RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

DOCUMENTATION OF RISIS DATASETS

European Social Innovation Database (ESID)

Abdullah Gök and Roseline Antai

v1.0 23/02/2021

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Version Information

Document Name	Documentation of RISIS datasets – European Social Innovation Database (ESID)
Version Number	Version 1.0
Document Date	13/03/2021
Main Author (s)	Abdullah Gok, Roseline Antai
Changes	This is the initial version of the document. Some parts were replicated from the ESID Manual in the KNOWMAK project.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table of Contents

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

1 Introduction

This document describes the data sources, structure, operationalisation and outputs of the European Social Innovation Database (ESID), a comprehensive and authoritative source of information on social innovation projects and actors in Europe. ESID was initially developed as part of the EU Funded KNOWMAK Project and currently it is being developed as part of the EU funded RISIS 2 project.

The development of ESID as part of the KNOWMAK project resulted in ESID v1.0, which were ultimately integrated into the KNOWMAK tool[1]. The current iteration of ESID as part of RISIS is called ESID v2.0. Further versions and the complete ESID data is available as part of the RISIS project.

ESID utilises advanced machine learning and natural language processing techniques to collect information about social innovation projects and actors from the publicly available information on the web. ESID also uses some limited human annotation to train our machine learning models and to ensure the quality and the integrity of the data.

The document is organised into the following sections: In Section 2, we provide some background information on social innovation. We outline and justify the criteria used to classify and identify social innovation. We also provide an overview of the database content and its design, along with an overall structural description of our database, the data stored in the database, and detailed statistics of our data. This section also discusses data integrity and the steps taken to achieve this, such as creating a web interface for manual annotation of our project data. Lastly, we provide a walkthrough of the web portal interface, as well as the verification we performed before accepting and incorporating annotated data to our database.

In Section 2, we cover the technical specifications, including how the data was collected, analysed and enriched. We discuss the database system used and provide an Entity Relationship Diagram. We describe the ESID engine in detail and give the technical details behind our data collection, classification as well as the information extraction processes performed. We provide a breakdown of our variable types and specification. We also discuss the results obtained in the various stages of our work. Additionally, we discuss the integration of ESID with KNOWMAK.

2 Database Content

2.1 Background

Social innovation is part of the solution to the various challenges that European societies face. From aging populations and the inclusion of marginalised groups to globalisation, it is necessary to build capabilities for societies and citizens to flourish. Many of these innovative solutions are social innovations, which have facilitated a growth of interest in the subject.

[1] In KNOWMAK tool, social innovation projects can be viewed on the right hand pane when clicking on a region: https://www.knowmak.eu/dashboard

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

From a policy perspective, Social innovation is becoming increasingly important in the European Union. Around 2010 social innovation was recognised by policymakers in Europe and the United States as an important driver of change. President Obama established The White House Office of Social Innovation and Civic Participation in 2009, while the European Union launched its Europe 2020 strategy in 2010, identifying Social Innovation as a field that should be nurtured in the Innovation Union Flagship Initiative. In 2013, the Social Investment package was launched to support EU member states in renewing their social protection systems, with a special focus placed on social innovation projects. The EU Commission subsequently adopted a more systematic approach, propagating and encouraging social and open innovation through policies on open innovation and open science as well as social protection policies. Social innovation became an opportunity for experimentation in multiple domains of government and industry through the Horizon 2020 and Collective Awareness Platforms for Social Innovation and Sustainability (CAPS) programme (Addarii and Lipparini, 2017).

While the term "social innovation" has been in circulation since the early 19th century, in its earliest incarnation, the term was politically charged and associated with social reform and revolution (Godin, 2012). By the mid-20th century, it was used to describe the remainder of "technical innovation" (Cajaiba-Santana, 2014). This meaning closely corresponds to the distinction between (Nelson and Nelson, 2002, Nelson and Sampat, 2001) "social technology" and "physical technology".

While the term social innovation resurfaced in Michael Young's writings in 1970s and 1980s (Mulgan et al., 2007), it still lacks conceptual clarity in spite of its rapid take-up since the early 2000s. A review of a number of studies that surveyed different meanings of the modern concept reveals four main elements that define social innovation (Caulier-Grice et al., 2012, Choi and Majumdar, 2015, Dawson and Daniel, 2010, Ettorre et al., 2014, Grimm et al., 2013, Harrisson, 2013, Jessop et al., 2013, van der Have and Rubalcaba, 2016, Edwards-Schachter and Wallace, 2017). While nuances between definitions vary, these broad criteria generally apply:

i. **Objectives:** Social innovations satisfy societal needs - including the needs of particular social groups (or aim at social value creation) - that are usually not met by conventional innovative activity (c.f. "economic innovation"), either as a goal or end-product. As a result, social innovation does not produce conventional innovation outputs such as patents and publications.
ii. **Actors and actor interactions:** Innovations that are created by actors who usually are not involved in "economic innovation," including informal actors, are also defined as social innovation. Some authors stress that innovations must involve predominantly new types of social interactions that achieve common goals and/or innovations that rely on trust rather than mutual-benefit relationships. Similarly, some authors consider innovations that involve different action and diffusion processes but ultimately result in social progress as social innovation.
iii. **Outputs/Outcomes:** Early definitions of social innovation strongly relate it with the production of social technologies (c.f. innovation employing only "physical technologies") or "intangible innovation." This is complemented by some definitions, which indicate that social innovation changes the attitudes, behaviours and perceptions of the actors involved. Some other definitions stress

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

the public good that social innovation creates. Social innovation is often associated with long-term institutional/cultural change.

Many definitions include a combination of the above four elements. For instance, the widely used EU definition (European Commission, 2013) (i.e. "social innovations are new ideas that meet social needs, create social relationships and form new collaborations.") involves elements i (objectives) and ii (actor interactions) as outlined above.

The majority of these definitions emphasise novelty and innovativeness as essential characteristics of social or other types of innovation, while there are others (Rogers, 2010) who relieve this criteria for social innovation. The novelty criteria are often seen as one of the key distinguishing factors between social innovation and social entrepreneurship (Cunha et al., 2015, Phillips et al., 2015). Similarly, most definitions share other essential characteristics of the classical OECD definition of ("technological product and process") innovation, namely involving a distinguishable practical activity (i.e. idea to be implemented) and resulting in new products, processes, services and models (OECD and EUROSTAT, 2005).

These four elements of the definition of social innovation are summarised in Table 1.

Table 1: Elements of Social Innovation

Element of Definition	Criteria
Objectives	• satisfy societal needs including the needs of particular social groups or aim at social value creation • target needs not met by conventional innovative activity either as a goal or end-product
Actors and actor interactions	• involve actors who would not normally get involved in "economic innovation", including formal and informal civil society/third sector/NGO/social and grass-root movements (i.e. social actors) • create collaborations between "social actors", business and public sector • involve (predominantly new types of) social interactions towards achieving common goals, including user/community participation • rely on trust relationships rather than solely mutual-benefit • involve significantly different action and diffusion processes that ultimately results in social progress as social innovation
Outputs/Outcomes	• change attitudes, behaviours and perceptions of the actors involved • produce social technologies (c.f. innovation employing only "physical technologies"), i.e. new combination or configuration of social practices or new law, norm or rule • lead to long-term institutional/cultural change
Innovativeness	• involves "the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Element of Definition	Criteria
	business practices, workplace organization or external relations"(Develop, 1997)

2.2 Existing Data on Social Innovation

There are around 10 existing databases of social innovation projects and initiatives (i.e. Digital Social Innovation, SI-Drive InnovAge, MOPACT, SIMRA, ICT-enabled social innovation, Centre de Recherche sur les Innovations Sociales (CRISES), etc.). Most of these databases are EU-funded. While these databases provide rich information on some social innovation projects, they have the following shortcomings:

- All of them are thematically focused on the technologies that the projects utilise (e.g. digital social innovation) or the societal goals of the projects (e.g. social innovation on ageing).
- They are fairly small, ranging between 50 to 1000 projects.
- They contain limited information about the features of the projects and actors.
- They rely on only one source. Some of them utilise manual search and input by project teams, while others collect their data through self-registration by the actors. All of the databases therefore involve a time-consuming human coding process. This also limits the sustainability of these databases as human coding only prevails during the data collection phase of the projects and after that, data starts to age rapidly.
- All existing databases adopt different definitions of social innovation. This makes comparison between different databases difficult. It is also observed that some databases have inconsistent operationalisations of their preferred social innovation definitions since some of the projects they include do not meet the definitions they adopt.

Additionally, there are a number of supplementary data sources (around 80 identified) which contain indirect but useful information (e.g. European Social Innovation Competition). These databases are also affected by the above issues.

All of the existing social innovation databases collect their information through manual data input by project team members or social innovation organisations themselves. On the one hand, data entry by an expert (i.e. member of project team) has the potential advantage of increased precision (i.e. data entered into database will be free from errors). However, as this is a very time-consuming manual process, this method of data collection has limited recall (i.e. some entities would not be included). On the other hand, data input by social innovation organisations results in very low precision (as some of the entities might be entered erroneously or maliciously) while recall is potentially increased.

As will be discussed later in this document, ESID employs an alternative approach to manual data input. It collects data through semi-automated machine learning. It was built on the above-mentioned publicly available databases, but it verifies, extends and enriches them. Consequently, ESID forms a definitive and comprehensive information source on social innovation with much higher precision and recall than existing databases.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

ESID offers the following advantages:

- It is thematically more comprehensive. It covers a broad range of societal grand challenges and key enabling technologies, thanks to its full integration with the ontologies developed in KNOWMAK Project WP2.
- It has significantly more recall than existing databases (containing data from all identified and available databases and additional projects identified from other sources, such as crowdfunding platforms). ESID contains several thousand projects as opposed to the existing databases, which range from50 to 1000 projects.
- It has a more consistent and flexible conceptual structure in terms of the social innovation definition it adopts. As discussed above, rather than one definition, we identify projects based on a comprehensive and diverse set of criteria.
- It provides much richer information on the projects and actors.
- As it is based on semi-automatic and automatic information retrieval and knowledge discovery techniques, it is more sustainable than existing databases that rely on continued human coding. ESID is updated with minimal human supervision. Due to machine learning, the more data it includes, the more precise its data collection is, which requires less human supervision.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.3 Inclusion criteria

Our entry point to data collection is projects. The projects in ESID are social innovation projects and activities found in the identified data sources which also satisfy our inclusion criteria. ESID first identifies projects from known sources, enriches them and then identifies actors related to them.

Based on the literature review presented above, we identified that social innovation definitions usually incorporate four elements: objectives, actors and actor interactions, outputs and outcomes, along with the implicit requirement of innovativeness. Table 2 presents a set of operational rules that was developed for each criterion.

Table 2: Social Innovation Operational Definition Criteria

Element of Definition	**Criteria**
1. Objectives	Project primarily or exclusively satisfies (often unmet) societal needs, including the needs of particular social groups; or aims at social value creation. Often no price is involved for the beneficiary or the innovation is provided to the beneficiary at low cost without any profit motive. However, there are examples where a fee is involved.
2. Actors and Actor Interactions	• Satisfy one or both of the following: **i.** **Diversity of Actors:** Project involves actors who would not normally be involved in innovation as an economic activity, including formal (e.g. NGOs, public sector organisations etc.) and informal organisations (e.g. grassroots movements, citizen groups, etc.). This involvement might range from full partnership (i.e. project is conducted jointly), to consultation (i.e. there is representation from different actors). **ii.** **Social Actor Interactions:** Project creates collaborations between "social actors" (i.e. actors that are not conventional innovation creators, but engage in social innovation such as charities, social enterprises, public sector organisations), small and large businesses and public sector in different combinations. These collaborations usually involve (predominantly new types of) social interactions towards achieving common goals such as user/community participation. Often, projects aim at significantly different action and diffusion processes that will result in social progress. Often social innovation projects rely on trust relationships rather than solely mutual-benefit.
3. Outputs and Outcomes	Project primarily or exclusively creates socially oriented outputs/outcomes. Often these outputs/outcomes go beyond those created by conventional innovative activity (e.g. products, services, new technologies, patents, and publications), while conventional outputs/outcomes might also be present. These outputs/outcomes are often intangible and they might include the following but not limited to: • change in the attitudes, behaviours and perceptions of the actors involved and/or beneficiaries • social technologies (i.e. new configurations of social practices, including new routines, ways of doing things, laws, rules or norms) • long-term institutional/cultural change

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Element of Definition	Criteria
4. Innovativeness	The Project should include the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method. The project needs to include some form of innovative activities (i.e. scientific, technological, organisational, financial, and commercial steps intending to lead to the implementation of the innovation in question). Innovation can be technological (involving the use of or creating technologies) as well as non-technological. The innovation should be at least "new" to the beneficiaries it targets (it does not have to be new to the world).

The ESID engine relies on manual annotation for about 20% of the entities so that the machine learning model can automatically predict each of for the subsequent projects. For this purpose, we have prepared annotation guidelines to be used in the human annotation process that is described in Section 3.4.3. The criteria and guidelines were refined iteratively through a pilot of around 200 projects by several different coders.

Box 1: Annotation Guidelines

- Annotation:
  - Annotators annotate each of the four criteria in sentence or paragraph level (i.e. some criteria might span multiple sentences).
  - Some sentences might indicate multiple criteria.
  - Claims versus evidence: We will assess what the project claims and will not seek any evidence of achievement. Some of the things might not happen yet (still being planned), but it is enough for us if they are mentioned as plans.
  - Guidelines for Actors and Actor Interactions
    - If one of the criteria is fully satisfied please grade as 3.
    - If both of the criteria are partially satisfied (normally grade 2) please grade as 3 (since both of them are satisfied).
  - Guidelines for outputs/outcomes criteria:
    - It might be useful to assess outputs and outcomes by looking at objectives of the project (the project might not have any outputs or outcomes yet, but it might have plans)
  - Guidelines for innovativeness criteria:
    - If the project claims they are conducting an innovation by using the term, but they do not substantiate the nature of innovations described above, give mark 2.
    - The project does not have to use the term "innovation" explicitly. If it satisfies the above criteria but does not use the term at all, it still might be marked as 3.
- Overall marking:
  - At the very end of the document (i.e. project), annotators will give grades for the project based on the eight criteria. Grading:
    - 3: fully satisfies the meaning of the criteria
    - 2: partially satisfies
    - 1: very weakly satisfies
    - 0: no indication at all (no sentence level annotation should be inputted for this criteria if you mark 0)
  - Give the benefit of the doubt to the project when marking.
  - If a project is clearly spam and has no relevance to social innovation, please mark as "spam project".

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.4 Subsets of ESID

For each project, the ESID database contains scores ranging from 0 to 3 for the four social innovation criteria. This allows us to avoid adhering to a strict definition of social innovation, as the concept is prone to varying interpretations. In turn, the users of the database can filter the projects in ESID based on their exact definition of social innovation using the four criteria.

ESID contains two subsets:

- Curated projects subset: these are the projects we reported to the KNOWMAK tool. These projects adhere to a strict set of criteria such as:
    - All projects satisfy the EU definition of social innovation: at least partially satisfies the objectives criteria AND at least partially satisfies actors and actor interactions criteria AND at least partially satisfies innovativeness.
    - All projects are located in Europe.
    - All projects are related to at least one topic of key enabling technologies or societal grand challenges.
    - All projects have full information on a number of key variables such as project title, project URL, project location, project topic, project summary.
    - All projects in this subset were manually verified and if necessary corrected to ensure the data quality.
- Non-curated projects subset:
    - This subset includes all the other projects in the database.
    - These projects are located inside and outside of Europe.
    - Some of these projects do not have some of the information (e.g. they are not on topics we cover).
    - Some of these projects were manually annotated to train our machine learning models but some of them were not.
    - Some of the projects included in this subset are “negative” examples (i.e. projects that do not classify as social innovation), as our models require positive as well as negative examples.

2.5 Overall Structure of ESID

The ESID database includes two main entities stored in a database:

- Social innovation projects (reported in KNOWMAK)
- Actors involved in social innovation (these are partially implemented at present)

Various features about main entities are collected, such as their name, type, web page, social media profiles, etc. However, since there are multiple values for some attributes, it is necessary to allow this in the model with multiple tables related to the main entities. The features and variables with their connections to the main entities are presented in Figure 1. It is worth mentioning that some actors were discovered from initial databases, and as such have no linked projects even though they are in our database.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

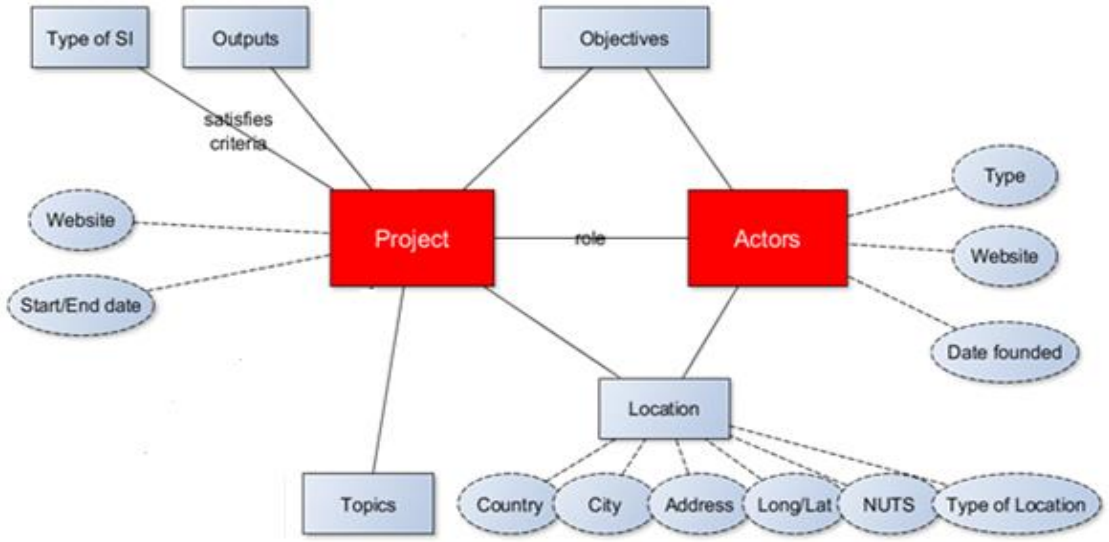


Figure 1: Project and actors variables (presented on a high level)

We also included relationships between entities in our model. Relationships can be between actors (e.g. subsidiary, umbrella organisation), between project and actor (e.g. actor executing or funding a project), or between projects (e.g. pilot, follow-up).

For every piece of information stored in the database, we also store the source from which that information was obtained. By doing so, we keep track of how the database is populated. Also, we assure the quality of data that is presented to the end user, as we will be able to track the information back to its source.

Furthermore, we also store all crawled pages and information. Crawled pages are stored in a document store, in MongoDB database. These pages are linked to the projects or actors for which they were crawled, by keeping the project or actor id in the MongoDB document with the information whether it is related to project or actor.

The information in our data model is populated in phases. In the first phase, we populated essential information for the main KNOWMAK database, such as project basic information, location, topics (SGCs and KETs). In the second phase (as part of the RISIS 2 Project), we will endeavour to populate more challenging information, such as information on actors, organisational structure data, funding information, outputs, outreach and impacts of the projects.

The ESID Schema shown in Figure 2 shows all the variables we have used in Phase 1 of our projects, as well as the intended ones for Phase 2. We should point out that only a section of the Actors variables has been utilized. The bulk of the work involving the Actors variables will be done in the second phase of the project.

Key variables in ESID 1.0 are

1. Project title – the name of the project
2. Project website – the main website of the project

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3. Project Facebook/Twitter/LinkedIn – Social media URLs of the project related pages
4. Project locations – the locations of the project that include the city, country, longitude and latitude. From these pieces of information, it is possible to infer FUA, NUTS1, NUTS2 and NUTS3 codes with RISIS geocoding tool.
5. Project description/summary – a brief description. Firstly, we utilised the descriptions given by the source databases. Then, at a later stage we went on to derive summarization of the project description from the project website as described in Section 3.6.2.
6. Topics – projects are assigned two classes of Topics: key enabling technologies societal grand challenges based on an ontology developed in the KNOWMAK project[2].
7. Social innovation inclusion criteria – The classification was made for each inclusion criteria (objectives, actors, outputs, and innovativeness). The classification was a combined effort of automatic scoring, as well as based on human annotations of the sample of the projects. The classification criteria follow a four-level gradation – 0 to 3, where '3' denotes that a criterion is most highly satisfied. As we explain further, these scores were condensed into a binary form where all scores which were greater than 1 were condensed to '1' – meaning criteria satisfied, and all '0' scores were left as were, showing that the criteria were not satisfied. At the moment we have a gradual classification (doesn't satisfy/partially satisfy/satisfy), with this condensation to a binary level, if required.

As ESID is based on unstructured information from the web, some features were not available for all observations (i.e. there was a degree of missing data). We have marked the mandatory variables (i.e. available for all observations) in the proceeding variable tables. The initial focus of the project was on retrieving information about projects.

