words "fishing", "fished", and "fisher" to the word stem "fish" [8].
Lemmatisation	Lemmatisation is about "*resolving words to their dictionary form*". The method's name comes from the term "lemma", which is used to refer to a word's "*dictionary or canonical form*" (Heidenreich, 2018). For example, the lemma of the words "am", "are", "is" is the word "be" [9].

Table 4: Indicative content extraction/mining methods

Content extraction/mining method	Method description
Word frequency count	As implied by its name, the "word frequency count" method is used in order to calculate the number of times each different word appears in a text corpus. Such a task may help to identify words of importance in specific text contexts.
Keyword extraction	As implied by its name, "keyword extraction" refers to techniques used with the aim to automatically retrieve text-related keywords. Extraction of keywords, combined with use of text summarisation techniques, can help towards obtaining a quick overview of text.
Named entity recognition	Named-entity recognition (NER) is about classification of named entities (for example, person names, location names, organization names, product names), found in text, into

[7] As characteristically mentioned at https://www.nltk.org/, NLTK is a leading platform for building Python programs to work with human language data.
[8] The word stemming example has been retrieved from Wikipedia (https://en.wikipedia.org/wiki/Stemming# Examples).
[9] Lemmatisation example source: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html

	pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc.[10]
Text summarisation	"Text summarisation" refers to techniques for automated generation of text summaries. They can be used with the aim to quickly and easily get text overviews and, thus, find out what a text is about.

WHO

Execution of the tasks involved in the "**Extraction of research-related content**" step takes place through the use of appropriately developed automated tools. The development, maintenance and tool use are responsibilities of the Agricultural University of Athens (AUA). The outputs of tasks involved in this step are reviewed by the **Scientific Editor** who makes the final decision as to whether there is a need for further iterations of content search and extraction.

INPUT

The input needed for the execution of the "**Extraction of research-related content**" step comes from:

- the list of rural-related topics produced as output of the "**Identification of topics**" phase;
- the lists of topic-related sub-topics and keywords (provided as output of the "**Identification of topics**" phase and revised at the "**Search for relevant research projects**" step);
- the list of projects linked to each rural-related topic (made available as output of the "**Search-Retrieve-Pool**" phase);
- the list of proposed project result types (provided as output of the "**Search-Retrieve-Pool**" phase);
- the set of recommendations and guidelines provided by the "**Research content extraction trial process**" step.

OUTPUT

The output of the "**Extraction of research-related content**" step is an appropriately structured summary report with content made available for each topic-related project. The structure is imposed by the template illustrated in Figure 3. Another output of the present step are project mapping matrices (providing mapping of projects with regard to types of results) made available on demand.

CHALLENGES THAT NEED TO BE ADDRESSED

There is a **wide range** and **number** of **text mining methods** and **tools** which have the potential to be exploited. The inventory of tools used within the context of the "**Extraction of research-related content**" step can be dynamically updated based on needs that may arise on the fly.

2.4.3. Step 3.3: Synthesis of research results

WHAT

The step of the "**Synthesis of research results**" precedes the phase of "**Reporting research-related results**", in which all project results that are related to one topic are reported in the **SHERPA Discussion Paper** with the aim of **providing inputs** to MAP-based interactions. This step is concerned with creation of appropriate **topic-related research summaries** which are suitable for the development of the SHERPA Discussion Paper. These summaries contain information about: (i) the problems that have been addressed

[10] Source: https://en.wikipedia.org/wiki/Named-entity_recognition

in reviewed projects; (ii) targeted objectives; (iii) employed research methodologies; (iv) context of research and challenges; and (v) produced results and their types. Topic-related summaries are produced via synthesis of the above-mentioned information made available per project (output of the "**Extraction of research-related content**" step).

WHERE

The input that is required for the creation of topic-related summaries is drawn from the SHERPA repository, in which content extracted from project-related documents is stored. **Project summaries** and **topic-related summaries** are also stored in (and, thus, can be retrieved from) the SHERPA repository.

HOW

Topic-related summaries are automatically produced through use of an **appropriately developed tool** that draws content from each project summary and provides it as inputs to the topic-related summary. This tool also makes use of **Natural Language Processing** (**NLP**) which are techniques relevant to enhancing the presentation of aggregated project-related information. In addition, the topic-related summary produced as an output of the automated tool is further reviewed and edited by the **Scientific Editor** to produce a more natural flow of the content of the text.