[2] Detailed information on this ontology can be found at https://gate.ac.uk/projects/knowmak/

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

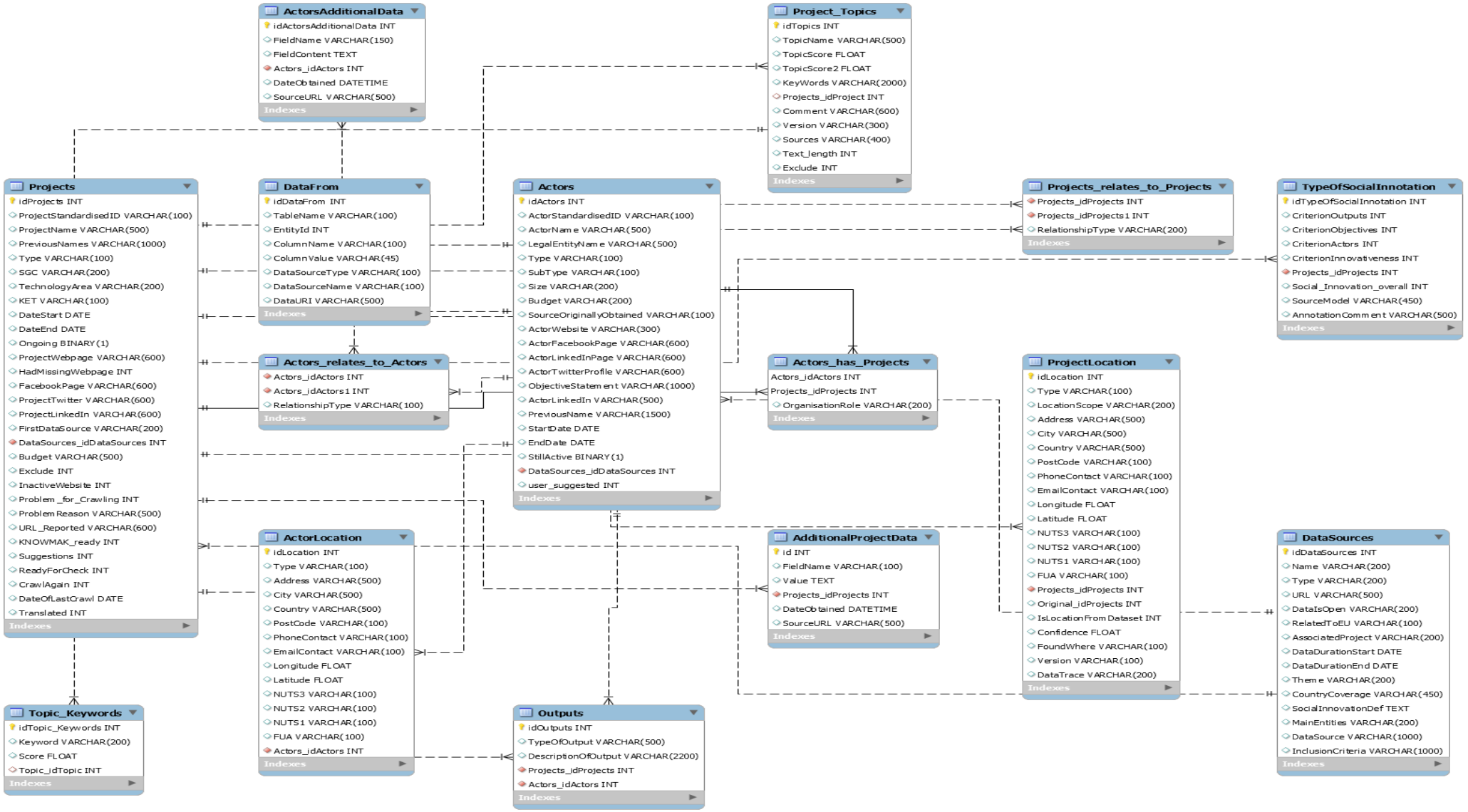


Figure 2: ESID Schema

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.6 Information on variables

2.6.1 ESID Tables

The ESID database has a number of tables and variables, and the relationships between these are shown in the Schema diagram in Figure 2. We provide here in Table 3 below, an overview of the different tables in the ESID database, as well what phase of the projects they are/will utilized in – 1 or 2. A detailed view of each table, with all the variables in the table is given in Appendix II of this document.

Table 3: ESID database Tables description

ESID Table	Table Description	Most Relevant Variables	Phase
Projects	The Projects table provides a detailed view of each project in the database. It holds information about each project, and also links to the other tables in the database, such as the Project Location, Actors, Project Topics, Type of Social Innovation and DataSource. It has sixteen (16) variables.	-idProjects (PK) Project Name -Project Type - ProjectWebpage -Date of Last Crawl	1 & 2
ProjectLocation	The Project Location table holds all the details on the location where the project was carried out. It holds details such as the project city and country details, as well as where the project location was found in the webpage of the project. It contains links to the Projects table and the actors table. It has sixteen (16) variables.	-Project_idLocation (PK) - City -Country -Longitude -Latitude	1 & 2
AdditionalProjectData	This table holds additional project data that did not fit in the structure of the Projects table variables. It contains information like the social media account	- AdditionalProjectData_id (PK) - Value - FieldName	1

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

	details of projects, as well the summarised description of these projects. It is linked to the Projects table, and has 6 variables.		
TypeOfSocialInnovation	This table holds data on how much each project satisfies our four (4) criteria of Social Innovation. It holds the output of the machine learning classifiers predictions scores for each projects level of satisfaction of each criterion. It is linked to the Projects table and has 6 variables	- idTypeOfSocialInnotation(PK) - CriterionOutput -CriterionObjectives -CriterionActors -CriterionInnovativeness -SourceModel	1
Projects_relates_to_Projects	This table holds the information on how one project might be related to another, for example, if one is a sub-project of another, or in the case of chain projects, a child project. It has 3 variables and links to Projects table	- Projects_idProjects(PK) -RelationshipType	1 & 2
Actors	The Actors table holds information on the different actors involved in a particular Social Innovation project. The information held includes details such as the type of actor, as well the website and social media accounts of these actors. This table has 19 variables and links to the Actor Location, Data Sources and Actor_has_projects tables. The table has 19 variables	-idActors -ActorName -ActorWebsite -SubType	1 & 2
ActorLocation	The ActorLocation table holds information on the	- Actors_idActors(PK) - City -Country	1 & 2

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

	location of the actors involved in the Social Innovation projects. It holds information such as the type of location, the city and country details of the actors. It has 15 variables and is linked to the Actors table		
ActorsAdditionalData	This table holds additional data on project actors that did not fit in with the variables in the Actors table. It holds information such as additional websites of the actors, as well as tags are held in this table. It has 6 variables and links to the Actors table only.	- idActorsAdditionalData (PK) -Actors_idActors	1
Actors_has_Projects	This table relates Actors with the projects they are involved in. It also specifies the actor's role in the project, such as the main partner role, or other. It has 3 variables, and links to both the Projects table and the Actors table	-Actors_idActors (PK) -Projects_idProjects -Organisation Role	1 & 2
DataSources	The DataSources table holds additional information about data sources from which information was collected. It holds information such as the database name, the URL, and if the data source is open or not. The table contains 14 variables and is linked to the Actors and Projects tables	-IdDataSources (PK) -Name - Type -URL - DataIsOpen -AssociatedProject -CountryProject -Theme -MainEntities -DataSource	1

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.6.2 Descriptive Statistics

In this section, we present a descriptive statistical analysis of the distribution of our variables through the database.

2.6.2.1 Projects

ESID contains in total 9,577 projects, of which 2,688 projects form the curated subset while 6,889 the non-curated subset. The curated dataset contains high quality data with integrity, which had been human annotated and corrected repeatedly to ensure its accuracy. As we display the analysis of our data below, we display the numbers that fall in the three categories: curated, non-curated, and full dataset.

2.6.2.1.1 Social Innovation Scores

As described above, ESID contains a score for each of the four criteria of social innovation. Scores range between 0 (no indication at all) and 3 (fully satisfies the meaning of the criteria). As discussed in Section 3, our machine learning models have differential performance of classifying projects into these four criteria. We show these scores and the number of projects that fall in each criterion in Table 4 below.

Table 4 :Social Innovation Scores Projects count

Criterion	Score	Total number of projects in ESID with SI Scores	Number of Projects in curated subset	Number of Projects in non-curated subset
Outputs	0	5,414	293	5,121
	1	1,913	1,434	479
	2	497	492	5
	3	475	469	6
Objectives	0	5,318	224	5,094
	1	1,851	1,345	506
	2	435	431	4
	3	695	688	7
Actors	0	5,598	388	5,210
	1	2,007	1,616	391
	2	287	283	4
	3	407	401	6
Innovativeness	0	5,377	256	5,121
	1	1,829	1,349	480
	2	556	552	4
	3	537	531	6
Total Number of Projects		8,299	2,688	5,611

For technical reasons outlined in Section 3, we also collapsed the four level scoring (0 to 3) into a two level scoring (0 or 1) in our curated subset (Table 5).

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 5 : Condensed Social Innovation Criterion Project count

Criterion		Number of Projects in curated subset	Total Number of Projects
Outputs	0	293	2,688
	1	2,395	
Objectives	0	224	2,688
	1	2,464	
Actors	0	388	2,688
	1	2,300	
Innovativeness	0	256	2,688
	1	2,432	

Table 6, Table 7 and Table 8 also present statistical distribution of projects in ESID based on the combinations of the four criteria of social innovation.

Table 6: Social Innovation gradation scores by criteria combination (Full Dataset)

Project Scores	Number of All Projects			
	All four Criteria	At least one criterion	At least two criteria	At least three criteria
= 0	5,262	-	-	-
>=1	2,594	3,037	2,969	2,644
>=2	597	1,169	1,132	662
= 3	375	787	526	410

Table 7: Social Innovation gradation scores by criteria combination (Curated Dataset)

Project Scores	Number of Curated Projects			
	All four Criteria	At least one criterion	At least two criteria	At least three criteria
= 0	200	-	-	-
>=1	2,225	2,488	2,458	2,268
>=2	589	1,157	1,120	654
= 3	371	779	519	379

Table 8: Social Innovation gradation scores by criteria combination (Non-Curated Dataset)

Project Scores	Number of non-Curated Projects			
	All four Criteria	At least one criterion	At least two criteria	At least three criteria
= 0	5,062	-	-	-
>=1	369	549	511	376
>=2	8	12	12	8
= 3	4	8	7	5

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

2.6.2.1.2 Projects with websites

Some projects in the website have their own independent websites, while some have both independent websites and social media sites. We show the distribution of projects in both of our curated and full datasets that have both or one of these (Table 9).

Table 9: Projects with Independent and Social media websites

Projects	All Projects	Curated Dataset Projects
Independent websites	9,294	2,686
Facebook page	1,854	1,219
Projects Twitter	1,664	1,184
Project LinkedIn	784	575
Independent website and Facebook page	1,698	1,219
Independent website and Project Twitter	1,663	1,184
No Independent website	283	2
No independent website or Social Media Account	127	2

2.6.2.2 Actors

Here, we present a Table of Actors, showing the number of actors in our database, and the number associated with projects. We also show their organisational role – main partner and other, as well as how many actors fall into these roles (Table 10).

We present a more comprehensive table of all our actors organised according to their subtypes in **Appendix IV. ESID Actors by Subtypes**.

Table 10: Actors by projects count

Projects	Number of Actors	Number of Actors Linked to projects	Main partner	Other partner	No organisational role specified
All Projects	6,666	3,912	841	2,210	835

2.6.2.2.1 Actors with websites

In the Table 11 below, we show the actors that have an independent website, as well as those that have just Social Media sites. We also show the number of actors who have no social media account or independent websites.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 11: Actors with Independent and Social media websites

Actors	All Actors	Actors with projects
Independent website	2,422	141
Facebook page	101	0
Actor Twitter profile	0	0
Actor LinkedIn page	0	0
Independent website and Facebook page	0	0
Independent website and Actor Twitter	0	0
No Independent website	4,244	3,258
No Independent website or Social Media Account	4,143	3,258

2.6.2.3 Location

Projects in our database have corresponding locations in the form of cities and countries. We also store coordinates derived from this data. Coordinates can then be used to infer a host of other geographical information including NUTS classifications. The projects in our database are located in 1,057 unique cities and 167 unique countries (Table 12).

Table 12: Projects by Location count

Projects	Total number of projects with locations (City or Country)	Total number of projects with locations (City AND Country)	Number of projects with no cities	Number of projects with no countries
All Projects	4,467	4,154	347	40
Projects in Curated Dataset	2,684	2,668	16	0
Projects in Non-curated datasets	883	632	285	40

Projects in ESID are geographically very diverse. Most of the projects are located in Europe (about 70%), while there are about 7% of our projects in North America and about 2% in Asia. The highest number of projects are located in the UK (412 projects, in our curated dataset, and 612 in our full dataset). Do the same description for the curated subset.

Some projects are located in more than one country, for instance some projects are located across EU while some other are located across a number of African countries. In

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

these cases, ESID records all the locations separately while we present them in this report as an entry called "Multiple Countries" (Table 13, Figure 3 and Figure 4).

Table 13: ESID projects by Country count (Curated and Non-Curated Dataset)

Country	Number of Projects in Curated	Number of Projects in Full Dataset
UK	414	614
Germany	272	369
Italy	182	217
USA	150	925
Netherlands	147	176
Spain	141	205
Belgium	128	142
Austria	94	122
Sweden	84	111
France	81	120
Denmark	50	69
Poland	48	56
Romania	41	50
Switzerland	41	58
Bulgaria	40	47
Ireland	38	47
Portugal	36	44
Turkey	35	48
Finland	35	41
India	32	71
Australia	27	48
Canada	27	48
Hungary	24	31
Russia	24	44
South Africa	24	32
Egypt	22	29
Greece	20	26
Latvia	20	23
Czech Republic	18	20
Kenya	18	18
Colombia	17	43
Lithuania	17	25
Croatia	15	24
Slovenia	14	19
Serbia	13	17
Brazil	12	18
Bosnia and Herzegovina	11	11
Estonia	11	12

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Albania	10	18
North Macedonia	10	10
Argentina	8	14
Chile	8	11
New Zealand	8	12
Uganda	8	13
Nigeria	7	8
Singapore	7	10
Japan	6	10
Kosovo	6	13
Mexico	6	9
Norway	6	9
Mali	5	7
Malta	5	5
Zambia	5	6
Oman	4	10
Slovakia	4	10
Tanzania	4	7
Tunisia	4	5
China	3	33
Cyprus	3	3
Ghana	3	3
Israel	3	3
Lebanon	3	4
Luxembourg	3	4
Malaysia	3	3
Mauritius	3	3
Montenegro	3	11
Nepal	3	4
Ukraine	3	3
Zimbabwe	3	3
Armenia	2	3
Bangladesh	2	3
Belarus	2	2
Cambodia	2	2
Ecuador	2	5
Georgia	2	3
Guatemala	2	2
Iceland	2	5
Morocco	2	2
Niger	2	2
Senegal	2	3
Togo	2	2
Afghanistan	1	1
Algeria	1	1
Azerbaijan	1	1

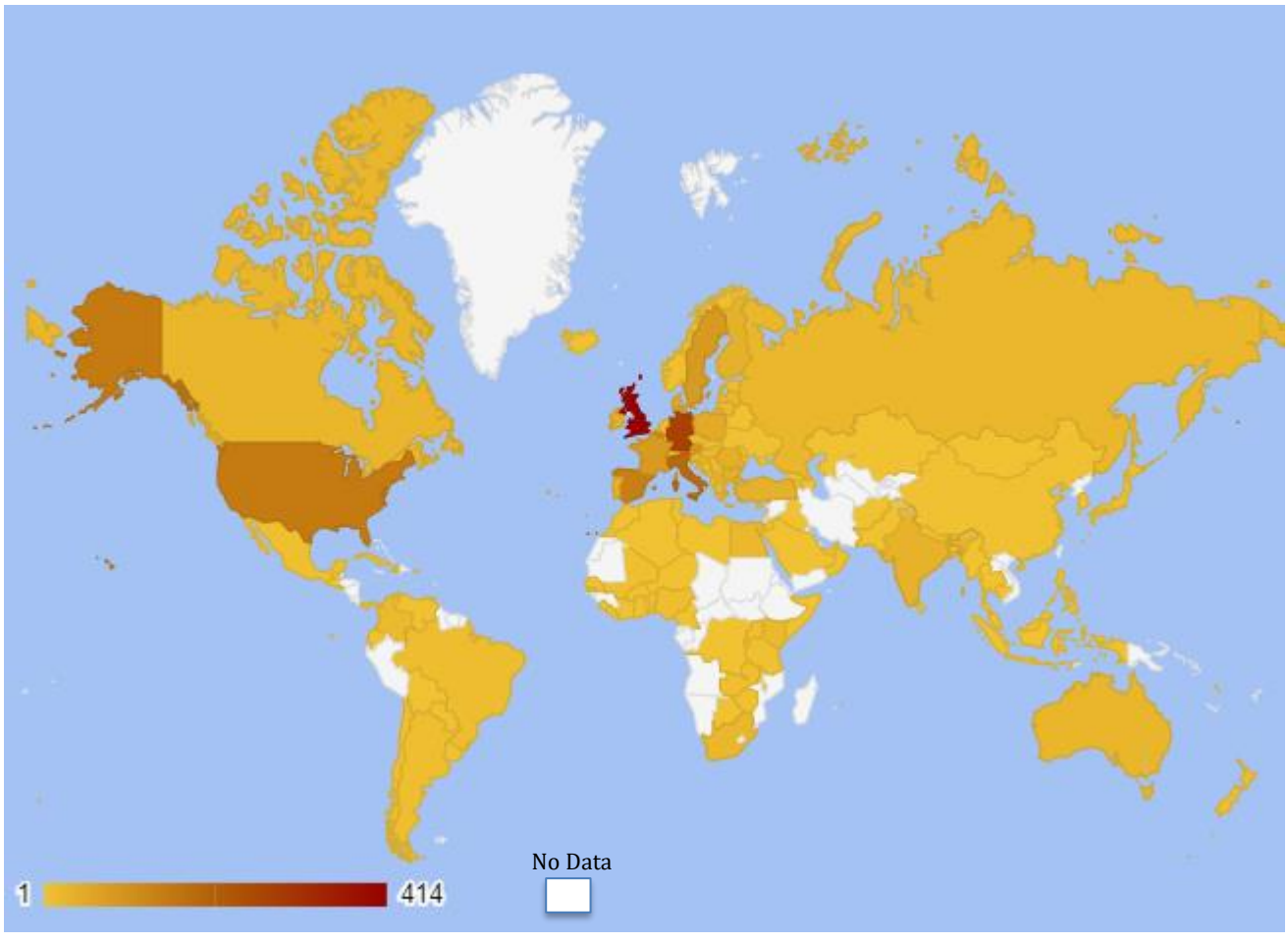
This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Barbados	1	1
Benin	1	1
Bhutan	1	1
Bolivia	1	3
Botswana	1	1
Burkina Faso	1	2
Burundi	1	1
Cameroon	1	1
Cocos(Keeling) Islands	1	10
Columbia	1	1
Cote d'ivoire	1	1
Cuba	1	1
Dominica	1	1
DR Congo	1	1
El Salvador	1	1
Gambia	1	1
Grenada	1	1
Haiti	1	2
Hong Kong	1	1
Indonesia	1	1
Iraq	1	2
Jordan	1	4
Kazakhstan	1	1
Liberia	1	1
Libya	1	1
London	1	1
Malawi	1	1
Moldova	1	1
Mongolia	1	1
Myanmar	1	1
Pakistan	1	7
Palestinian Territory	1	1
Panama	1	1
Paraguay	1	2
Philippines	1	1
Puerto Rico	1	1
Rwanda	1	1
Saint Lucia	1	1
Saint Vincent and the Grenadines	1	1
Saudi Arabia	1	2
Sierra Leone	1	1
Somalia	1	1
South Korea	1	2
Sri Lanka	1	2
State of Palestine	1	1

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Thailand	1	2
The Philippines	1	
Tobago	1	1
Trinidad	1	1
United Arab Emirates	1	3
Uruguay	1	1
Vanuatu	1	1
Venezuela	1	1
Virgin Islands	1	1
Peru		1
Kyrgyzstan		1
British Indian Ocean Territory		6
Costa Rica		1
Ascension Island		1
Iran		2
Isle of Man		1
Korea		1
Multiple Locations	42	70

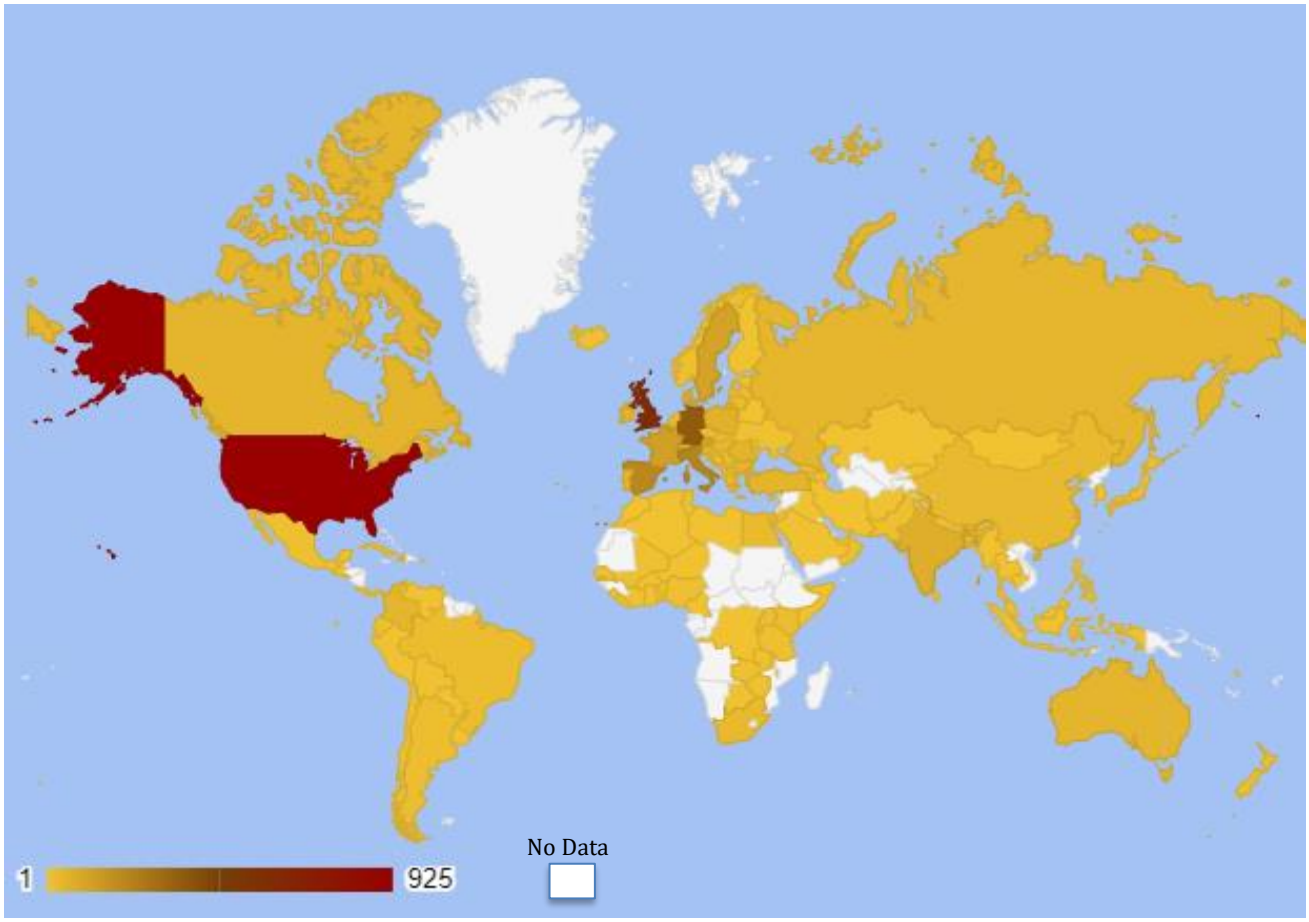
Figure 3: Heatmap showing Projects count in countries (Curated Dataset)



This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 4: Heatmap showing Projects count in countries (Full Dataset)



2.6.2.4 Project Descriptions

We also show the distribution of projects with descriptions in our dataset, and the average length of these descriptions, given that these were summaries created from our summarization process (Table 14).