WHO

The step of "**Synthesis of research results**" step the responsibility of:

- the Agricultural University of Athens (AUA), which develops and maintains the summary production tool; and
- the Scientific Editor who reviews and edits the summary from the automated extraction process.

INPUT

The input that is required for the execution of the "**Synthesis of research results**" step comes from the summary reports that are available for each topic-related project. Summary reports are produced on-demand per project by content extracted from project-related documents and stored in the SHERPA repository.

OUTPUT

The output of the step of "**Synthesis of research results**" is a report summarising topic-related information, retrieved from reviewed projects, covering: (i) problems that have been addressed in the projects reviewed; (ii) targeted objectives; (iii) research methodologies employed; (iv) context of the research and the challenges addressed; and (v) the results produced and their types.

CHALLENGES THAT NEED TO BE ADDRESSED

Topic-related summaries need to be clear and comprehensible in their meaning, and not mere aggregations of project-related summary information. Thus, further reviewing and editing of automatically generated topic-related summaries, by the Scientific Editor, guarantees the provision of information in an easy-to-read and comprehendible way.

2.5. Phase 4: Reporting research-related results

The final phase (phase 4) of the SHERPA methodology relates to development of the **SHERPA Discussion Paper**, a document that is used to provide input to the interactions that take place in Multi-Actor Platforms (MAPs). The "**Reporting research-related results**" phase comprises **two** (**2**) **steps**, namely:

- **Step 4.1**: Authoring the SHERPA Discussion Paper; and
- **Step 4.2**: Peer-review, revision and finalisation of the SHERPA Discussion Paper.

Tasks involved in each step of this methodology phase are presented similarly to steps of previous phases.

2.5.1. Step 4.1: Authoring the SHERPA Discussion Paper

WHAT

The first step of the phase on "**Reporting research-related results**" is the **process of writing** the SHERPA Discussion Paper. This is the step in which the Scientific Editor produces the document that provides an overview of the rural-related topic being considered, including the findings of relevant projects, and is used to facilitate interactions in the context of the Multi-Actor Platforms (MAPs).

WHERE

The development of the SHERPA Discussion Paper is based on an **appropriately designed template** that aims to capture all topic-related information and insights so as to facilitate MAP-based interactions and exchanges in the most effective way possible. The **SHERPA discussion paper template** is made available in the Annex of the present document.

HOW

Development of the SHERPA Discussion Paper is based on the following structure (SHERPA Discussion Paper template):

- Topic Overview

  Definitions and background information are made available regarding the rural-related topic being considered. The aim of providing this type of input is to help shape the context for the discussion intended to take place as part of MAP-based interactions.

- Guidance Questions

  Guidance questions serve as a follow-up to the initial orientation section. They are intended to steer the focus towards issues and findings that have been identified through the review of relevant research projects.

- Input from Research

  This section summarises what research has reported. It contains a summary of the findings of relevant projects. Some easy-to-understand and interpreted statistics are also provided. It also includes content extracted from project documents and reported in the previous phase (i.e. in the "**Extraction of research content and synthesis of outcomes**" phase).

- Results and Recommendations

  Results and recommendations made available by relevant research are described on the basis of a fit-for-purpose categorisation (for example, "insights/knowledge/evidence", "recommendations" and "data/tools"). This section includes content extracted from project documents and reported in the previous phase (i.e. in the "**Extraction of research content and synthesis of outcomes**" phase).

- Potential Topics to Discuss

  This section contains a list of issues that can be discussed as part of MAP-based interactions.

WHO

The **Scientific Editor** is in charge of the process of writing the SHERPA Discussion Paper.

INPUT

The inputs required for the step of authoring the "**SHERPA Discussion Paper**" are: (i) the **topic-related summary report** produced as the final output of phase 3 (the phase of "**Extraction of research content and synthesis of outputs**"); and (ii) the **SHERPA Discussion Paper template**.

OUTPUT

The output of the "**SHERPA Discussion Paper authoring**" step is a first version of the Discussion Paper for each rural-related topic developed on the basis of the template provided.

CHALLENGES THAT NEED TO BE ADDRESSED

The "**SHERPA Discussion Paper authoring**" step requires considerable human efforts to integrate the available input into a coherent document and messages (i.e. the SHERPA Discussion Paper), suitable for use in guiding and supporting MAP-based interactions. The Scientific Editor needs to compile input made available from topic-related summary reports and integrate it in the document, together with their own inputs. Submitting the paper produced into a review process, provides the quality control of the SHERPA Discussion Paper.

2.5.2. Step 4.2: Peer-review, revision and finalisation of the SHERPA Discussion Paper

WHAT

The second and final step of the "**Reporting research-related results**" phase is about the process of reviewing the SHERPA Discussion Paper prior to its finalisation. The tasks involved in this step of producing the SHERPA Discussion Paper are: (i) peer-review; (ii) revision; and (iii) finalisation.

WHERE

Reviews and revisions to the SHERPA Discussion Paper Document are implemented.

HOW

The initially developed Discussion Paper is assigned to Review Editor(s), who are SHERPA partner members, aiming to **review it** and **provide comments**. These comments relate mostly to the **content** of the report and assessment of the degree to which it can **actually facilitate discussion** and interactions in the context of the addressed topic. The peer-reviewed Discussion Paper is then assigned to the Communication Editor, who proceeds to comment and make editorial suggestions related to **provided information** and **layout issues**. All comments and feedback are made available to the Scientific Editor, who proceeds to the necessary document revision with the help of Support Staff.

WHO

Those involved in the production of the SHERPA Discussion Paper, for the process of reviewing, revising and finalisation are the:

- Scientific Editor;
- Review Editor;
- Communication Editor; and
- Support Staff.

INPUT

The inputs to the process implemented in the step for the "**Peer-review, revision and finalisation of SHERPA Discussion Paper**" are the initial version of the SHERPA Discussion Paper.

OUTPUT

The output of the review, revision and finalisation process is the final version of the SHERPA Discussion Paper, ready-to-be-used by regional MAPs.

CHALLENGES THAT NEED TO BE ADDRESSED

Formal peer-review processes require a considerable amount of time in order to be implemented and a high quality product finalised. More than one review and several iterations revision may be needed. Therefore, there is a need for **careful time scheduling** of the process given the inter-dependency of phases before and after the production of the Discussion Paper.

3. Networking with ongoing rural-related research projects

Apart from rural-related research projects that have already been implemented and can be proved to be useful for research undertaken in the context of SHERPA, there also needs to be a focus on projects that will be initiated and implemented during SHERPA's life cycle. Identification of such efforts in order to obtain more contemporary rural-related results is based on an organised plan involving frequent monitoring for relevant projects. This monitoring process is not limited to search in project programme databases and other web sources, but also includes direct contacts and investigation in partnership networks of **DG-AGRI**, **SHERPA partners** and organisations belonging to **SHERPA partner networks**. In case of identification of projects of interest, coordinators of these projects will be contacted directly by email or phone.

Monitoring for newly funded project efforts is going to take place on an **annual basis** to ensure that new project efforts will actually be identified and taken into consideration. Information related to these projects is going to be stored in the SHERPA repository exactly the same way as in the case of completed projects. Project-related documents for relevant content extraction will be either retrieved from project programme databases, and other web-based sources, or made available by project coordinators (in case that they are in a pre-publication phase). Identification of projects being in progress during SHERPA's life cycle will contribute to the creation of a **network around SHERPA** able to facilitate exchange of ideas and knowledge regarding rural-related topics. Networking will take place according to SHERPA's Communication Plan. Initial release of networking results will take place in Month 12 and records will be updated on an annual basis. Keeping track of evolvements with regard to projects addressing rural-related issues will allow for a consistent renewal of the content and information to be utilised for offering input to interactions at the Multi-Actor Platform level.

4. Conclusions

To achieve the goal of identifying and extracting relevant research content from past and ongoing research, the SHERPA project employs a four-phase methodology. This starts with identification and definition of rural-related topics, continues with a systematic process of project search and retrieval, followed by research content extraction, and concludes with the reporting of research results. Work Package 4 (i.e. "Stocktaking of relevant project results") provides sound methodological foundations for the tasks involved.