Table 14: ESID Projects with Descriptions (Curated, Non-Curated and Full Dataset)

Number of Projects	Projects with Descriptions	Average Length of Descriptions
All Projects	9,416	608.6
Curated Dataset Projects	2,687	228.3
Non-Curated Dataset projects	6,729	811.7

2.6.2.5 Project Topics

Table 15 and Table 16 show the count of projects in our dataset, per topic. We show how many of our projects fall under Key Enabling Technologies (KETs) and which ones fall under Societal Grand Challenges (SGCs).

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 15: ESID Projects with Topics count (Curated, Non-curated Dataset and Full Dataset)

Number of Projects with Topics	8,472
Number of Projects with Topics in curated dataset	2,684
Number of Projects with Topics in non-curated dataset	5,788

Table 16: ESID Projects Topics counts by KET and SGC (Curated and Full Dataset)

	Topics	Frequency counts	
		Full Dataset	Curated Dataset
Key Enabling Technologies (KET)	Industrial Biotechnology	767	442
	Nanoscience and Technology	425	237
	Optics and Photonics	47	32
	Micro-and Nano-electronics	1	1
	All Projects with a KET	1,240	712
Societal Grand Challenges (SGC)	Society	3,479	2,141
	Bioeconomy	23	15
	Transport	249	126
	Health	147	97
	Climate change and the Environment	268	171
	Energy	33	24
	All Projects with SGC	4,199	2,574

2.7 Quality and accuracy of data

As ESID relies on the initial data sources, especially in ESID1.0, there were some missing projects. Similarly, while the ESID Engine tried to balance between precision and recall as much as possible, there was inevitably, a recall loss due to our classification scheme. In order to address these and increase our recall and precision, we employed the following manual mechanisms:

- Extensive manual checks: We manually checked all of the projects with an aim to increase the precision of ESID in terms of the projects' alignment with our definition of social innovation. We also checked the information provided on the projects, for instance, if a website was taken over by another organisation, or if the summaries reflects the projects, etc.
- Guided search on the basis of an initial analysis of the database: We analysed the database in terms of the major groupings of project themes (for instance, poverty reduction, refugee integration, environmental protection). For currently underrepresented groupings we investigated new sources or we manually identified projects through web search.
- Quality control by stakeholder: We held a virtual workshop with stakeholders including our partners at the KNOWMAK project where each stakeholder investigated ESID1.0's curated subset data in detail. Stakeholders were asked to

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

review projects in the database and to make suggestions on the addition of missing projects, or exclude irrelevant projects. This was then reviewed by the team and adjusted in ESID 1.0 curated subset.

- Hackathon: We invited scholars and PhD students working on social innovation to a Hackathon day. They were then encouraged find errors and omissions.

For the quality check, we developed a web interface through which these reports were made. All the submissions to this interface were then subsequently reviewed manually. We extensively discuss these final checks in Section 3.4.3 of this document, including a detailed overview of the web interface portal we designed to enable us complete this task. As discussed above the resulting fully manually checked data constituted the curated subset.

2.7.1 Manual reviews of the extracted data

In order to facilitate the quality of the information presented in ESID, we have adopted manual reviews as a final step, before marking the data as ready for presentation. Data points for the projects were reviewed manually by the domain experts as soon as they were deemed accurate and had integrity. This way, we were able to assure the best quality of the presented information.

To facilitate manual checks, we have developed a web interface, in which we present a list of projects to the domain experts, who reviewed the information collected for each project, edited the incorrect information (while all the edits are logged) and marked the project as ready for presentation in the ESID database. The web interface facilitates registration of the users, with approval from the administrators, so non-invited users cannot view or edit projects. In order to review projects, users need to log-in. User then can search projects that they want to review, or they can access curated list by administrators of the projects that need to be reviewed in the current review phase.

The interface was developed using python and Flask framework for web applications. The various screens on the user interface is presented in the following figures (5 – 9). Users can - login, search, list projects to review, view projects, and edit projects.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

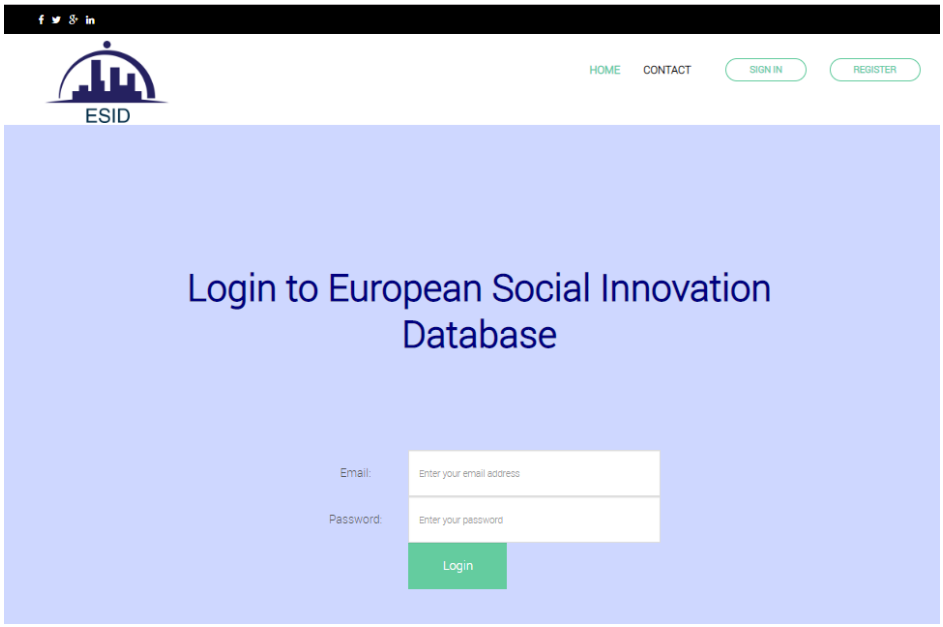


Figure 5: Log-in screen of web interface

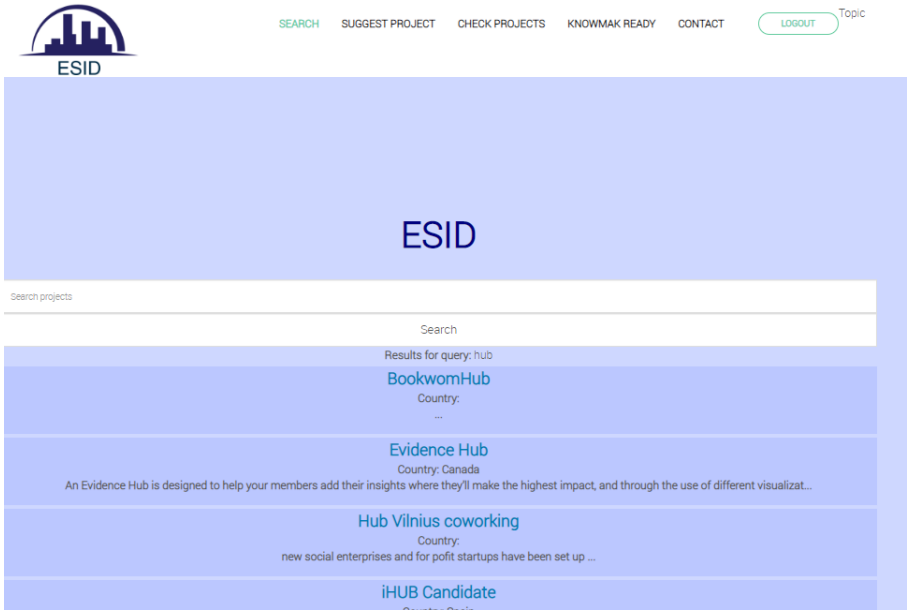


Figure 6: Search view of web interface

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

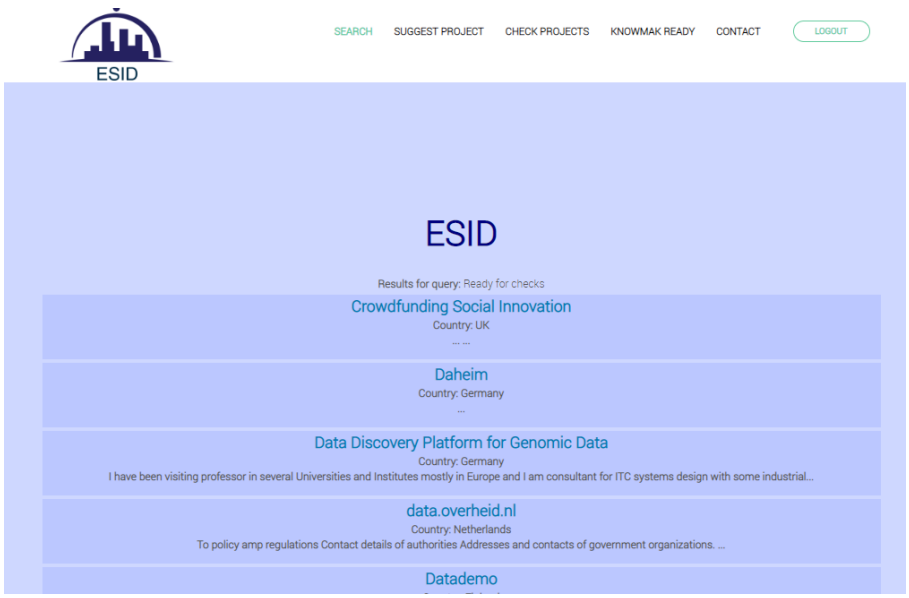


Figure 7: List of projects that need to be checked in the current phase

Figure 8: Project review screen

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 9 : Project edit screen

3 Technical Specifications

3.1 ESID Architecture and methodology

ESID contains multiple elements:

1. ESID engine, which is responsible for crawling the web, classifying the data, processing and extracting the information;
2. The Structured database in MySQL, which is utilized for storing the essential, structured and processed information which is presented in the interface
3. The Unstructured database, which is utilized for storing crawled data and translations of the documents that were not in English. The unstructured database is MongoDB, a NoSQL database program.
4. Connector/exporter to KNOWMAK, which was responsible for exporting and sending

database updates to the main KNOWMAK database.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 10 presents the architecture of ESID components and the relationships between data stores in KNOWMAK and ESID.

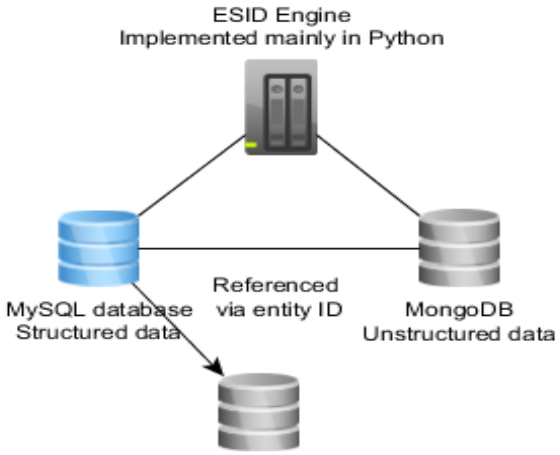


Figure 10: Relationships between data stores in KNOWMAK and ESID

3.2 Phases of ESID

Data collection and processing in ESID is complex and is therefore split into two basic phases:

- **Phase 1:** Entities were identified from the existing databases. This provided us with an entry point into social innovation. As discussed above, the information contained in the existing databases was uneven and, in most cases, incomplete. However, most existing databases included additional sources of information on entities, for example their website or Facebook page. The existing databases were enriched by text mining these additional data sources for the entities identified.

  Even though they were manually coded, the quality of the data collected from the existing data was not optimal because of varied definitions, conventions and approaches which the existing databases have adopted. To overcome this quality issue and to guide the machine learning process, Phase 1 involved substantial human annotation effort. As we are now ensured that our seed data is of high quality and trained our machine learning model, we will now rely on less on manual coding in the RISIS 2 Project.

  At the end of Phase 1, we have the first version of ESID (ESID 1.0). This seed database is based on entities identified in the existing databases, but includes much more information with substantially higher degree of internal and external consistency.

  A list of the existing databases and data sources which were used in Phase 1 is presented in Appendix I. List of Existing Databases and Data Sources

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

- **Phase 2 (to be implemented in RISIS 2):** In this phase, more entities will be discovered, as well as information on these entities from fully unstructured data sources. Based on the seed database and the machine learning model we created in Phase 1, we will text mine information contained in a number of data-sources. Some of these data sources include chain project sites like FabLab, Kickstarter, and other crowdfunding databases, databases of social enterprises, websites of the universities in ETER, major actors of social innovation etc. As a result, we will be able to identify more entities and the required information on them. This will help us extend the seed database substantially. We will also be able to keep the data live as opposed to most of the existing databases which operate within a limited time window. Keeping the data live will involve implementing some form of dynamic crawling, which will allow us to keep our crawled data up to date with any changes made to these project websites.

  In conjunction with the continuous manual data quality assurance, the additional entities discovered in Phase 2 will facilitate the training of the text-mining model, and keep it up to date. This will enable our text-mining model to self-learn continuously with minimal supervision.
  Phase 2 will also include two data quality control mechanisms. Firstly, it will involve some limited human coding to fine-tune the text-mining model. This effort will be considerably smaller, in comparison to the coding programme in Phase 1. However, low-level continuous human coding is necessary to assure quality. Secondly, a mechanism for entity consultation will be implemented. For the pre-production and production database we will include an interface in which organisations and/or people connected with the social innovation projects and actors will have the opportunity to suggest amendments to the data on them, held in ESID.

The summary of the phases that we are looking to implement in this Phase 2 include:

- Extension: This will involve increasing the number of projects we have to about 20,000 projects. This will be achieved by sourcing for additional projects from:
  - Chain projects: Social Innovations projects which are connected to each other, either through having different branches, or through the same actor and initiative. As such, we will be looking at including these projects as related projects in our database
  - Additional sources: In addition to incorporating chain projects in our database, will also be looking at including projects from other Social Innovation sources, including social enterprise websites, crowdfunding sites, to mention a few
  - Kickstarter: Kickstarter is a crowdfunding site which contains projects which are not specifically Social Innovation, but there are some Social Innovation projects as well. These will be incorporated in our database.
- Expansion: The expansion phase will involve improving our existing models; the Social Innovation classification models, the summarisation models as well as the location models, through the use of manually annotated data. We will

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

rerun our existing algorithms on the dataset we have, which is an improved training set due to it being human annotated. We will also work on model improvement and enhancing the efficiency of the models. This will involve making tweaks, and optimisation of the models. Another aspect of this phase will involve implementation of other techniques and state of the art NLP methods, as well as improved crawling techniques. We will be looking into utilizing more expansible crawlers, which will not only lead to an improvement in the efficiency of the crawling process in terms of time, but also in terms of the quality of the crawled data. The addition of more features will also form part of this phase of the project. We hope the addition of more informative features will improve our model quality and performance, hence, helping us answer some data analysis questions in relation to social innovation. Some of the features we will be looking to include are:

  - Actors (including type of actors, project owner, funder, partner, etc)
  - Mission, products/services, outputs/impact

- Dynamic retrieval: The aim of this is to ensure that we keep the data on these projects in our database up to date. As such, what we will be looking to do is perform biannual crawls. This means we will have to implement our crawling methods more effectively. This will make up the first stage of our dynamic retrieval implementation. As another part of it, other techniques will be explored, to enable us implement a timely semi-automatic crawl using our algorithms. Revisits to project websites to ensure that the data we have in our database is up to date, and that webpages are still active, are other goals of the dynamic retrieval phase.

A summary of different data-sources and Phases of the ESID is presented in Figure 11.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

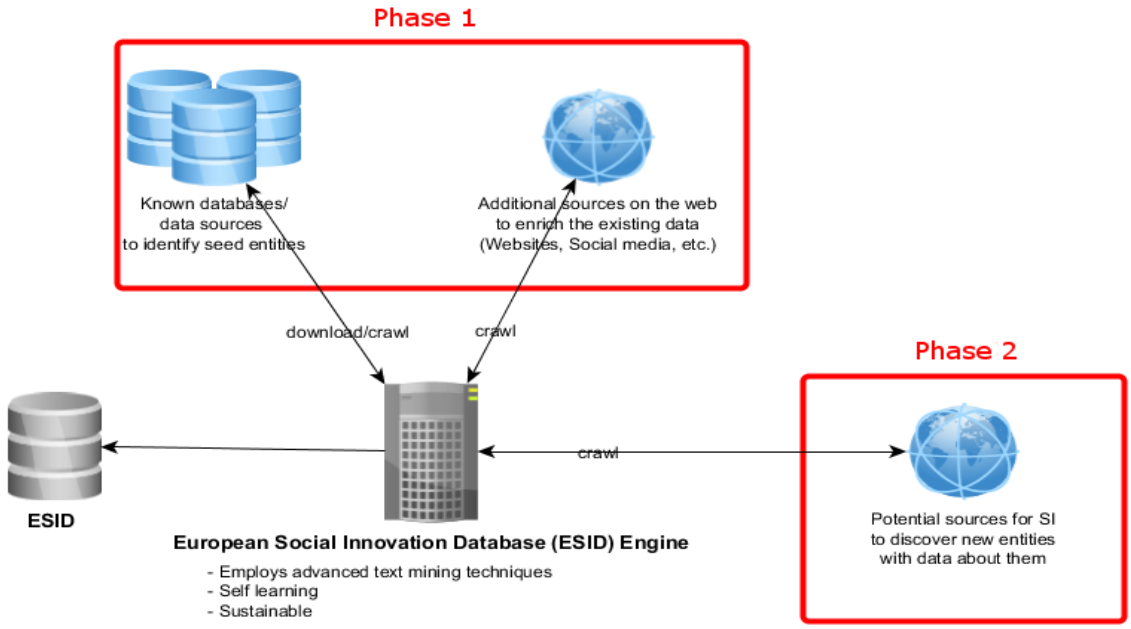


Figure 11: ESID General Architecture

3.3 The ESID Engine

The ESID engine is a web crawling, natural language processing and machine learning system for discovering social innovation projects and creating, editing and updating the ESID database. The ESID engine contains the following components:

- A set of web crawlers
- A set of natural language pre-processing tools and named entity recognisers which are used for location detection and classification
- A set of machine learning classifiers which are used to perform the classification into social innovation or not
- A combination of natural language and machine learning classifiers for performing the summarization task
- Web interfaces for adding and editing data stored in the ESID database; to provide annotated text
- A set of tools for data normalization

This conglomeration of the abovementioned complex system and tools we named the ESID engine. It presents a system that is able to obtain, process, and enrich information from the seed information sources (as shown in Figure 18).

In the following subsections, we outline the plan we followed for the development of the ESID engine and ESID database.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.4 Data collection

The seed social information data sources are manually identified. These data sources are then scraped in order to obtain initial data about social innovation actors and project. The ESID engine scrapes known data sources and stores the available information about social innovation actors/projects into the database. Most of the data sources cannot be downloaded, but can be viewed over the web interface. We developed a set of crawlers and data transformers in Python, mainly using Scrapy library, to extract this information and store them in the database. The data sources that we downloaded data from, are databases or project/organisation lists that are often developed during other EU projects.

In this stage, we collected data that can be mined in the further steps. This was mainly done through the use of web crawlers. However, other tools for collecting data for different kinds of documents, such as PDF were also used. In this part, textual data was manually collected, however, when crawling data sources, some structured information was also be retrieved and immediately stored in the database.

Information stored in the databases were not consistent, and in some cases, not complete as well (some databases link organisations to projects, while other did not, some databases contained full addresses and coordinates, while some contained just the country, etc.). Since the information was incomplete and not normalised, it was necessary to scrape additional information sources to find the missing information and to normalise available information to the same format. Most of the data sources contained a web address, a Facebook page or other pointers to additional sources. All the websites that were found in data sources were crawled. We also searched for the names of the projects or organisations on DBPedia, search engines (Bing and Google API), Twitter, Facebook and LinkedIn and extract available information. Additional information, in the form of text, is stored in MongoDB database and linked with the project or organisation id in our MySQL database.

3.4.1 Phase 1: Data Sources Identification

97 potential sources for social innovation were identified. Generally, sources fell into the following categories:

- Curated social innovation databases – These include the databases that were created either as a result of certain EU or other government funded project that collect social innovation in a certain area. Social innovation databases could also be created by a certain NGO organisation who is interested in the area. These databases could have different sizes, ranging from 10-2,000 entities. They presented the primary source for obtaining initial information in Phase 1 about social innovation projects. However, some of these databases contain a high number of false positives (in some cases it could be up to 60%). The number of false positives depends on the mode of data entry. The largest number of databases with false positives were user inputted, while those with much lower false positive rates had expert curated and inputted databases.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

- Social innovation prizes – This is usually given by governmental bodies or EU agencies, who publish on their websites a list of semi-finalists or finalists. There are usually 10-30 projects listed each year for different prizes (European Social Innovation Prize, European Investment Bank's Social Innovation Tournament). These prizes present also one of the main initial information sources and usually these competitions contain low levels of false positive projects listed (usually less than 10%).
- Case study reports – produced by different academic, NGO or governmental organisations, case study reports usually provide a relatively small number of curated projects in a certain area. Information in these reports is usually unstructured.
- Funding organisation databases – organisations that provide grants for projects often publish a database of the projects they funded. However, these databases are not as reliable as previous sources, as only some of these projects may be social innovation projects
- Social enterprise databases – databases listing social enterprises. Companies listed in these databases may be involved in social innovation; however, while many companies provide solutions for social issues, they are not necessarily innovative.

As previously mentioned, 97 sources were identified, which fall under these categories. Some of these sources contain several thousands of projects and actors, while others contain only a few (some contain only 3 entities). The main challenges regarding the data sources were:

- Data sources were inconsistent
- Different structures of data source websites required a crawler for each database
- Some data sources were not publicly available
- Information in some data sources overlapped
- The data sources contained a different wealth of information

We further explain these challenges which we encountered:

The data sources were not consistent. Some of the data sources contained only actors, some contained only projects, while some contained both projects and actors. Out of the databases that contain both projects and actors, some contained linkages between actors and organisations, while others did not. This presented a challenge, as the final database was required to be consistent and contain information about relationships between projects and actors. Therefore, some of the seed data sources did not contain enough information and had to be enriched using natural language processing techniques, which may have made them less reliable. Also, discovering relationships between projects and actors is quite a challenging task. Only few databases contain information about relationships.

Crawlers for each data source. Each data source contained some basic information about the project. However, this information was differently positioned on the page and often, different data sources may contain different information (e.g. some may contain coordinates, some addresses, some may have no location at all, etc.). Some data sources are not structured, but present their projects in the form of case studies. This posed a

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

challenge, as information had to be retrieved using text mining and natural language processing techniques. Because of the different structures of pages and databases, for each data source, a custom crawler had to be developed.

Unavailable data sources. Some of the identified databases from the literature were not publicly available. Though a valuable source of information may be identified based on the literature, we were unable to access some of these even upon contacting authors.

Overlap between databases. On the other hand, some projects were included in multiple databases. As we previously mentioned, the data sources were generated by different entities. The inclusion criteria vary; the requirement was to have no data sources with complete overlaps, however, some projects could be found in multiple databases (e.g. included in thematic databases, received certain prices, therefore included in competition database, certain organisations wrote a case study on them). Duplicate projects were excluded from the database.

Different information wealth. As it was already mentioned, data sources have different information stored in them. Some data sources contained quite detailed information about the project, including name, location, description, enabling technologies, etc. On the other hand, other data sources contained only project names and maybe one or two additional attributes. The challenge in this sense was normalizing data, so each project had similar information presented about them.

3.4.2 Phase 1: Crawling

A web crawler (also known as scraper or spider) is a program or automated script which browses the World Wide Web in a methodical, automated manner and collects the content of the visited web pages (in full or targeted parts of them). During the data collection phase of the project a number of crawlers had to be developed in order to obtain data from identified data sources. The data sources could not be directly downloaded, however, they had databases accessible on the web, and therefore web crawlers could methodically visit all entity pages and obtain data about them.

However, while crawling, we faced a number of challenges. Some of the challenges have already been discussed as challenges regarding data sources. However, the main challenge we encountered during the crawling phase was that the web sources did not have consistent structures, and wealth of information. Our approach to crawl these data sources was to obtain targeted information that was included in the data source about a certain entity. In order to achieve this, we needed to develop separate crawlers for each data source that were able to locate information of interest on the page.

The development of crawlers can be time-consuming. We started with developing crawlers for the biggest databases. Certain data sources had information in unstructured PDF documents and therefore crawlers were not helpful in these cases. These sources were fairly small, containing case studies, and therefore were entered manually. We have implemented a web interface for adding data into the database. We discussed this in Section 2.7.1 of this documentation.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

We have collected data from over 45 data sources, containing over 3,500 projects and over 6,000 organisations. This presents more than 95% of all entities in the identified data sources.

The crawled data sources and the number of their entities are presented in the following table.

Table 17: ESID Datasource with Projects and Actors count

Data source	Number of projects	Number of actors
BENSI	29	0
Berlin Startup list	3	0
CAPPSI projects	37	0
Citizen science projects	9	0
Customer related projects	10	0
Digital Social Innovation	1,017 (2,203 before cleaning)	2,007
EFESEIIS case studies	52	0
EMPATIA case studies	23	0
EMPATIA pilots	6	0
European Social Innovation Competition	90	0
Global Innovation Fund projects	32	0
MoPAct	140	0
ICT for social innovations	21	0
ICT-enabled social innovation initiatives	39	0
ICT-enabled social innovation: cases	114	0
Innovage	153	0
Impact Hub Stockholm	33	0
ImPRovE cases	7	0
Kennisland	5	0
LIPSE cases	16	0
MAKE IT case studies	10	0
Making Sense campaigns	6	0
Manual search	18	0
Marias World Foundation	6	0
MAZI Pilot Studies	6	0
SIMRA	9	28
European Investment bank social innovation tournament	72	0
Open Knowledge International	14	0
P2P Value case studies	4	0
P2P Value directory	382	0
SI-drive case studies	35	0
SI-drive	1,005	2,624
SINGOCOM Cases	17	0
Social Innovation Generation	6	256
Social Investment case studies	5	0
TRANSIT social innovation initiatives	73	0
TRANSIT social innovation networks	22	0
TRANSITION cases	17	0

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

TRANSITION projects: social innovation warehouse	16	0
TRANSITION success cases	9	0
WEBCOSI: civil society initiatives	6	0
WILCO cases	42	0
Bill and Melinda Gates Foundation	0	444
Social Enterprise UK	0	687
Total	3,493	6,046

The presented list is a list of all projects and organisations found in the discussed data sources. Some of these projects may not be social innovations and therefore they will be removed in the future. Also, not all projects are captured by the given data sources. ESID will discover additional projects on other, general web sources, such as crowdsourcing platforms.

Apart from the crawlers that obtain targeted information from database sources, we also developed a general website crawler. This crawler is used to obtain additional information about projects and actors from their websites. The crawler takes URLs of the projects from the database, visits them, obtains all the textual content from the pages and stores it into the Mongo database. The link between the project in the MySQL database and the page in Mongo database is established through a project id in the MySQL database, which is stored as one of the document's attribute in the MongoDB. The pages and description texts from the databases will be used in order to determine whether the project is social innovation project or not.

The development of a crawler for websites was very challenging. It was necessary to determine the depth of crawling. In certain cases, the crawler was crawling many pages without relevant content. In order to overcome this issue, we limited the crawling time for a single website to 10 minutes. Should a crawler require more than 10 minutes to get data, it probably means it went too deep and is obtaining irrelevant pages. Also, some pages would have redirects, for which the domain limitation would not work or be suitable. Some websites may contain descriptions of many different projects, out of which only one or few are social innovations. In order to deal with these kinds of websites, for the projects having a certain page in database as a website, the crawler crawled only that page. In the case where the whole domain name is referenced as the project's website, the whole website would be crawled, with crawling depth of 3. Also, certain projects or actors had big web portals, with blogs and news, while the others have relatively small websites. Certain projects used Wix or other platforms for creating their websites. Crawling content from some of these platforms can be challenging, because of the content loading mechanism that those platforms use in hiding the content from the page source. For these sorts of sites, we had to retrieve them via JavaScript. Crawling of the websites for all our projects and actors can be also very time-consuming, as it may take several days.

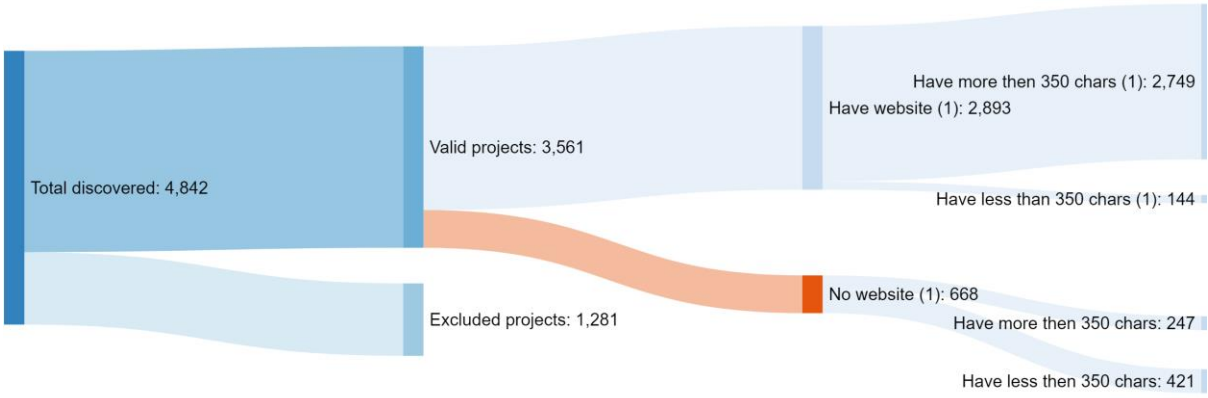
Preliminary results of the first stage of data collection are presented in Figure 12. We discovered 4,842 projects from source databases. We excluded 1,281 projects because they were subsequently excluded from these source databases, for instance Digital Social Innovation Database dropped some of the projects in their database after data cleaning. These dropped projects are also useful in terms of providing "negative examples" to our

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

machine learning algorithm. A small number of projects (about 80) were excluded because they were duplicates of already existing projects.

Of the remaining 3,561 projects, about 565 projects did not have a running website. We excluded another 169 projects whose websites returned 404 error (page not found). Another 12 projects were excluded because their websites had a placeholder page selling their domain. 787 projects had less than 350 characters. In total 2,593 projects had more than 350 characters in the text describing them (website crawled text and description text from data source). These projects could be classified by ESID engine.

Figure 12: ESID Preliminary Results of the First stage Data Collection



3.4.3 Phase 1: Human Annotation Workshops

Once the pages about projects were crawled, we aimed to classify the projects based on whether they satisfy our social innovation criteria or not. This can be performed using supervised machine learning. However, it was necessary to guide (supervise) the algorithm in order for it to learn what projects satisfy social innovation criteria. In order to achieve a sufficient level of supervision we organised two human annotation workshops to annotate around 10% of the crawled projects.

The data for the annotation workshop was obtained by crawling websites which were pointed at by the social innovation projects in the crawled data sources. A project in the data source usually contained some website. These websites were crawled and merged into a single file for annotation. We included only projects that contained between 500-10,000 words into the first annotation task data set and projects from EU Social Innovation Challenge and European Investment Bank's Social Innovation Tournament that contain between 500-20,000 words into the data set for the second annotation task. Annotators were instructed to use only the presented text in drawing conclusions about projects and whether they satisfy the given criteria.

However, these can be challenging both in terms of dataset generation and for annotators. Since the majority of the projects came from social innovation databases, it is assumed that they would satisfy a major part of the criteria. However, according to the descriptions, this was not always the case. Some project focused on innovation, while others focused on social objectives, omitting information about innovativeness. Using

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

background information to infer certain inclusion criteria would be unfair towards the machine learning classifier, as it cannot access that background information, while not marking these things may reduce the recall of the method significantly. Also, often one website could contain descriptions about multiple smaller events or projects, under the umbrella of the main project. These may introduce a certain level of noise for the classification and information extraction.

The first annotations workshop was organised on 16th September 2017. During this event, 6 annotators annotated about 40 projects each. During the workshop, we first presented the annotation schema to the annotators. The annotation schema was developed in a number of iterations and it was assured backward compatibility. In order to develop a common understanding between annotators, two projects were annotated as an example and discussed with the rest of the annotators. Afterwards, annotators annotated documents on their own. 15-20% of documents were shared between two annotators. These documents were used for calculating inter annotator agreement. The annotators had to annotate sentences in the text that explain why the project satisfies a certain criterion. At the end of the document, annotators gave scores for each of the criteria. These scores indicate how well the project satisfies given criteria. Marks used were in range 0-3. The annotations were performed using *Brat annotation tool*[3].

The second workshop was organised in the week between 30th October and 3rd November 2017. Three annotators annotated 43 documents, while one annotator annotated 30 documents. All four annotators participated in the first workshop, therefore, presenting the annotation schema again in order to develop a common understanding, was not necessary. The annotations were performed using the *Brat annotation tool*. All annotators were PhD students and researchers at the Alliance Manchester Business School, The University of Manchester. All the documents were shared between the annotators, so we could perform more exhaustive examination and statistics of inter-annotators' agreement. During the workshops, we annotated about 10% of the data available to us at the moment of the annotation workshops. In cases where two annotators annotated the same document and disagreed, we asked for additional annotation to be performed by the experts at the Austrian Centre for Social Innovations (ZSI). Aggregate conclusions from both annotation workshops are:

- Agreement for detecting social innovation or false positive projects was about 85% (spam project was defined as a project which an annotator marked with zeros for all criteria or hit the spam button)
- Agreements for annotating each inclusion criteria were the following:

[3] http://brat.nlplab.org/

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 18: ESID Annotation agreement count

Inclusion criteria/level of annotations	Paragraph level annotations	Document level annotations
Objectives	37.5%	69.3%
Actors and Actor interactions	17.2%	63.8%
Outputs	18.9%	65.7%
Innovativeness	19.5%	66.8%

Agreement was calculated as overlapping annotations over the total number of annotations done by both annotators. During the annotation task, we did not identify any outlier annotator. Annotators performed similarly on both paragraph and document level annotations.

We have performed an additional document annotation task in collaboration with ZSI. In this annotation task, two independent annotators from Austrian Zentrum für Soziale Innovation (ZSI) were tasked to annotate in total 451 projects. Each annotator, similar to the previous task, gave marks for a defined criterion, as well as annotated sentences which describe a given criteria. In case marks between two annotators differed by more than 2 (scale of marks was 0-3), a third independent annotator would give a final decision. At the end of this annotation task, we obtained 728 annotated documents.

Marking of the documents was performed relatively well. We decided to create a machine learning model using document level marks/annotations, as there was satisfactory agreement between annotators. However, since annotators were not agreeing on marks, we decided to drop marks, and transform scores to a binary metric. In case an annotator marked a certain criterion as 1 or higher, we assumed it was satisfying the criteria. On the paragraph level, the inter-annotator agreement was relatively low. This is due to the fact that phrases or sentences explaining objectives, actors, outputs and innovativeness are not exact, especially in social projects. Therefore, annotators may mark different sentences explaining the same criteria, while reaching the same conclusion. However, sentence level annotations were used to assure annotators were reading the documents, and as a measure of quality control.

The dataset generated from the annotation task is used for machine learning classifiers for social innovation criteria. Inter-annotator agreement will be used as a performance benchmark.

During the process of human annotation, we noted the amount of false positive entries in the crawled databases. For each data source, we mapped projects that were annotated as a sample from it, and the number of false positive projects annotated by our annotators. The percentage of false positives is presented in the following table:

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 19: Datasource project counts with false positives

Database	Number of projects	Percentage of false positive projects
European Social Innovation Competition	90	12.9% (8/62)
MoPAct	140	7.3% (3/41)
Innovage	153	30% (6/20)
Digital Social Innovation (as of November 2017)	2,200	58% (105/188)
European Investment bank social innovation tournament	72	6.8% (6/87)
SIMRA	9	0% (0/2)

Note that this is representation of the datasets from September/October 2017. In November/December Digital Social Innovation performed significant cleaning and almost halved the number of projects, thereby reducing the number of false positive entries.
The number of false positive projects is calculated as number of files marked by annotators as false positive. Multiple annotators were annotating the same file (at least 2, in some cases up to 4), and it is calculated as average (total number of annotations as false positives/total number of files from the given data source).

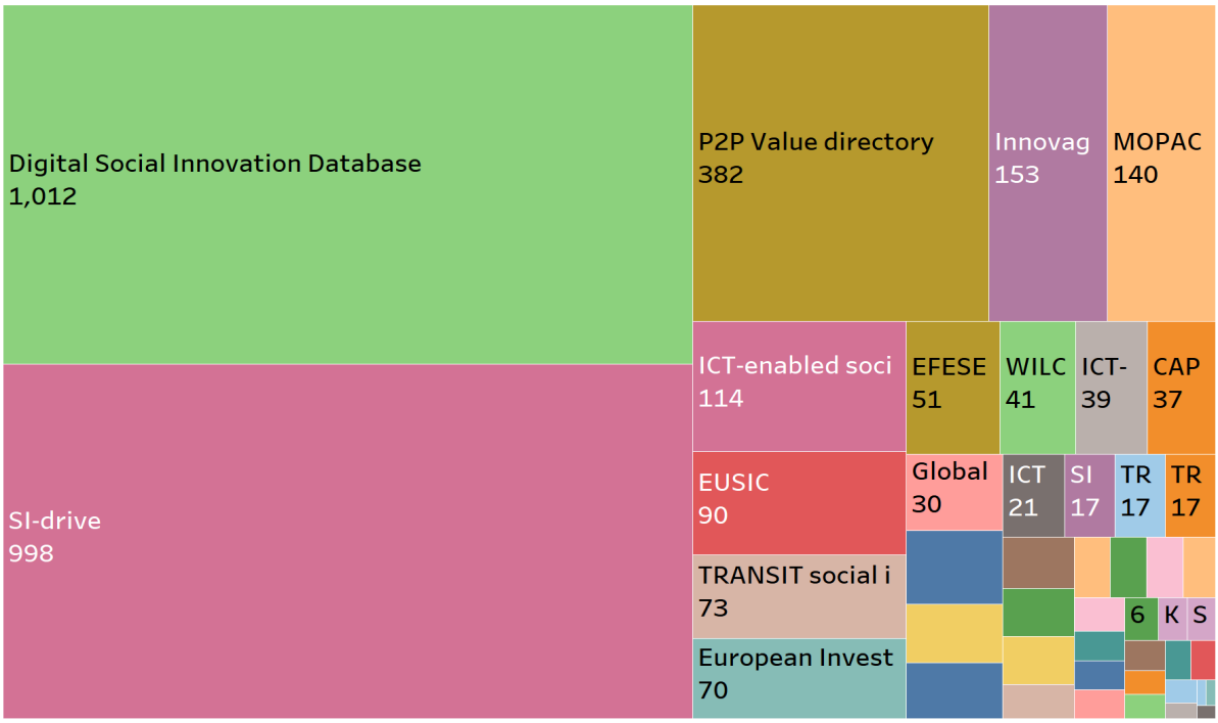


Figure 13: Proportion of the project depending on data source

Every annotation in the database was checked and confirmed (or disputed) by at least one additional annotator. Disputed annotations were resolved by the third annotator. We obtained 986 annotated projects from three annotation workshops out of a total of 3560

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

projects initially included in the ESID. Out of 986 annotations, 881 included annotations that were based on an overall understanding.

As previously stated, the projects that were currently annotated consisted mainly of social innovation projects. For example, 820 projects were marked as 1 or more in overall social innovation criteria. The table with distribution of scores can be seen in Table 20. In order to allow supervised machine learning that needs relatively balanced datasets, we have crawled European projects listed on Kickstarter[4]. We have collected 4297 projects from Kickstarter.

These projects are usually not social innovation projects, and therefore we have automatically labelled them as such in our database in order to create a set of negative examples.

Table 20: The distribution of scores per each annotated criteria in the original set of annotated projects

Criteria	Total projects	Mark > 0	Mark >1	Mark > 2
Social innovation	881	820	790	741
Objectives	986	934	894	776
Actors	986	890	800	626
Outputs	986	909	856	697
Innovativeness	986	924	876	711

For each classifier we selected an equal number of positive and negative sample projects from our database. The annotated projects contain description text (short text describing the project) and text from the project website. Negative example projects, obtained from Kickstarter contain text description collected from Kickstarter only. These descriptions are usually long and could be equivalent in length to the content of the project website.

3.5 Data Classification

Data classification involves creating models and classifying the collected data and text according to the four social innovation criteria (objectives, actors and actor interactions, outputs and innovativeness). This is performed mainly using machine learning.

After the additional sources are downloaded and stored in the database, we perform human annotation of the data via the web interface. Expert annotators from ZSI,

[4] https://www.kickstarter.com/

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

University of Strathclyde and University of Manchester annotated the project as social innovation or not social innovation, based on our four inclusion criteria. For each criterion, the annotator gave a mark. These data were used to create a machine learning system with the ability to recognise social innovations projects as well as entities and metadata, which the method is looking for. Each project was annotated by at least two independent annotators. Where there is significant disagreement between annotators, a third independent annotator also annotated those projects.

We generated several classification models that were able to categorise whether an article is about social innovation or not based on the inclusion criteria. We tested a number of classification algorithms, including Decision trees, Support Vector Machines (SVM), Naïve Bayes and Deep neural networks. Ngrams were used, ranging from unigrams to bigrams, and stopwords were removed for some of the classifier runs in order to create cost sensitive algorithms. Based on the preliminary results, we have around 80%-90% of precision, recall and F1 scores. This also results in around 80% of the projects being classified as social innovation (i.e. ESID engine predicts that around 80% of projects satisfies at least one of the four criteria discussed above).

3.5.1 Phase 1: Classification of Social Innovations

For classification of social innovation criteria (objectives, actors and actor-interactions, outputs and innovativeness), we created a number of approaches. We have utilized:

1. Rule-based approach
2. Machine learning-based approach using Sklearn
3. Deep neural network-based approach

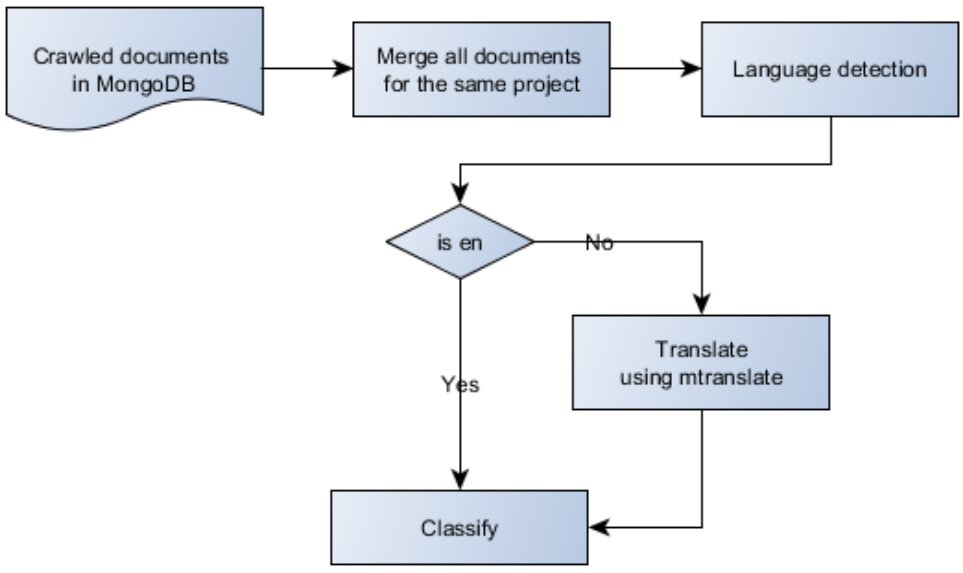


Figure 14: Workflow of the project classification methodology

The overall methodology for creating documents for classification is presented in Figure 14 above. First, the crawled pages were merged into one large document. We used language detection library langdetect[5] in order to detect whether the text was in English

[5] https://github.com/Mimino666/langdetect

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

or not. If the language was not English, we used the mtranslate library which uses Google Translate[6] , in order to translate the text to English. English version of the text were stored in MongoDB and used for classification.