The description of all steps involved in the SHERPA methodology has focused on "**what**" (i.e. description of main tasks involved), "**where**" (i.e. references about sources of their input and/or outputs), "**how**" (i.e. explanation of the process of each step), and "**who**" (i.e. details about the roles involved). Details are provided of the **inputs** required for the execution of each step and the **outputs** produced. The challenges identified which require to be addressed in the context of each step denote points to be considered during execution of the methodology.

The process presented contributes to a comprehensive and systematic methodology. It presents the work required for the development of the SHERPA Discussion Paper, reporting all topic-related research findings

and results, which is required as inputs to MAP-based interactions. The Discussion Paper provides the interface between Work Packages 4 and clear.

The steps involved in all phases of the SHERPA methodology are executed through the use of human and technology-supported solutions. Appropriately designed and developed tools are utilised to facilitate and support tasks related to the search, extraction and storage of relevant information and content. SHERPA-related work and research, outlined in this document, will be supplemented by subsequent deliverables, which will focus on the details of the design of the SHERPA repository, the implementation of the text mining and provision of its results.

5. References

Ellis, D., Cox, D., & Hall, K. (1993). A comparison of the information seeking patterns of researchers in the physical and social sciences. *Journal of documentation*, *49* (4), pp. 356-369.

Ellis, D., & Haugan, M. (1997). Modelling the information seeking patterns of engineers and research scientists in an industrial environment. *Journal of documentation*, *53* (4), pp. 384-403.

Heidenreich, H. (2018). Stemming? Lemmatization? What? Retrieved from https://towardsdatascience.com/stemming-lemmatization-what-ba782b7c0bd8.

Hotho, A., Nürnberger, A., & Paaß, G. (2005). 'A brief survey of text mining'. *Ldv Forum*, *20* (1), pp. 19-62.

Kuhlthau, C. C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science*, *42* (5), pp. 361-371.

Ma, E. (2018). NLP Pipeline: Sentence Tokenization (Part 6). Retrieved from https://medium.com/@makcedward/nlp-pipeline-sentence-tokenization-part-6-86ed55b185e6

Mitchell, R. (2018). *Web scraping with Python: Collecting more data from the modern web*. " O'Reilly Media, Inc.".

Rosenfeld L, Morville P, Arango J (2015). Information architecture: for the web and beyond. O'Reilly Media, Inc.

6. Annex: SHERPA Discussion Paper template

Topic Overview

Definitions and background information with regard to the rural-related topic at hand. The aim of providing this kind of input is to help shape a context for the discussion intended to take place as part of MAP-based interactions.

Guidance Questions

Guidance questions serve as a follow-up to the initial orientation section. They intend to steer the focus towards issues and findings that have been identified through review of relevant research projects.

Input from Research

This section summarizes what research has to say. It is the section in which a summary of what has been made, in the context of relevant projects, is provided. Some easy-to-understand and interpret statistics are also being provided. This section includes content extracted from project documents and reported in the "**Extraction of research content and synthesis of outcomes**" phase.

Results and Recommendations

Results and recommendations made available by relevant research are described on the basis of a fit-for-purpose categorization (e.g. "insights/knowledge/evidence", "data/tools" and "recommendations"). This section includes content extracted from project documents and reported in the "**Extraction of research content and synthesis of outcomes**" phase).

Potential Topics to Discuss

This section contains a list of issues that can be discussed as part of MAP-based interactions.

Annex

Topic:

topic

number of research projects addressing the topic – related keywords

time period in which the topic has been mostly addressed

Distribution of topic across Project Funding Schemes

☐ FP6

☐ H2020

topic popularity

☐ topic 1

topic 2

☐ …

project partners per sector

☐ academia

☐ NGO

Results per research project on the topic

Project title:	
Start/end year:	
Involved actors:	
Addressed Problem/Subtopic:	
Objectives:	
Context and constraints:	
Methodology:	
Result:	
Result type:	

Project title:	
Start/end year:	
Involved actors:	
Addressed Problem/Subtopic:	
Objectives:	
Context and constraints:	
Methodology:	
Result:	
Result type:	

summary

Addressed Problem(s): Results: Recommendations:

SHERPA

Rural Science-Society-Policy Interfaces