During the evaluation, for all approaches, we measured precision, recall and F1-score.

- Precision (positive predictive value) – number of true entities over number of entities predicted as true

$$Precision = \frac{TP}{TP + FP}$$

- Recall (sensitivity, true positive rate, hit rate) – predicted true instances over the total amount of true instances

$$Recall = \frac{TP}{TP + FN}$$

- F1 – score – combination of precision and recall.

$$F1 = \frac{2 * precision * recall}{precission + recall}$$

In above formula, TP are true positives (truly in the positive class and predicted as positive class), FP are false positives (truly negative class, but predicted as positive class) and FN are false negatives (truly positive, but predicted as negatives.

The rule based approach we employed was keyword matching. For example, for innovativeness, the rule based approach was looking for words such as “innovation”,” innovativeness” and “novelty”. The other rule for innovativeness was looking at whether words like” technology”, “product”, “process”, “service”, “way”, “practice” were close to words like “new”, “novel”, “improved”, “better”, “alternative”. This approach classified as positive about 77% of the projects, however its precision was 0. 55, recall 0.715, while F1-score 0.62.

Initially, we tested various pre-processing techniques in combination with machine learning algorithms. The algorithms that were evaluated were Naïve Bayes, Random Forests, Decision trees, SVM and dense neural networks (with input, one hidden layer with 256 neurons and one output layer) with pre-trained ELMo embeddings. ELMo embeddings were used as one of the recent state-of-the-art techniques in word representations. The pre-processing techniques were applied only on traditional algorithms (Naïve Bayes, Decistion trees, Random Forests and SVMs) and they included 1-grams and combinations of 1-,2- and 3-grams together, with and without stemmer, stopwords, and TF-IDF transformations.

Initial evaluation was done on the actor criteria only in order to find the best performing combination of pre-processing and best performing algorithm. The results are presented in Table 21.

[6] https://github.com/mouuff/mtranslate

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 21: Classification results for actor class using various algorithms and pre-processing techniques.

Algorithm	Precision	Recall	F1-score
Naïve Bayers			
1 gram	0.62	0.94	0.75
1-gram with stemmer	0.64	0.95	0.76
1gram with stemmer and stopword	0.58	0.95	0.72
1 gram, TF-IDF	0.76	0.95	0.84
1-3 grams	0.68	0.98	0.80
1-3 grams with stemmer	0.63	0.94	0.75
1-3 grams with stemmer, stopwords	0.74	0.81	0.77
1-3 grams, TF-IDF	0.77	**0.97**	0.86
1-3 grams, TF-IDF, stemmer	0.83	**0.97**	0.89
1-3 grams, TF-IDF, stemmer, stopwords	0.80	0.95	0.87
1-3 grams, TF-IDF, stemmer, string cleaning	**0.86**	0.95	**0.90**
Random forest			
1 gram	0.75	0.83	0.79
1-gram with stemmer	0.82	0.82	0.82
1gram with stemmer and stopword	**0.86**	0.84	0.85
1 gram, TF-IDF	0.80	0.74	0.77
1-3 grams	0.84	0.76	0.80
1-3 grams with stemmer	0.82	0.84	0.83
1-3 grams with stemmer, stopwords	0.80	0.82	0.81
1-3 grams, TF-IDF	0.84	0.77	0.80
1-3 grams, TF-IDF, stemmer	0.81	0.80	0.81
1-3 grams, TF-IDF, stemmer, stopwords	0.81	0.79	0.80
1-3 grams, TF-IDF, stemmer, string cleaning	0.86	0.79	0.82
Decision trees			
1 gram	0.78	0.80	0.79
1-gram with stemmer	0.82	0.81	0.82
1gram with stemmer and stopword	0.78	0.84	0.81
1 gram, TF-IDF	0.72	0.74	0.73
1-3 grams	0.78	0.76	0.77
1-3 grams with stemmer	0.83	0.79	0.81
1-3 grams with stemmer, stopwords	0.79	0.82	0.80
1-3 grams, TF-IDF	0.79	0.79	0.79
1-3 grams, TF-IDF, stemmer	0.84	0.76	0.79
1-3 grams, TF-IDF, stemmer, stopwords	0.77	0.79	0.78
1-3 grams, TF-IDF, stemmer, string cleaning	0.76	0.75	0.76
SVM			
1 gram	0.85	0.58	0.69
1-gram with stemmer	0.79	0.56	0.66
1gram with stemmer and stopword	0.79	0.62	0.69
1 gram, TF-IDF	0.00	0.00	0.00
1-3 grams	0.73	0.20	0.32
1-3 grams with stemmer	0.83	0.57	0.68
1-3 grams with stemmer, stopwords	0.80	0.59	0.68
1-3 grams, TF-IDF	0.47	1.00	0.64

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

1-3 grams, TF-IDF, stemmer	0.00	0.00	0.00
1-3 grams, TF-IDF, stemmer, stopwords	0.00	0.00	0.00
1-3 grams, TF-IDF, stemmer, string cleaning	0.48	1.00	0.65
Dense NN with ELMo embeddings	0.82	0.90	0.86

As can be seen from the Table 21, the best results were yielded by Naïve Bayes followed by ELMo embedding pre-trained neural networks and random forests. This may be due to the data-set's small size and relatively noisiness. On such datasets Naïve Bayes is able to generalize fairly well (John & Langley, 1995).
We have used Naïve Bayes with text cleaning, stemming and TF-IDF transformation on other criteria. The results are presented in Table 22.

Criteria	Precision	Recall	F1- score
Objectives	0.81	1.00	0.89
Actors	0.86	0.95	0.90
Outputs	0.76	0.99	0.86
Innovativeness	0.82	1.00	0.90
Social innovation	0.78	0.97	0.87

Table 22: Classification results for all four social innovation criteria and general single criteria.

As it can be seen from Table 22, all 4 social innovation criteria (objectives, actors, outputs, innovativeness) preform in similar range (0.86-0.90 F1-score). Similar to these, within the same range is a single social innovation criterion (0.87 F1-score). We have further analysed these results in our error analysis, analysing errors by the language and topics of the projects.

3.5.2 Error analysis

The dataset was tested on 319 projects in total. Out of these projects, 133 projects were added to balance the dataset, as non-social innovation projects, mainly collected from Kickstarter. The projects were in 15 languages, of which 275 projects were in English, 11 in Italian, 8 in German, 4 in Danish, 4 in Russian, 3 in Dutch, 2 in Spanish, 2 in French, 2 in Serbo-Croatian, 2 in Romanian. There were 147 projects with society as a topic, 229 projects with health as a topic and 20 with security as a topic.

In terms of models predicting the actor criteria, there were 9 false negatives (predicted negative, actually positive). Out of these 9 projects, only one project was in Lithuanian, while the rest were in English. Three projects were located in UK, two in USA, and one in Lithuania, South Africa, Belgium and Mexico. Only 2 projects had a description of about 500 characters long, while other projects had descriptions longer than 9000 characters (some having web pages with over 1,000,000 characters of crawled text).

There were 37 false positives (predicted as satisfying the criteria, while they did not). Out of these, 5 were located in UK, 7 did not have a defined location (out of which 6 were Kickstarter projects) and Denmark had 3 false positive projects. Other countries had up to two misclassified projects. Six of the false positive projects had less than 400 characters (out of which five were false projects from Kickstarter), while the other 31 had more than 2000 characters of description and crawled text. Out of these, 37 were falsely

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

predicted as positive projects, 7 were added from Kickstarter. Less than a third of these projects covered society related topics (11/37, 2 of which had less than 500 characters text, all were in English and 3 were from Kickstarter), one had security as a topic (and the project was originally in German), while 23 projects that were misclassified were of the health topic (24/37, 4 of which had under 500 characters of text, while 17 were in English and 5 from Kickstarter).

For the objective criteria, there were no false negatives (Recall = 1), while there were 39 false positives. Of these, 22 were from Kickstarter, 35 were in English, while 2 were in Spanish and another 2 in Danish. Projects location were UK (4), USA (2), Denmark (2), Spain (2), Netherlands, Poland, Canada, and Kenya. Of these misclassified projects, 11 had description lengths smaller than 500 characters. Society topics were represented in 15 of them (2 had less than 500 characters, only one was non-English), health topic was represented in 22 projects (6 had less than 500 characters, one was non-English), and security only in 2 projects (out of which one had less than 500 character, both were English). In the set for objectives, there were 170 projects in total having society as a topic, 238 having health as a topic, and 28 having security as a topic, of 318 projects considered.

For the objective criteria, there were 48 false positive instances (out of which 16 were from Kickstarter). Out of these 48 projects, 44 are originally written in English, while 2 projects had an Italian website, and one a Danish one, and another, a French website. There were 6 projects with less 500-character long descriptions (smallest has 208). Most projects were in the UK (10), followed by the US (4), Italy (2), etc. Topic-wise, 17 projects were of the society topics (out of 159 in total), 31 were of health topic (out of 229 in total), and 6 of security (out of 27 in total). For the objective criteria, there was only one false negative, which had a Polish website and had societal and health related topics.

For the innovativeness criteria, there were 321 projects in total, out of which 38 were false positive instances and 0 false negatives. Most of these misclassified projects were in English, apart from 1 in German, Spanish and French. Seven false positive projects had less than 500 characters in the classified text. Six projects were located in UK, while the rest were in Switzerland, Italy, France, Spain, Netherlands, etc. There were 18 misclassified projects with society as a topic (172 in total), 31 projects with health (233 in total) and 2 with security (19 in total).

For the generic social innovation criteria, there were 317 projects in total, with 43 false positives and 4 false negative instances (out of which 3 were in English originally, one in Lithuanian, one had society and health as a topic, the rest had no topics in health, society and security category). Out of the false positive projects, 3 had less than 500 characters, 27 were in English, 3 in German, and there were projects in Turkish, Spanish, Romanian, etc. Regarding the topics, 27 false positives had society as a topic (167 in total), 36 had health (235 in total), and 4 had security (17 in total).

Apart from the flexibility given to the users, the machine learning algorithm did not really benefit from breaking up the criteria into more concrete concepts. The distribution of topics in misclassified instances is similar to the overall distribution of those of the same topics. A similar situation is observed with languages and locations.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.6 Information Extraction

Information extraction involves extracting metadata about the project, such as information about the organisation involved in the project, where it is located, as well as its topics. This part was performed using rule-based and machine learning-based named entity recognition tools and information extraction tools.

Once the projects and organisation were classified, ESID enriched the database entries with additional features of the projects and actors, such as their locations, relationships, etc. This was performed using a number of named entity recognition tools and tools for relation and information extraction. The labelled instances were post-processed and normalised. For these tasks, we used some of the existing tools, however, for certain features, the necessary tools do not exist. Also, domain specific (social innovation domain) challenges were addressed. For the features where there were no available tools and where domain-specific challenges could not be addressed using existing tools, we developed extractors either by creating rule-based systems or machine learning systems, depending on which was more appropriate. For the machine learning based information extraction tool, we utilised Conditional random fields (CRF). The algorithm which produced the best performing model was used in the production system.

Information extraction has the aim to extract metadata and additional information about projects and actors. The methodology should extract mainly information about actor-project relationships, project locations, actor locations, topics, aims and objectives of the project. There are a number of additional variables that may be added in the future, but were not in the scope of this first phase of the ESID project. Some of these include funding information, relationships between projects or actors, size of organisations, etc. All these variables were extracted from text that is either retrieved as project/actor description in the source data source, or crawled from the project/actor website.

In this section, we present the methodology for each of the main variables required by both KNOWMAK and ESID databases. The focus of the ESID database is on social innovation projects, and through the projects, relevant actors are found. As such, we start with variables related to projects.

3.6.1 Project location

For our ESID database, the project location is an important aspect, and hence, we need to account for the project locations as well. We started the process of project location discovery by looking at the actors' location and by so doing, inferred the location of the project and hence attributed knowledge creation. Social innovation is different in a sense that projects have their location.

Certain data sources, such as SI-Drive or Digital Social Innovation include locations of the projects. We trusted these locations and utilized them as presented. However, many other data sources did not include data about the project location or their connection with the actors. In this case, it was necessary to text mine this information from the available text about the project.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

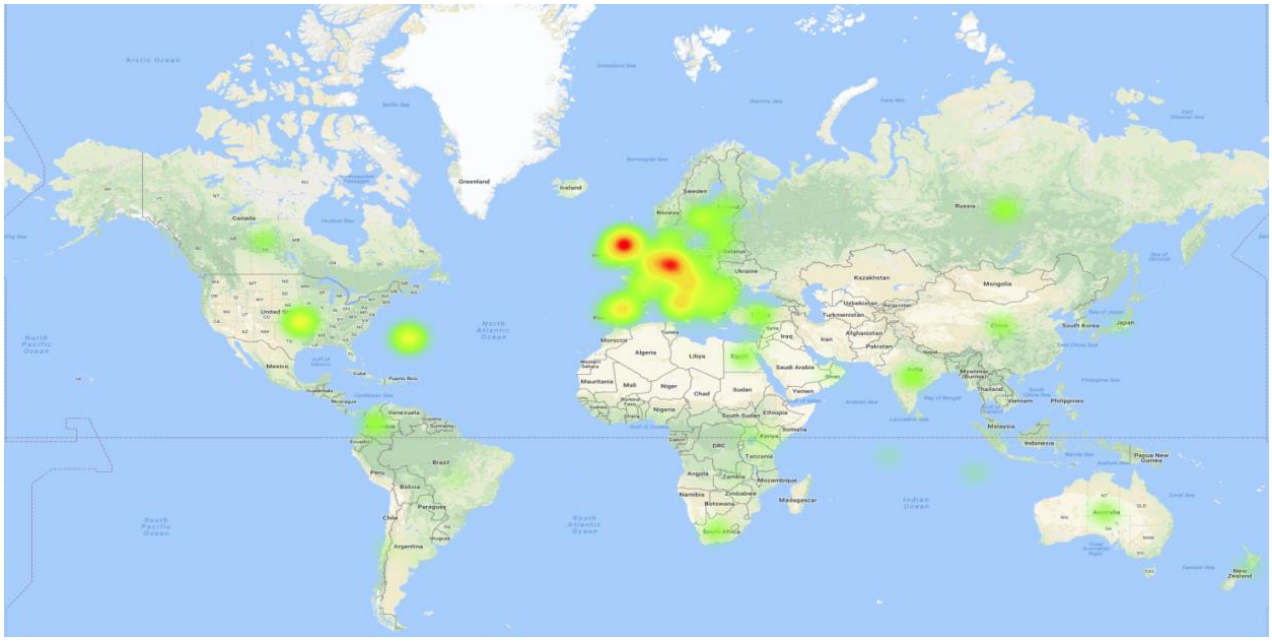
The level of detail for both KNOWMAK and ESID projects are city level and up (country). Therefore, each project was required to have one or more instances of cities and countries with attributed geographical coordinates.

One way of extracting project locations is by using Stanford NER, which contains locations. Stanford NER is capable of highlighting locations; however, it is necessary to normalize the recognised locations. For example, it often recognises the city, but not the country. Using GeoNames[7], it was possible to attribute these cities to the corresponding countries and geographical coordinates.

Often, texts about projects would mention many locations, giving examples or just referring to the problems or solutions in these regions. Projects would have most often only one location, therefore, we adopted a methodology where we select the most frequently mentioned location in the text as the main location of the project. Where it was not possible to find a city or even a country in the text, we adopted an alternative methodology that looks at the domain name extension of the website.

We utilized Stanford NER and a set of heuristics on the relevant pages where the location could likely be found. We do not report the performance of Stanford NER, as it has been widely tested in other literature.

We performed some of the experiments using domain name and the most frequent location extraction. The initial heat-map with the project distribution is presented in Figure 15.
As it can be seen from this initial heatmap, ESID covers the projects in Europe quite extensively. However, it also does have some projects from other continents. Our aim is to grow ESID, and while its primary focus is Europe, it will incorporate projects from all over the world. This is what we hope to achieve in Phase 2, in the expansion phase of our implementation.



[7] http://www.geonames.org

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Figure 15: Initial heat map of project locations in ESID database

3.6.1.1 Location extraction methodology

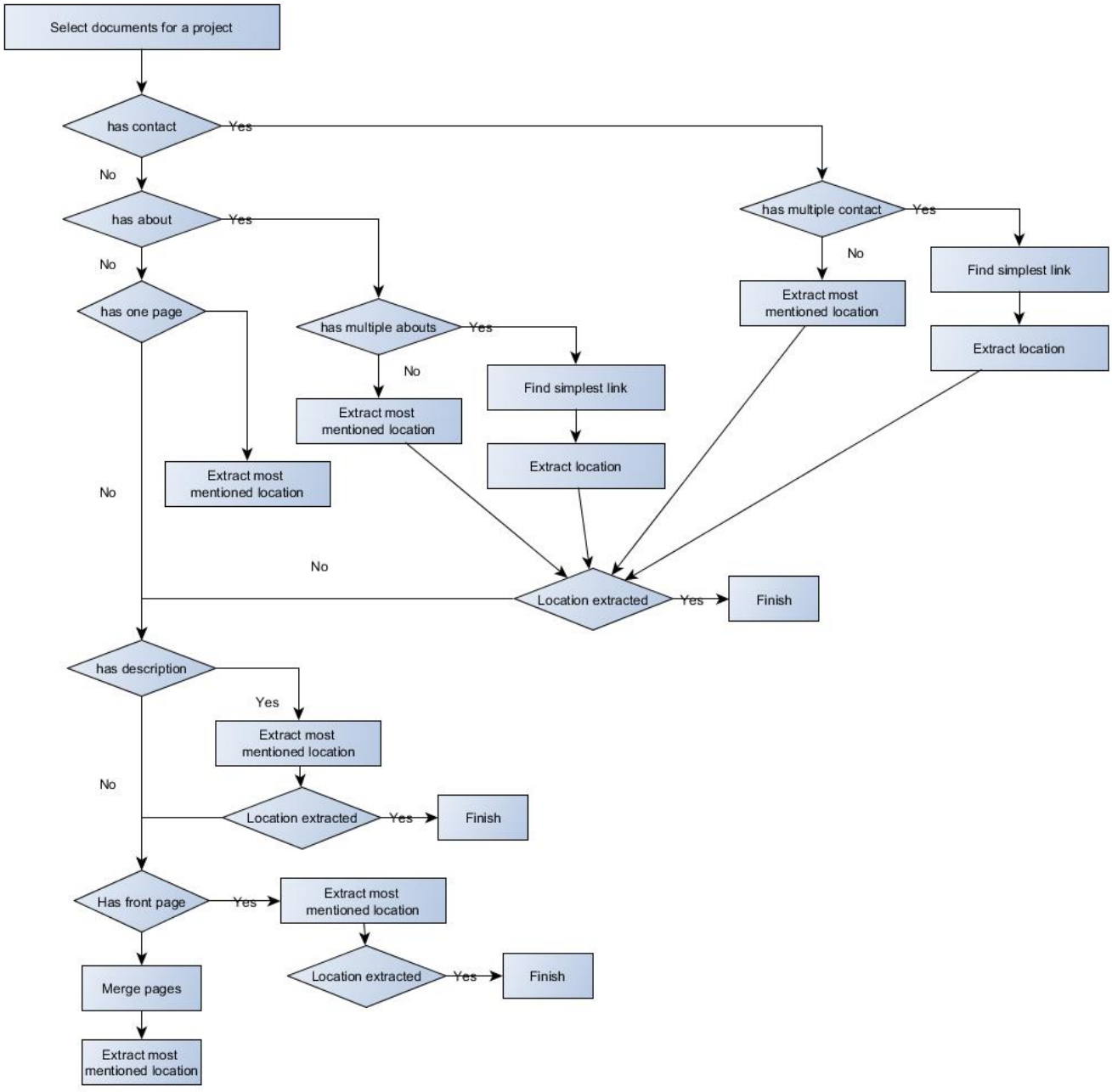


Figure 16: Location extraction workflow

Location extraction is one of the focus areas of information extraction in ESID. As many projects did not report their locations in the original sources, we developed a methodology for extracting locations which relies on assigning a confidence score based on the page of the website where the location was found.

Our Location extraction methodology first used Stanford NER in order to tag the location in the text. We also used Semanticon to map the identified location in order to assign countries where there are none, and map cities. We used a process which systematically

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

looked through the crawled pages of a website, starting from pages which are more likely to have the actual location information of the project. We started with the most reliable page of the website where the location could be found, which is the *Contact* page. Where the location was not found on the *Contact* page, it looked for the location in the A*bout* page. If the location was found in the *About* page, it would have a lower confidence score, but the location was likely right. We went through each page in turn, the *Contact* page, the *About* page, *One* page, the *Main* page, and the *General* page. As we moved to these pages, we lowered the confidence score assigned to the location identification in these pages. This is because we were working on the belief that the likelihood of the actual location information being in these pages is lower. In our methodology, we went through the pages in the set order that is represented in the workflow diagram in Figure 16, lowering the confidence every time it moved on to another page.

From the workflow, we see that the first step was to select documents related to a project from our crawled data, we checked if there was a *Contact* page first. If there was a *Contact* page, we checked if there were multiple contacts. Where there were, selected the simplest link and extracted the location from there. Where there weren't multiple contacts, we extracted the most frequently mentioned location as our actual location. Where we were unable to track the location on the *Contact* page, or if there was no *Contact* page, we moved on to the *About* page. From here again, we check if there are multiple about pages, and if there are, just as we did with the *Contact* page, we find the simplest link and extract the location from there. Where there was just a single *About* page, then we selected the most frequently mentioned location on that page and extracted it as our location. With this location, the confidence level was lower than the confidence level we assigned to the location obtained from the *Contact* page. Where there was no *About* page, we checked the *One* page available and simply extracted the most frequently mentioned location on that page as our location, with a lower confidence than that of the *About* page.

Where none of these pages existed, or where we were unable to find a location on these pages, we moved on to the project description, where we extracted the location with the most mentions. In the case where we were still unable to find the location there, we then checked the *Front* page of the webpage, and extracted the most mentioned location as the actual location. Where all of these failed, and we were unable to determine a location, we merged all the available pages of the website and then selected the location with the most mentions in these aggregated pages and extracted that as our actual location.

In Table 23 andTable 24 below, we show the distribution of where the location of our projects was extracted from on the project webpage. For some of the projects, we extracted the location information from the different Social Innovation databases we used as sources for our projects. However, not all the projects had the location information, and some had part of the location information, such as the country only. We give an overview of the project webpage sections where we extracted the location information from in the tables and break this down to show the distribution of projects where this information had to be mined from the crawled data with our location extraction technique. For some of the projects, the location information was extracted from the domain extension of the project webpage.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Table 23: Project location source page

Location Found	Number of Projects
Description	197
About	149
General	136
Main page	149
AfterAll	89
One page	409
Contact	27
Domain extension	120

Table 24 : Project Location Source distribution

Project Location Source	City and Country text mined	City text mined, Country from Data Source	City and Country from Data Source
Description	23	174	0
About	56	92	1
General	25	109	2
Main Page	36	113	0
AfterAll	51	38	0
One page	67	342	0
Contact	5	22	0
Domain Extension	0	0	120

3.6.2 Summarization

Projects in ESID need to have a description of the project. In order to automatically facilitate creation of description, we utilized methods of automated summarization of the text from the project websites.

In order to gather data for training summarization algorithms, we performed a set of annotation tasks in which annotators annotated sentences that described how each project satisfies some of the following social innovation criteria:

- **Social objective** - project addresses certain (often unmet) societal needs, including the needs of particular social groups; or aims at social value creation.
- **Social actors and actor interactions** - involves actors who would not normally engage in innovation as an economic activity, including formal (e.g. NGOs, public sector organisations etc.) and informal organisations (e.g. grassroots movements,

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

citizen groups, etc.) or creates collaborations between "social actors", small and large businesses and the public sector in different combinations.

- **Social outputs** - creates socially oriented outputs/outcomes. Often these outputs go beyond those created by conventional innovative activity (e.g. products, services, new technologies, patents, and publications), but conventional outputs/outcomes might also be present.
- **Innovativeness** - There should be a form of "implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organisational method".

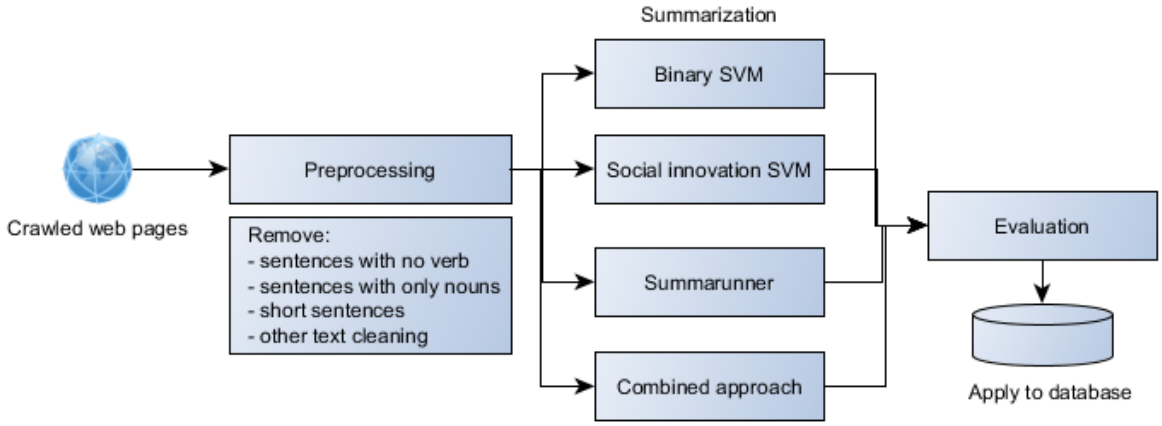


Figure 17: Summarization methodology

The distribution of annotated sentences is presented in Table 25. Annotated data, descriptions from the original data sources and crawled websites were used for training and evaluating summarization approaches.

We also provide an overview of the summarisation methodology in Figure 17. The first step we performed was to pre-process the documents and remove sentences which have no verbs, have only nouns, are too short, as well as other forms of text cleaning. We discuss these in the following subsection. Then we passed this pre-processed data through our chosen summarisation algorithm. The produced summaries were then evaluated before being passed on to the database.

Table 25: Number of sentences satisfying social innovation criteria

Criteria	Number of sentences
Social innovation criteria	
Objectives	374
Actors	217
Outputs	309
Innovativeness	256
Not satisfying any criteria	3167
Binary (inside/outside summary)	
Inside	2459
Outside	12962

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.6.2.1 Data Collection and Dataset Generation

The initial set of social innovation projects was collected using pre-existing databases of social innovation. These databases included MOPACT, Digital Social Innovation, InnovAge, SI-Drive, etc. These data sources contained structured data, with project descriptions which were created by humans, websites and social media. A set of crawlers were created which were able to locate and crawl the structured data points on these pages and store them in our database. The small number of data sources which contained descriptions were used for the creation of our training set (Milošević, 2019).

A total of 3560 projects were collected, and out of these, 2893 projects had websites which were identifiable. A crawler was then created to collect text from the websites. Annotations were then performed according to the social innovation criteria outlined above, and we have presented the breakdown of the outcome of this in Table 25.

3.6.2.2 Data cleaning

The data from the websites could be quite noisy, as the crawler was collecting all the textual information, including menus, footers of the pages and at times, advertisements. Additionally, many pages contained events and blog posts that were not relevant for describing the core of the project. Therefore, we performed some data cleaning before proceeding with training of the summarizers.

In order to reduce the amount of irrelevant text in the form of menus and footers, we performed part of speech tagging and excluded sentences that do not contain verbs. For further summarization, only *Main* pages, *About* pages and project *description* pages were used. Where the page was not in English it was first translated, using Google Translate.

3.6.2.3 SVM based summarizer

The first summarization approach we employed assumed that the summarization task could be modelled as a classification task, where sentences would be classified as being part of a summary or not. It was hypothesized that words in a sentence could indicate whether it described the project (e.g. "project aims to...", "the goal of the project is to...", etc.) or not.

In order to create a training data set, we utilized projects that had both project description in the original data sources and the crawled websites. As the descriptions were created by humans, they usually could not be matched with the sentences from the website. In order to overcome this issue, we generated sent2vec embedding vectors of the sentences in both the description and the crawled text. We then computed cosine similarities between the sentences from the description and those from the crawled text. If the cosine similarity is higher than 0.8, the sentence was labelled as part of the summary, otherwise it was labelled as a sentence that should not be part of the summary.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

These sentences were used as training data for the SVM classifier. Before training, we balanced the number of positive (sentences that should be part of the summary) and negative (sentences that should remain outside the summary) instances. The bag-of-words transformed to TF-IDF scores, the position of a sentence in the document (normalized to the score between 0-1) and keywords were used as features for the SVM classifier. The keywords were extracted using KNOWMAK ontology API[8], which for the given text returns grand societal challenge topics and a set of keywords that were matched for the given topic and text[9].

3.6.2.4 Social innovation criteria classifier

The social innovation criteria classifier utilized an annotated dataset. In this dataset, sentences that were marked as explaining why a project satisfies any of the social innovation criteria (objectives, actors, outputs, innovativeness), were used as positive training instances for the SVM (Support Vector Machine) classifier. The classifier used a bag-of-words transformed to TF-IDF scores as its set of features.

3.6.2.5 Summarunner

Summarunner is an extractive summarization method developed by IBM Watson that utilizes recurrent neural networks (GRU). If compared using ROUGE metrics, the algorithm outperforms state-of-the-art methods. The method visits sentence sequentially and classifies each sentence by whether or not it should be part of the summary. The method uses a 100-dimensional word2vec language model. The model was originally trained on a CNN/DailyMail data set. The social innovation data set that we created was quite small and not sufficient for training a neural network model (about 350 texts compared to over 200,000 in DailyMail data). However, we performed a model fitting on our social innovation data set.

3.6.2.6 Stacked SVM-based summarizer and Summarunner

Our final summarization method was developed as a combination of SVM-based method and Summarunner (Milošević, 2019). We noticed that the binary SVM model produced quite long summaries and as such could be efficient for the initial cleaning of the text. Once the unimportant parts were cleaned up by the SVM-based classifier, Summarunner shortened the text and generated the final summary.

3.6.2.7 Evaluation methodology

The evaluation of summarization techniques is a challenging process, therefore, we employed several techniques. In order to evaluate our methodologies and select the best performing model we used ROGUE metrics, human scoring and two topic-based evaluation methods.

[8] https://gate.ac.uk/projects/knowmak/
[9]

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

ROUGE metrics are the most popular and widely used summarization scoring approaches which were presented back in 2004. As such, we utilized them as well.

A good summary should include the most important topics from the original text, hence, topic-related metrics can be devised. We used two topic based metrics: one was based on KNOWMAK ontology and the proportion of matched topics related to EU defined Grand Societal Challenges[10] and Key Enabling Technologies[11] in the original and summarized text. The other method was based on latent Dirichlet allocation (LDA). We extracted 30 topics using LDA from the merged corpus of original texts and summaries and then we have calculated the proportion of topics that match. In order to prevent favouring long summaries, we normalized the scores, assuming that the perfect summary should be no longer than 25% of the length of the original text (longer texts were penalized) (Milošević, 2019).

As our SVM classifiers utilize classification, we calculated their precision, recall and F1-scores. These are measures commonly used for evaluating classification tasks. These metrics were calculated on a test (unseen) data set, containing 40 documents (286 sentences labelled as inside summary, 2014 sentences as outside). The results can be seen in Table 26.

Classifier	Precision	Recall	F1-score
Binary SVM	0.8601	0.7130	0.7594
Objectives SVM	0.8423	0.5601	0.6226
Actors SVM	0.8821	0.4687	0.5659
Innovativeness SVM	0.8263	0.4456	0.5166
Outputs SVM	0.8636	0.6284	0.7089

Table 26: Evaluation based on classification metrics (precision, recall and F1-score) for classification-based summarizers (binary and social innovation criteria-based)

The data set for training these classifiers was quite small, containing between 200-400 sentences. It is interesting to note that the criteria classifiers containing a larger number of training sentences, perform with a better F1-score (Objectives and Outputs). This indicated that scores could be improved by creating a larger data set. The classifiers performed with quite good precision, which meant there were few false positive sentences (the majority of the sentences that ended up in the summary were correct).

As aforementioned, ROUGE metrics are commonly used in summarization literature, hence, we evaluated all our summarization approaches with ROUGE 1, ROUGE 2 and ROUGE-L metrics. The evaluation was performed again on an unseen test set, containing 40 documents and their summaries. The results can be seen in Table 27 .

Classifier	ROUGE 1	ROUGE 2	ROUGE-L
Binary SVM	0.6096	0.5544	0.5553
Social innovation SVM	0.6388	0.6140	0.5846
Summarunner	0.6426	0.5788	0.5762
Binary SVM + Summarunner	0.5947	0.5197	0.5279

[10] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges
[11] http://ec.europa.eu/growth/industry/policy/key-enabling-technologies_en

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Binary SVM + Summarunner Relative Length	0.5496	0.4731	0.4668

Table 27: ROUGE scores for the developed summarization methodologies

Summarunner had the best performance based on unigram ROUGE (ROUGE-1) score. However, the social innovation SVM-based summarizer performed better in terms of bigram ROUGE (ROUGE-2) and ROUGE-L score (measuring longest common token sequence). Based on these results, it was possible to conclude that a specifically crafted classifier for the problem would outperform a generic summarizer, even if it was trained only on a small data set. Stacked binary SVM and Summarunner performed worse than the single summarizers on their own, in terms of ROUGE.

In order to further evaluate the methodologies used, we used an LDA-based metric. The assumption behind using this approach was that a good summarizer would have a high number of topics in the summary/description and the original text matching (Milošević, 2019). The results of the LDA topic similarity evaluation can be seen in Table 28.

Classifier	LDA Topic Similarity
Binary SVM	0.2703
Social innovation SVM	0.2485
Summarunner	0.2398
Binary SVM + Summarunner	0.2683

Table 28: LDA topic similarity scores for the developed summarization methodologies

The most matching topics were found with the binary SVM classifier. However, this classifier also produced the longest summaries. Stacked SVM and Summarunner showed a similar performance in terms of matches, with much shorter summaries being generated.

The second topic-based approach utilized topics about grand societal challenges and key-enabling technologies retrieved from the KNOWMAK topic-modelling tool. The results can be seen in Table 29.

Classifier	KNOWMAK Topic Similarity
Binary SVM	0.3725
Social innovation SVM	0.3625
Summarunner	0.3025
Binary SVM + Summarunner	0.3025

Table 29: Topic similarity evaluation using KNOWMAK ontology topics

The binary SVM summarizer showed the best performance according to this metric. It was closely followed by the social innovation summarizer.

Finally, summaries were scored by human annotators. Human scorers were presented with an interface containing the original text and a summary for each of the three methods (binary SVM, social innovation SVM and Summarunner). For each of the summaries they could give a score between 0-5. In Table 30 averaged scores made by the human scorers are presented. We also averaged the scores in order to account for document length. In order to do that we used the following formula:

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

$$LengthAveragedScore = \frac{docLen - summaryLen}{docLen} * human\_score$$

Classifier	Number of ratings	Human Score	Length averaged human score
Binary SVM	23	2.7391	0.8647
Social innovation SVM	20	2.4500	1.6862

Table 30: Human scores for the developed summarization methodologies

The best human scores were for binary SVM. However, this classifier excluded only a few sentences from the original text, and it was generally creating longer summaries. If the scores are normalized for length, the best performing summarizer was that based on social innovation criteria, followed by Summarunner. At the time of the manual scoring, the stacked approach consisting of binary SVM and Summarunner was not yet developed, so results for this approach are not available.

We used stacked (SVM+Summarunner) and social innovation classifier in order to generate summaries for our database. The Stacked model was used as a fall-back, in case summary based on social innovation model was empty or contained only one sentence. The approach was summarizing and generating project descriptions where either the description was too long (longer than 1000 words), or was missing (Milošević, 2019).

The Binary summarizer performed well over a number of metrics, and since adding more data would improve the performance of the algorithm, we scaled the dataset to contain about 500,000 sentences. The initial dataset was not balanced, but we have performed experiments with both balanced and non-balanced data. The results are presented in Table 31 below:

Algorithm	**Precision**	**Recall**	**F-score**
Naïve Bayes - unbalanced	**0.93**	**0.93**	**0.91**
in summary	0.89	0.35	0.50
outside summary	0.93	0.99	0.96
Naïve Bayes - balanced	**0.88**	**0.87**	**0.87**
in summary	0.81	0.96	0.88
outside summary	0.95	0.79	0.86
Random forests - unbalanced	**0.97**	**0.97**	**0.97**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

in summary	0.96	0.78	0.86
outside summary	0.98	1.00	0.99
Random forests - balanced	**0.94**	**0.94**	**0.94**
in summary	0.96	0.92	0.93
outside summary	0.92	0.96	0.94
CNN – balanced	**0.92**	**0.92**	**0.92**
in summary	0.89	0.93	0.91
outside summary	0.94	0.90	0.92
CNN – unbalanced	**0.97**	**0.96**	**0.97**
in summary	0.76	0.89	0.82
outside summary	0.99	0.97	0.98

Table 31: Summarisation results with Precision, Recall and F-Scores

This model initially generated summaries that were later manually reviewed and are part of KNOWMAK interface.

3.6.3 Topic Classification

Topic classification was performed using the KNOWMAK ontologies. KNOWMAK's topic classification algorithm was used, and we implemented this through API calls to KNOWMAK's ontologies to match the topics. The two ontologies which were used to define the topics were KNOWMAK's Key Enabling Technologies (KET) topics and Grand Societal Challenges (SGCs).

Topic scores were returned and a threshold was arbitrarily set at 0.7 and all topics that exceeded this threshold were assigned as suitable topics for the project in question.

The KNOWMAK project aims at creating a web-based tool which gives a provision of interactive visualisations and state-of-the-art indicators on knowledge co-creation in the European Research Area (ERA). One of the integrative elements the KNOWMAK tool is structured around is Research topics, which it handles through the development of ontologies around Societal Grand Challenges and Key Enabling Technologies.

In KNOWMAK, the ontology acts as a bridge between the users and the underlying data, as it on the one hand enables users to browse topics, access related topics, and widen

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

their search, while on the other hand enabling connection between the data sources and the relevant topics by means of annotation (tagging).

The ontology[12] structure of KNOWMAK is built in three layers, where the first layer corresponds to KET and SGC. The ontology structure is hierarchical, but also allows for multiple inheritance. KNOWMAK implements this structure under the belief that multiple inheritance is required, as terms and concepts are not unique to single KETs and SGCs. There is a lot of overlap in technology between these areas, and this needs to be reflected in the system.

Topics within the ontology are associated with vocabularies, which are sets of words which are associated with one or more topics. These vocabularies have two functions in KNOWMAK;

- They link topics with data items, for example, publications. Data items are annotated with keywords, and through them, can therefore be associated to topics
- They link topics with search queries by users

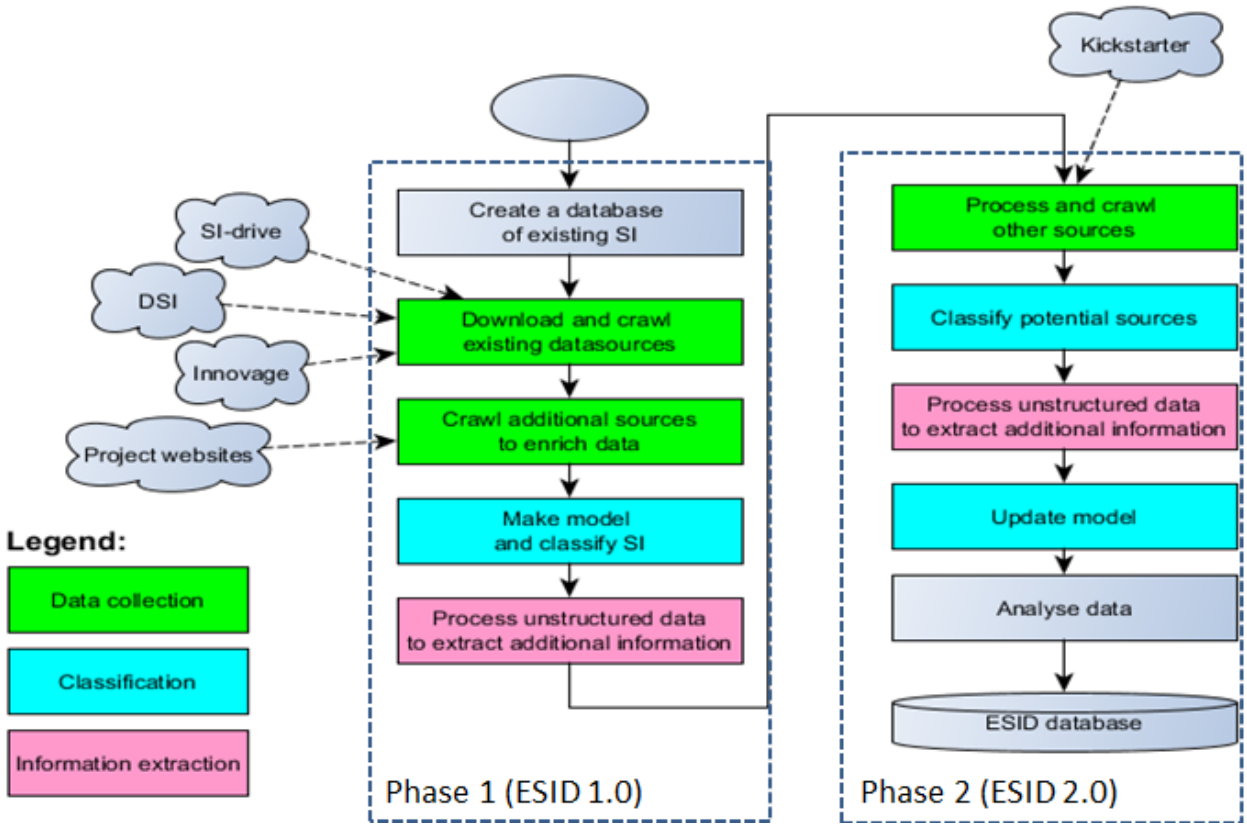
KNOWMAK's three classical data sources include publication, patents and European projects, which are linked with the ontology, based on word frequencies in project title and summaries, patent description and associated codes for patent classes, and publication titles and summaries

The basic idea is that each data item is attributed to topics that have received the highest scores – possibly adopting a minimum threshold for the score. The assignment of data to topics will allow constructing indicators at different aggregation levels, for example counting the publications attributed to a given geographical space and to a particular topic.

Figure 18: The ESID Engine

[12] https://project.knowmak.eu/integrative-elements/ontology/

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091



This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

3.7 Data integration with KNOWMAK

This task relates to the integration of ESID into KNOWMAK. This included data cleaning, data harmonisation and negotiating requirements. All development work was done by Manchester, while the conceptual and the design work done by Strathclyde.

More specifically, this involved the following steps:

- Identification and transmission to the KNOWMAK central database of the list of standardized actors including the descriptors foreseen in the KNOWMAK manual.
- Transmission to KNOWMAK of the list of social innovation projects and of the related indicators required for the KNOWMAK tool, more specifically:
    - The project identifier
    - The project title
    - The project website.
    - A standardized project summary where available (based on the existing data sources or project objectives described above).
    - Any ontology classes attributed to the project
    - The project location(s).
    - A dummy variable to identify European-level projects than cannot be localized precisely.
    - The list of involved actors in the project where available. When applicable, these actors are linked with the standardized actors' table for the classical research actors and for the standardized social innovation actors.
    - Scores for the four social innovation criteria.
- Coordination with AIT for the preparation and integration of the data.
- Support for testing the indicators are results provided by the KNOWMAK tool.

The aim of Knowledge in the Making (KNOWMAK) project is to develop an interactive tool which allows selected groups of users to visualise and analyse the production of knowledge in the European Research Area, with a particular focus on knowledge related to Societal Grand Challenges (SGC[13]) and Key Enabling Technologies (KET[14]).

The tool was based on three existing data sources on knowledge production in the European Research Area (ERA), i.e. scientific publications derived from the Web of Science database (CWTS-WoS database), patents derived from PATSTAT (UPEM-PATSTAT database) and European projects derived from CORDIS (AIT-EUPRO database). Additionally, the project will develop additional datasets, one concerned with social innovation projects and actors, the other with user attention as observed through social media.

[13] https://ec.europa.eu/programmes/horizon2020/en/h2020-section/societal-challenges.
[14] https://ec.europa.eu/programmes/horizon2020/en/area/key-enabling-technologies.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

The goal of the KNOWMAK project was to develop a web-based tool, which provided interactive visualisations and indicators on knowledge co-creation in the European research area. The tool was structured around three integrative elements:

- Research topics, by developing an ontology on Societal Grand Challenges (SGC) and Key Enabling Technologies (KET);
- Research actors, both "conventional" and social actors;

Geographical spaces and, more specifically, countries and regions

3.7.1 Social innovation data in KNOWMAK tool

Social innovation is a part of KNOWMAK. Social actors and social innovation projects are thus incorporated into the KNOWMAK database and, through that, into the KNOWMAK web application. The handling of data on social innovation in the KNOWMAK database builds upon that of data on the classical innovation indicators using projects, publications, and patents.

The information necessary for KNOWMAK database was transferred from ESID to KNOWMAK. Information necessary for KNOWMAK was stored and generated from the ESID MySQL database. These were extensively annotated by humans to ensure data integrity and accuracy before being transferred over to KNOWMAK. All projects whose data were transferred to KNOWMAK are marked as "knowmak_ready" in the database. These form the curated dataset of ESID.

The social innovation indicators are described as extending beyond what is planned for the traditional indicators. We sketched out a database design in two stages, first paralleling the structure used for the classical innovation sources (especially EU projects) and second enriching the design with the additional information needed for the extended set of indicators.

ESID database follows the data formats for actors and projects prescribed by the KNOWMAK project. The database stores mandatory data required by KNOWMAK, such as project/actor name, project/actor identifier, actor/project website, project/actor type and subtype.
In Figure 19, we show the simplified design. We incorporate tables for social actors and social innovation projects, and a table of the specific sites at which social actors engage in knowledge creation. KNOWMAK and ESID database also hold the locations of the projects.

The tables are linked to one another, showing in which project the social actors take part and at what locations. The tables are as well connected to the existing tables of geographical and topical information, which link in turn to tables of territorial statistics and ontological properties (not shown). Two additional tables are provided, showing which social innovation projects involve international or intercontinental collaborations; these can be computed from the geographical information.

Social actors are linked to the table of standardized key actors. The key actor table

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

incorporates additional data on social actors, but the table structure does not change. The actors table though have not all been linked to the projects in ESID, and more of the work on the actors will be done in the second phase of the ESID project.

Geographical information for social innovation activities can be handled as for the classical data sources. For each social innovation project, the latitude and longitude of the project is used to determine first the country and adapted NUTS region, which together constitute a geographical identifier. The same is done for social actors. The geographical coordinates only need to be accurate to the city level, rather than the street address, to produce reliable country and region assignments.

The structure for social innovation in the simplified design is essentially identical to that for European Framework Programme project participations. Thus, volume-based indicators can be computed for social innovation exactly the same way as for EU projects.

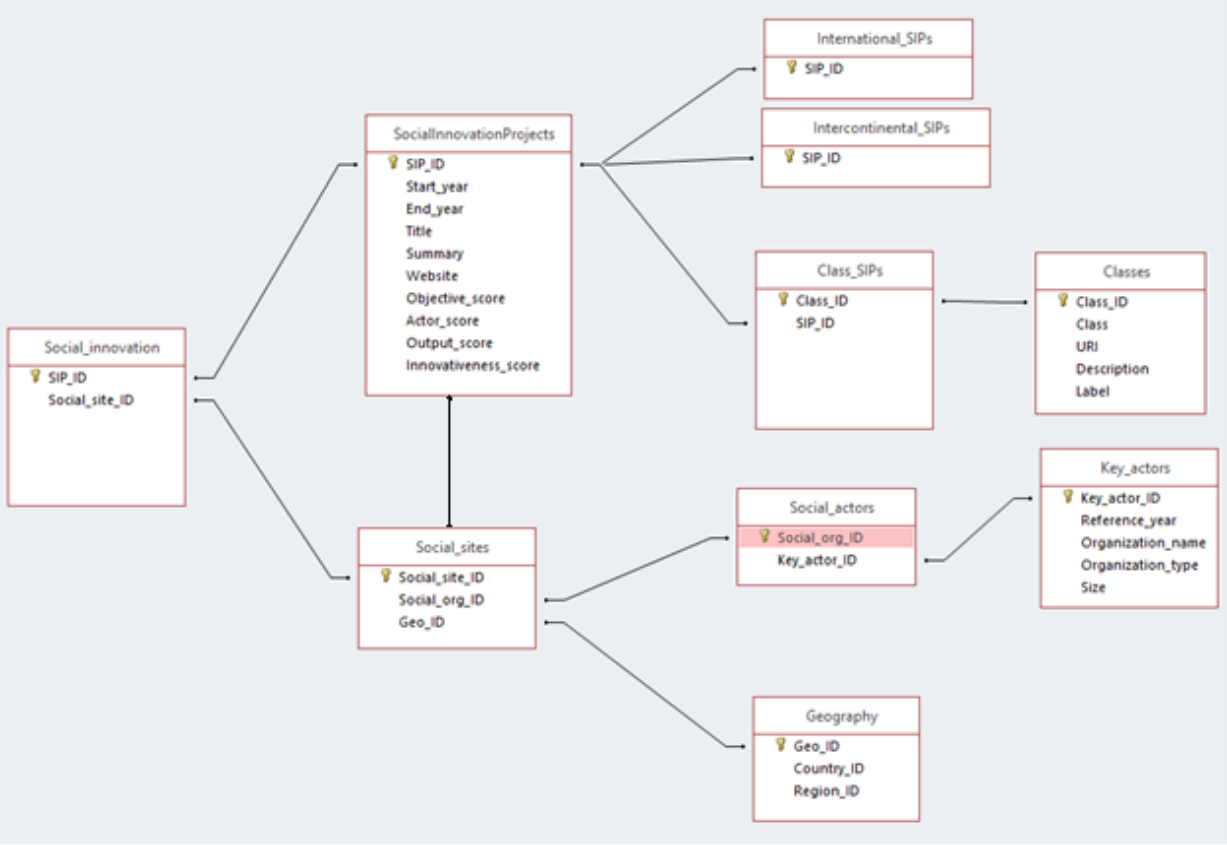


Figure 19: Enriched database structure

The additional project information can also be used to limit volume-based indicators to social innovation projects with desired characteristics, e.g., only those projects with sufficiently high scores for innovativeness.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

In addition to conventional research actors, the KNOWMAK standardized actors' list includes a subset of the social innovation actors identified in the ESID database. Social innovation actors include the following types of actors:

- Non-governmental organizations at the European, national and regional level.
- Public-sector entities, particularly at the city level, like municipalities.
- Grassroot organizations like patients' organizations.

Social innovation actors display, by definition, some level of structure, like having a legal form, some level of continuity over time, some level of visibility (for example having an informative website).

More specifically, three criteria will be adopted to decide which actors will be standardized:

- Stability over time: years of existence.
- Geographical outreach: covering a broader space than a single city/region, with a priority to European-level actors.
- Extent of activity: actors involved in more than one social innovation project and with a lasting engagement in social innovation activities as monitored by ESID.

Social innovation actors are parallel with respect to classical actors, i.e. public sector research and higher education organizations and firms. It is however possible that the latter are also engaged in social innovation projects.

The ESID database contains in total 9577 projects (including some negative sampling data and excluded projects). The relevant data that has been checked and verified include 2688 projects, which form the curated dataset. These projects, as previously mentioned, are marked as "knowmak_ready" in the database. However, not all of the projects are social innovation and some of them are located outside of Europe. However, we have over 2688 social innovation projects (satisfying EU definition) that are located in the European Economic Area and these have been presented in the KNOWMAK tool. This makes ESID database the biggest and most comprehensive database of social innovation built up to date.

KNOWMAK presents social innovation data in two views:

- On a map, there is a list of social innovation projects for each region or country
- A Detailed view is presented when a social innovation project is clicked from the list. This detailed view contains the information about project name, project summary, website, and location.

Examples of data presentation of social innovation projects are presented in Figure 20 and Figure 21.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

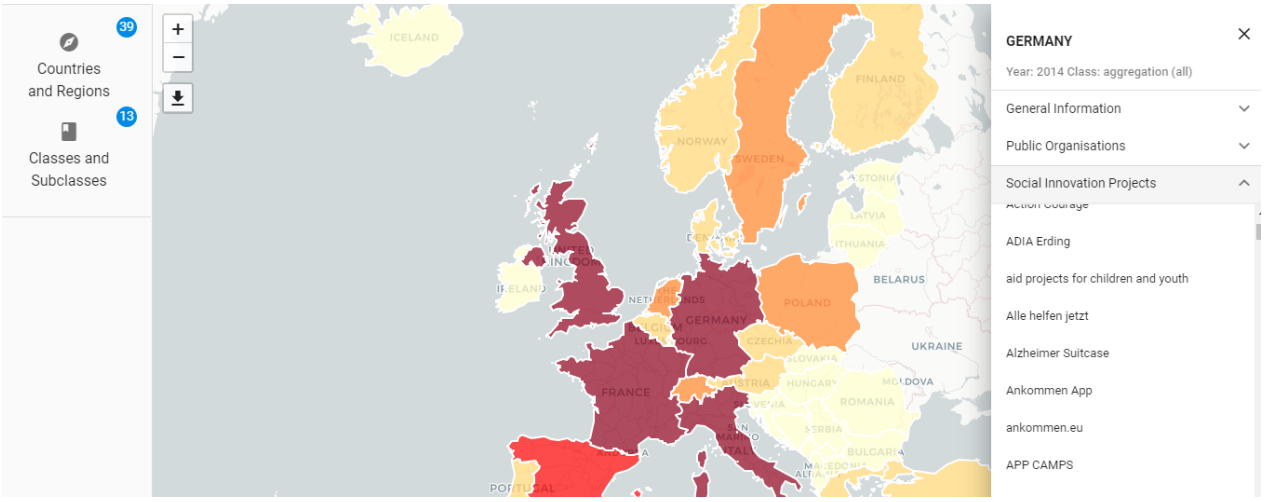


Figure 20: List of social innovation projects presented in KNOWMAK tool (right). Example of Germany

Figure 20 presents KNOWMAK tool's map view. When a user clicks on a certain country (in the example, Germany), a list of social innovation projects can be seen in the right sidebar with other country specific information. If the user clicks on some project from the list, a project detail view will be presented, as is shown in Figure 21.

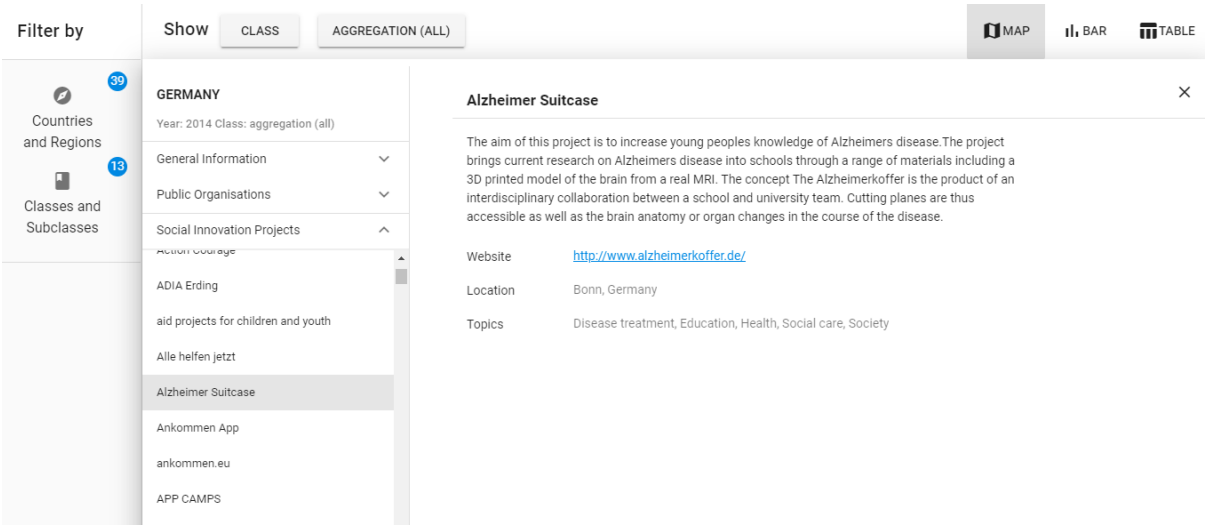


Figure 21: Information about project presented when clicked on the project from the list

3.7.2 Interfaces for access and to other infrastructures

Data was exchanged with KNOWMAK in CSV format. The Database tables for KNOWMAK are presented below. This data was transferred over to the KNOWMAK tool. The ESID data will also be available as SQL dumps and/or CSV files within RISIS.

Social innovation

Variable	Data type	Remarks
SIP_ID	Integer	Project identifier from ESID
Start_year	Integer	
End_year	Integer	
Title	Text	Enriched DB

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Summary	Text	Enriched DB
Website	URL	Enriched DB
Objective_score	Integer	Score of 0 to 3. Enriched DB
Actor_score	Integer	Score of 0 to 3. Enriched DB
Output_score	Integer	Score of 0 to 3. Enriched DB
Innovativeness_score	Integer	Score of 0 to 3. Enriched DB
Latitude	Numeric	Accurate to city level
Longitude	Numeric	Accurate to city level

Table 32: Social Innovation variables and data types exported to KNOWMAK

Topical classes

Variable	**Data type**	**Remarks**
SIP_ID	Integer	
Class	Text	Could instead be URI for class.

Table 33: ESID Topical Classes exported to KNOWMAK

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

References

ADDARII, F. & LIPPARINI, F. 2017. Vision and trends of social innovation for Europe. Technical Report). *Brussels: European Commission.*

CAJAIBA-SANTANA, G. 2014. Social innovation: Moving the field forward. A conceptual framework. *Technological Forecasting and Social Change,* 82**,** 42-51.

CAULIER-GRICE, J., DAVIES, A., PATRICK, R. & NORMAN, W. 2012. Social Innovation Overview: A deliverable of the project: "The theoretical, empirical and policy foundations for building social innovation in Europe" (TEPSIE). Brussels: The Young Foundation.

CHOI, N. & MAJUMDAR, S. 2015. Social innovation: Towards a conceptualisation. *Technology and Innovation for Social Change.* Springer India.

CUNHA, J., BENNEWORTH, P. & OLIVEIRA, P. 2015. Social entrepreneurship and social innovation: A conceptual distinction. *Handbook of Research on Global Competitive Advantage through Innovation and Entrepreneurship.* IGI Global.

DAWSON, P. & DANIEL, L. 2010. Understanding social innovation: a provisional framework. *International Journal of Technology Management,* 51**,** 9-21.

DEVELOP, O. F. E. C.-O. A. 1997. The measurement of scientific and technological activities: Proposed guidelines for collecting and interpreting technological innovation data: Oslo manual. *OECD*.

EDWARDS-SCHACHTER, M. & WALLACE, M. L. 2017. 'Shaken, but not stirred': Sixty years of defining social innovation. *Technological Forecasting and Social Change,* 119**,** 64-79.

ETTORRE, D., BELLANTUONO, N., SCOZZI, B. & PONTRANDOLFO, P. 2014. Towards a new definition of social innovation. *Organizational Innovation and IT Governance in Emerging Economies.* IGI Global.

EUROPEAN COMMISSION 2013. Guide to Social Innovation. Brussels: European Commission.

GODIN, B. 2012. Social Innovation: Utopias of Innovation from c.1830 to the Present. *Project on the Intellectual History of Innovation.*

GRIMM, R., FOX, C., BAINES, S. & ALBERTSON, K. 2013. Social innovation, an answer to contemporary societal challenges? Locating the concept in theory and practice. *Innovation,* 26**,** 436-455.

HALL, M., FRANK, E., HOLMES, G., PFAHRINGER, B., REUTEMANN, P., & WITTEN, I. H. 2009. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P.The WEKA data mining software: an update. *Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H.ACM SIGKDD explorations newsletter***,** 10-18.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

HARRISSON, D. 2013. Social innovation: What is coming apart and what is being rebuilt? *Challenge Social Innovation: Potentials for Business, Social Entrepreneurship, Welfare and Civil Society.* Springer-Verlag Berlin Heidelberg.

JESSOP, B., MOULAERT, F., HULGÅRD, L. & HAMDOUCH, A. 2013. Social innovation research: a new stage in innovation analysis? Cheltenham, UK: 'Edward Elgar Publishing, Inc.'.

MILOSEVIC, N., GOK, A. & NENADIC, G. Classification of Intangible Social Innovation Concepts. 23rd International Conference on Applications of Natural Language to Information Systems (NLDB2018), 2018 Paris, France. Springer, 407-418.

MULGAN, G. 2006. The Process of Social Innovation. *Innovations: Technology, Governance, Globalization,* 1**,** 145-162.

MULGAN, G., TUCKER, S., ALI, R. & SANDERS, B. 2007. Social Innovation: What it is, Why it matters and hoW it can be accelerated. *Skoll centre for social entrepreneurship Working Papers*.

NELSON, R. R. & NELSON, K. 2002. Technology, institutions, and innovation systems. *Research Policy,* 31**,** 265-272.

NELSON, R. R. & SAMPAT, B. N. 2001. Making sense of institutions as a factor shaping economic performance. *Journal of Economic Behavior & Organization,* 44**,** 31-54.

OECD AND EUROSTAT 2005. *Oslo Manual: The measurement of scientific and technological activities, Proposed guidelines for collecting and interpreting innovation data,* Paris, OECD.

PENNINGTON, J., RICHARD SOCHER, AND CHRISTOPHER MANNING 2014. Glove: Global vectors for word representation. *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*.

PHILLIPS, W., LEE, H., GHOBADIAN, A., O'REGAN, N. & JAMES, P. 2015. Social Innovation and Social Entrepreneurship: A Systematic Review. *Group and Organization Management,* 40**,** 428-461.

ROGERS, E. M. 2010. *Diffusion of innovations*, Simon and Schuster.

VAN DER HAVE, R. P. & RUBALCABA, L. 2016. Social innovation research: An emerging area of innovation studies? *Research Policy,* 45**,** 1923-1935.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Appendix I. List of Existing Databases and Data Sources

Name	Type	Data is open	Main Entities	# Entities
InnovAge	Database	Open but not downloadable	Projects	153
MOPACT social innovations	Database	Open but not downloadable	Projects	150
Centre de Recherche sur les Innovations Sociales (CRISES)	Database	Not open	Innovations	>300
SIMRA (Social innovations in marginalised rural areas)	Database	Open but not downloadable	Projects; Organisations	46
Social Innovation Generation (SIG): Social Innovation in Canada database	Database	Open but not downloadable	Organizations	258
Social Innovation Generation (SIG): case studies	Database	Open but not downloadable	Projects; Organizations	10
Seforïs	Database	Not open	Organizations	1000
European Association for Information on Local Development projects	List of projects	Open but not downloadable	Projects	16
European SI competition: semifinalists	List of projects	Open but not downloadable	Projects	30
KENNISLAND projects	List of projects	Open but not downloadable	Projects	35
Berlin SI startups	List of organizations	Open but not downloadable	Organizations	3
Social Innovation and Homelessness	List of organizations	Open and downloadable	Projects; Organisations	30
Social Innovation Europe	Network	Open but not downloadable	Projects; organisations; networks	>3000
Swearer Centre Social Innovation Initiative	News portal	Open but not downloadable	Projects; Organisations	45
BENISI (Building a European Network of Incubators for Social Innovation)	List of projects	Open but not downloadable	Organizations	369
Social Innovation Tournament	List of organizations	Open but not downloadable	Projects	15 yearly
OECD Observatory of Public Sector Innovation	List of projects	Open but not downloadable	Organizations	402
TRANSITION project: social innovation warehouse	List of organizations	Open but not downloadable	Organizations	17
SI DRIVE mapping results	Mapping	Not open	Projects; Organizations	>1000
SIMPACT	Mapping	Not open	Organizations	94
ICT-Enabled Social Innovation	Mapping	Not open	Projects	613 (595 in EU)
LIPSE SI mapping	Mapping	Not open	Projects	245
Digital Social Innovation Database	Mapping	Open and downloadable	Projects; Organisations	1,115 projects; 1,905 Organisations
Latin American SI network: partners	Network	Open but not downloadable	Universities	13

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

ICT for social innovations	Network	Open but not downloadable	Organizations	33
Young Foundation Ventures	List of organizations	Open but not downloadable	Organizations	11
Impact hub	Network	Open but not downloadable	Organizations	86
EUCLID network	List of organizations	Open but not downloadable	Organizations	19
Design for Europe	List of organizations	Open but not downloadable	Organizations	28
SIC partners	List of projects	Open but not downloadable	Organizations	12
Partners of Young Foundation	List of organizations	Open but not downloadable	Organizations	81
British Columbia Partners for Social Impact	List of organizations	Not open	Organizations	?
Partners of European Association for Information on Local Development	List of organizations	Open but not downloadable	Organizations	82
Wollongong SI network	Network	Not open	Organizations	?
Social entreprise UoM	Network	Not open	Organizations	?
Social entreprises ecosystem	Document	Open and downloadable (each country separately)	Organizations	?
Social entreprise UK	Network	Open but not downloadable	Organizations	>300
Center for Social Innovation	University	-	-	?
Social Firms Scotland	Network	Open but not downloadable	Organizations	186
Partners of Social Entreprise Academy	List of organizations	Open but not downloadable	Organizations	12
Social Entreprise Scotland	List of organizations	Not open	Organizations	?
Scottish Council for Voluntary Organisations	List of organizations	Not open	Organizations	1800
Melting Point: Scotland's Centre for Social Innovation and coworking	Network	Not open	Organizations	?
Ashoka Changemaker Campuses	List of universities	Open but not downloadable	Universities and colleges	37
Glasgow Social Entreprise Network	List of organizations	Open but not downloadable	Organizations	120
YY Foundation: list of universities related to social business	List of universities	Open and downloadable	Universities	21
Social Finance database	List of projects	Open but not downloadable	Projects	87
Open Knowledge International	List of projects	Open but not downloadable	Projects	15
Global Innovation Fund projects	List of projects	Open but not downloadable	Projects	27
Bill Melinda Gates Foundation: community grants	List of projects	Open but not downloadable	Organizations	2371
European Volunteering Service organizations	List of organizations	Open but not downloadable	Organizations	5790
Ontario Social Entreprises	List of organizations	Open and downloadable	Organizations	>1000
KEEP EU projects	List of projects	Open and downloadable	Projects	1204
European Social Fund	List of projects	Open but not downloadable	Projects	435

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

United Nations Democracy database	List of projects	Open but not downloadable	Projects	>700
European Investment Bank projects	List of projects	Open but not downloadable	Projects	>1000
Community Indicator Concortium	List of projects	Open but not downloadable	Projects	294
EFESEIIS case studies	List of organizations	Dropbox, EFESEIIS report, Annex 1	Organizations	55
CAPPSI projects	List of projects	Open but not downloadable	Projects	37
MAZI Pilot Studies	List of projects	Open but not downloadable	Projects	4
Making Sense campaigns	List of organizations	Open but not downloadable	Projects	7
MAKE IT case studies	List of organizations	Open and downloadable (text), page 31	Projects and Organizations	10
EMPATIA case studies	List of projects	Open and downloadable (text), page 13	Projects	26
EMPATIA pilots	List of projects	Open but not downloadable	Projects	9
Citizen science projects	List of projects	Open but not downloadable	Projects	9
Consumer-related projects	List of projects	Open but not downloadable	Projects	9
WEBCOSI: civil society initiatives	List of projects	Open and downloadable (text), page 12	Projects	8
P2P Value case studies	List of projects	Open and downloadable (text), page 8	Projects	4
Social Investment case studies	List of projects	Open and downloadable (text), page 6	Projects	20
TRANSIT social innovation initiatives	List of projects	Open but not downloadable	Projects	76
TRANSIT social innovation networks	List of projects	Open but not downloadable	Organizations	20
European Investment Bank 2017 finalists	List of projects	Open and downloadable	Projects	15
European Investment Bank 2016 finalists	List of projects	Open and downloadable	Projects	15
European Investment Bank 2015 finalists	List of projects	Open and downloadable	Projects	16
European Investment Bank 2014 finalists	List of projects	Open and downloadable	Projects	15
European Investment Bank 2013 finalists	List of projects	Open but not downloadable	Projects	16
SI DRIVE case studies: education and lifelong learning	List of projects	Open and downloadable (text), pages 8 - 102	Projects	20
SI DRIVE case studies: employement	List of projects	Open and downloadable (text), pages 9 - 58	Projects	14
SI DRIVE case studies: environment and climate change	List of projects	Open and downloadable (text), pages 15 - 48	Projects	9
SI DRIVE case studies: energy supply	List of projects	Open and downloadable (text), pages 13 - 44	Projects	7

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

SI DRIVE case studies: mobility and transport	List of projects	Open and downloadable (text), pages 21 - 70	Projects	9
SI DRIVE case studies: health and social care	List of projects	Open and downloadable (text), pages 17- 81	Projects	15
SI DRIVE case studies: poverty reduction and sustainable development	List of projects	Open and downloadable (text), pages 24 - 91	Projects	13
ImPRovE cases	List of projects	Open and downloadable (text)	Projects	31
SINGOCOM cases	List of projects	Open but not downloadable	Projects	30
WILCO cases	List of projects	Open and downloadable (text)	Projects	about 70, needs assessment
TRANSITION success cases	List of projects	Open and downloadable (text)	Projects	10
TRANSITION cases	List of projects	Open and downloadable (text)	Projects	17
LIPSE cases	List of projects	Open and downloadable (Appendix 3, page 113)	Projects	15
P2P Value directory	List of organizations	Open and downloadable	Projects and Organizations	383
ICT-enabled social innovation: cases	List of projects	Open and downloadable (text, from page 85)	Projects	132
RegioStars Awards	List of projects	Open but not downloadable	Projects	24
European Social Fund grantees	List of projects	Open but all in separate pdfs	Projects	100 documents
SIC search: projects	News about projects	Open but it is a list	Projects and Organizations	
SIMPACT cases	List of projects	Open, text, Appendix 1 page 132	Organizations	91
ICT-enabled social innovation initiatives	List of projects	Table, page 50	Projects and Organisations	40

Appendix II. Variables and data types in the ESID database

1.1 Projects

Projects table describes project entities in detail.

Variable name	**idProjects**
Variable description	Primary key
Format	int(11)
Phase	Phase 1
Coverage	Mandatory

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Remarks	

Variable name	**ProjectStandardisedID**
Variable description	Standardised id for the project. SI-P+country code+6 digits (e.g. SI-P-FR-000001)
Format	varchar(100). Format: SI-P+country code+6 digits (e.g. SI-P-FR-000001)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**ProjectName**
Variable description	The name of the project
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**PreviousNames**
Variable description	The list of previous names of the project, if they exist
Format	varchar(1000)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**SGC**
Variable description	Societal Grand Challenges that the project addresses
Format	varchar(200)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**TechnologyArea**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	Technology area in which project operates
Format	varchar(200)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	Types to be decided later

Variable name	**KET**
Variable description	Key Enabling Technologies that project is using
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**DateStart**
Variable description	Start date of the project
Format	year
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**DateEnd**
Variable description	End date of the project
Format	year
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**Ongoing**
Variable description	Variable indicating whether the project is ongoing
Format	binary(1)
Phase	Phase 2
Coverage	Missing data is allowed

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Remarks	

Variable name	**ProjectWebpage**
Variable description	The main website of the project.
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**FacebookPage**
Variable description	The main Facebook page of the project, if available
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ProjectTwitter**
Variable description	The main Twitter profile of the project, if available
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ProjectLinkedIn**
Variable description	The LinkedIn page of the project
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**FirstDataSource**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	The name of the data source initial information about project was retrieved from
Format	varchar(200)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**DataSources_idDataSources**
Variable description	Reference to the DataSource in which project was initially identified.
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.2 ProjectLocation

Project location table presents detailed information about the location where the project was executed.

Variable name	**Project_idLocation**
Variable description	Primary key
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Project_LocationType**
Variable description	Type of location
Format	varchar(100). Categorical. Possible values: • Primary • Secondary (to be further detailed at a later stage)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**LocationScope**
Variable description	Scope of the project in term of locations
Format	varchar(200)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Project_Address**
Variable description	Address of the project's contact
Format	varchar(500)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Project_City**
Variable description	City in which project is executed
Format	varchar(500)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Project_Country**
Variable description	Country in which project is executed
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Project_PostCode**
Variable description	Post code of project's contact person/team
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**PhoneContact**
Variable description	Phone number of responsible person for the project
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**EmailContact**
Variable description	Email address of responsible person for the project
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Project_Longitude**
Variable description	Longitude of the location where project is execute
Format	float
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Project_Latitude**
Variable description	Latitude of the location where project is execute
Format	float
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS3**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	The third level of NUTS3-level locational classification by EUROSTAT, related to the location of the project
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS2**
Variable description	The second level of NUTS3-level locational classification by EUROSTAT, related to the location of the project
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS1**
Variable description	The first level of NUTS3-level locational classification by EUROSTAT, related to the location of the project
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**FUA**
Variable description	Functional Urban Area of the project's location
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Projects_idProjects**
Variable description	Reference to the relevant project

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.3 AdditionalProjectData

This table presents additional data about the project that was structured, and was not fitting in defined project related variables. Examples could include domains, or tags that are presented in some data sources or additional web sites or social media profiles (Instagram, Pinterest, etc.)

Variable name	**AdditionalProjectData_id**
Variable description	Primary key, identifier of the additional data
Format	int(11)
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

Variable name	**FieldName**
Variable description	The name of the new variable.
Format	varchar(100)
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

Variable name	**Value**
Variable description	The value of the variable, named in FieldName
Format	text
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	Projects_idProjects
Variable description	Reference to the related project in the Project table
Format	int(11)
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

Variable name	DateObtained
Variable description	Date and time when the information was obtained
Format	datetime
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

Variable name	SourceURL
Variable description	URL of the page from which the variable was extracted from.
Format	varchar(500)
Phase	Phase 1
Coverage	The entry for this table is not mandatory, however, if it exists value for this value is mandatory
Remarks	

1.4 TypeOfSocialInnovatation

TypeOfSocialInnovation stores information on whether the project satisfies social innovation criteria. The output of machine learning classifiers are stored in this database

Variable name	idTypeOfSocialInnotation
Variable description	Primary key
Format	int(11)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**CriterionOutputs**
Variable description	Score for criterion Output
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	Scores: • 3: fully satisfies the meaning of the criteria • 2: partially satisfies • 1: very weakly satisfies 0: no indication at all

Variable name	**CriterionObjectives**
Variable description	Score for criterion Objectives
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	Scores: • 3: fully satisfies the meaning of the criteria • 2: partially satisfies • 1: very weakly satisfies 0: no indication at all

Variable name	**CriterionActors**
Variable description	Score for criterion Actors
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	Scores: • 3: fully satisfies the meaning of the criteria • 2: partially satisfies • 1: very weakly satisfies 0: no indication at all

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**CriterionInnovativeness**
Variable description	Score for criterion Innovativeness
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	Scores: • 3: fully satisfies the meaning of the criteria • 2: partially satisfies • 1: very weakly satisfies 0: no indication at all

Variable name	**Projects_idProjects**
Variable description	Reference to the relevant project
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.5 Projects_Relates_to_Projects

Project_Relates_toProjects describes relationship between projects.

Variable name	**Projects_idProjects**
Variable description	Reference to the first project in the relationship
Format	int(11)
Phase	Phase 1
Coverage	Projects do not need to relate to other projects. However, if they do, this variable is mandatory
Remarks	

Variable name	**Projects_idProjects1**
Variable description	Reference to the second project in the relationship
Format	int(11)
Phase	Phase 1

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Coverage	Projects do not need to relate to other projects. However, if they do, this variable is mandatory
Remarks	

Variable name	**RelationshipType**
Variable description	The type of relationship between the projects (to be finalized later)
Format	varchar(200)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

1.6 Actors

Actor tables describes social innovation actor, with basic details such as its name, size, websites, etc.

Variable name	**idActors**
Variable description	Identifier of the project
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**ActorStandardisedID**
Variable description	Standardised ID as defined by KNOWMAK manual
Format	varchar(100) O+ISO country-code + four digits for OrgReg (O-FR001). F+ISO country-code + six digits for FirmReg (F-FR00001). S+ISO country-code + four digits for social innovation actors (S-FR001).
Phase	Phase 1
Coverage	Mandatory
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**ActorName**
Variable description	The official name of the actor in English or, if not available, in national language.
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Actor_LegalEntityName**
Variable description	Legal entity name as registered in national register of organizations if available
Format	varchar(500)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_Type**
Variable description	The broad identification of the type and subtype of actors.
Format	Literal Type codes: O = public sector organizations. F = firms. S = social innovation actors.
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Actor_Subtype**
Variable description	A more fine-grained delineation of subtypes of actors.
Format	Numeric. Subtype codes. a) For Orgreg: 1 = Higher Education Institutions. 2 = Public Research Organizations 4 = Research Hospital 5 = Public Administration 6 = Private Non Profit b) For firmreg

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

	11 = start-ups 12 = fast-growing mid-size firms c) For social innovation (to be finalized later): 21 = Formal non-governmental organisations 22 = Informal non-governmental organisations
Phase	Phase 1
Coverage	Mandatory
Remarks	Multiple subtypes are allowed.

Variable name	**Actor_Size (Importance)**
Variable description	Classification of the actor by size
Format	varchar(200) Categorical. Possible values (to be finalized later) Could be defined as an importance, based on the actor's visibility, date of foundation, social media presence, activity, size of the website
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_Budget**
Variable description	Yearly budget of the organization if available
Format	varchar(200). Categorical. (to be finalized later)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**SourceOriginallyObtained**
Variable description	The data source from which the entity was initially obtained
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Remarks	

Variable name	**ActorWebsite**
Variable description	The main website of the actor
Format	varchar(300)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ActorFacebookPage**
Variable description	The page on Facebook about the actor
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ActorLinkedInPage**
Variable description	LinkedIn page or profile of the actor's organization, if available.
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ActorTwitterProfile**
Variable description	Twitter profile of the actor, if available
Format	varchar(600)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**ObjectiveStatement**
Variable description	Objective or mission statement of the organization, if available
Format	varchar(1000)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**PreviousName**
Variable description	A list of previous names of the actor, in case it was renamed
Format	varchar(1500)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**StartDate**
Variable description	The date when the actor's organization was created.
Format	date
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**EndDate**
Variable description	Date when actor closed or stopped existing
Format	date
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**StillActive**
Variable description	Binary variable indicating whether the actor is still operating
Format	binary(1)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	DataSources_idDataSources
Variable description	Link to the DataSources table, with additional information about a data source, given actor was initially obtained from
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.7 ActorLocation

ActorLocation provides detailed information about locations of the actor. As there may be multiple locations of the actor, including headquarters and branches, we store all these information. The KNOWMAK database is only interested in headquarters' location. This information is integrated with KNOWMAK based on coordinates and location of the headquarters.

Variable name	Actor_idLocation
Variable description	Primary key
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	Actor_LocationType
Variable description	Type of location that is stored in the entry
Format	varchar(100). Categorical. Possible values (to be finalized later): • Headquarters • Branch
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	Actor_Address
Variable description	Address of the office if available
Format	varchar(500)
Phase	Phase 1

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_City**
Variable description	The name of the city of the actor's location entry.
Format	varchar(500)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_Country**
Variable description	Country of the actor's location
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Actor_PostCode**
Variable description	The post code of the actor's location
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_PhoneContact**
Variable description	Telephone number of the office that is described by the entry, if available
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_EmailContact**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	Email contact address of the office that is described by the entry, if available
Format	varchar(100)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Actor_Longitude**
Variable description	Longitude of the actor's location described in the entry
Format	float
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Actor_Latitude**
Variable description	Latitude of the actor's location described in the entry
Format	float
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS3**
Variable description	The third level of NUTS3-level locational classification by EUROSTAT, related to the location of the actor
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS2**
Variable description	The second level of NUTS3-level locational classification by EUROSTAT, related to the location of the actor

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**NUTS1**
Variable description	The first level of NUTS3-level locational classification by EUROSTAT, related to the location of the actor
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**FUA**
Variable description	Functional Urban Area code of the actor's location
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Actors_idActors**
Variable description	Reference to the relevant actor in Actors table
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.8 LegalEntityRegister

LegalEntityRegister presents variables related to formal legal registration of a given actor in a certain country.

Variable name	**LegalEntityRegister_id**
Variable description	Primary key

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Format	int(11)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**NameOfLegalNameRegistry**
Variable description	Name of the legal name registry agency
Format	varchar(500)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**LegalNameOfEntity**
Variable description	Legal name of the actor, as registred in the legal name registry
Format	varchar(500)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**LegalUniqueID**
Variable description	The unique ID provided by a legal entity name registry
Format	varchar(500)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**RegistryCountry**
Variable description	The country of registraction
Format	varchar(500)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**Actors_idActors**
Variable description	Reference to the relevant actor, for which the registration was performed.
Format	int(11)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

1.9 ActorsAdditionalData

This table presents additional data about the actor that was structured, and was not fitting in defined actor related variables. Examples could include domains, or tags that are presented in some data sources or additional web sites or social media profiles (Instagram, Pinterest, etc.)

Variable name	**idActorsAdditionalData**
Variable description	Primary key, identifier of the additional data
Format	int(11)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**FieldName**
Variable description	The name of the variable.
Format	varchar(150)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**FieldContent**
Variable description	The value of the variable, named in FieldName
Format	text
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**Actors_idActors**
Variable description	Link to the related actor
Format	int(11)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**DateObtained**
Variable description	Date and time when the information was obtained
Format	datetime
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**SourceURL**
Variable description	URL of the page from which information was obtained
Format	varchar(500)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

1.10 OrganisationStructure

Organizational structure is related to project and describes the roles of people or organizations in the project and therefore the organizational (management) structure of the project.

Variable name	**idOrganisationStructure**
Variable description	Primary Key
Format	int(11)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**Type**
Variable description	The type of organizational structure (to be finalized later)
Format	varchar(100)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**Name**
Variable description	Name of the organization or person
Format	varchar(500)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

Variable name	**Role**
Variable description	Role of the organization or person in the given project (to be finalized later)
Format	varchar(500)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	Projects_idProjects
Variable description	Reference to the relevant Project in Projects table
Format	int(11)
Phase	Phase 2
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

1.11 Actors_has_Projects

Actor_has_Projects links Actors with the projects. It also specifies the role of the actor in the project.

Variable name	Actors_idActors
Variable description	Link to the actor
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	Projects_idProjects
Variable description	Link to the project
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	OrganisationRole
Variable description	Role of the actor in the project (to be finalized later)
Format	varchar(200).
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

1.12 Actors_Relates_to_Actors

This table creates links between actors.

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**Actors_idActors**
Variable description	Link to the first actor in the relationship
Format	int(11)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**Actors_idActors1**
Variable description	Link to the second actor in the relationship
Format	int(11)
Phase	Phase 1
Coverage	Entry for this table is not required, however, if it exists this variable is mandatory
Remarks	

Variable name	**RelationshipType**
Variable description	The type of the relationship (to be finalized later)
Format	varchar(100)
Phase	Phase 2
Coverage	Missing data is allowed
Remarks	

1.13 DataFrom

DataFrom table keeps track of the sources from which certain variables were obtained from.

Variable name	**idDataFrom**
Variable description	Primary key.
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**TableName**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	The name of the table in which the new variable's value was obtained
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**EntityId**
Variable description	Reference ID of the entity in to which value was added
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**ColumnName**
Variable description	The name of the variable that was obtained
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**CoulumnValue**
Variable description	The value of the variable
Format	varchar(45)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**DataSourceType**
Variable description	Type of the data source. Can be: • Database • Website • Social media
Format	varchar(100)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**DataSourceName**
Variable description	The name of the Data source
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**DataURI**
Variable description	URL to the page from which the newly added value was obtained from
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

1.14 DataSources

DataSources table provides additional information about data sources from which information was collected.

Variable name	**idDataSources**
Variable description	Primary key
Format	int(11)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**Name**
Variable description	The name of the data source
Format	varchar(200)
Phase	Phase 1
Coverage	Mandatory
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable name	**Type**
Variable description	Type of the data source. Can be: • Database • Web resource • Case study • Social media
Format	varchar(200)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**URL**
Variable description	URL to the data source that is described
Format	varchar(500)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**DataIsOpen**
Variable description	Variable presenting information whether data is open and downloadable.
Format	varchar(200)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**RelatedToEU**
Variable description	Variable presenting information whether the data source is related to European Union or some EU funded project.
Format	varchar(100)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**AssociatedProject**

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Variable description	The name of the Project that generated data source
Format	varchar(200)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**DataDurationStart**
Variable description	Starting data of the data collection process and from when the entities in the data sources are collected.
Format	date
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**DataDurationEnd**
Variable description	End date of the data collection and the date until which the data source has been updating
Format	date
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**Theme**
Variable description	Theme of the data source, if avaialble
Format	varchar(200)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**CountryCoverage**
Variable description	Geographical coverage of the data source
Format	varchar(450)

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**SocialInnovationDef**
Variable description	The definition of social innovation that was used in order to create data source.
Format	text
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

Variable name	**MainEntities**
Variable description	The names of main entities that are collected in the given database.
Format	varchar(200)
Phase	Phase 1
Coverage	Mandatory
Remarks	

Variable name	**InclusionCriteria**
Variable description	Inclusion criteria for including actors and projects in the data source
Format	varchar(1000)
Phase	Phase 1
Coverage	Missing data is allowed
Remarks	

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Appendix III. ESID Data Sources and Number of Projects

DataSource	Projects
MOPACT	140
Simra	9
EUSIC	90
Innovage	153
European Investment Bank Finalists	72
Digital Social Innovation Database	2201
CAPPSI projects	36
TRANSIT social innovation initiatives	73
P2P Value directory	382
SI DRIVE case studies: employment	6
SI DRIVE case studies: environment and climate change	4
SI DRIVE case studies: energy supply	1
SI DRIVE case studies: mobility and transport	1
SI DRIVE case studies: health and social care	6
SI DRIVE case studies: poverty reduction and sustainable development	2
SINGOCOM cases	17
TRANSITION cases	17
LIPSE cases	16
ImPRovE cases	7
WILCO cases	42
TRANSITION success cases	9
ICT-enabled social innovation: cases	114
ICT-enabled social innovation initiatives	39
Social Innovation Generation	6
Kennisland	5
Berlin Startups list	3
BENISI (Building a European Network of Incubators for Social Innovation)	29
TRANSITION project: social innovation warehouse	16
ICT for social innovations	21
Open Knowledge International	14
Global Innovation Fund projects	31
EFESEIIS case studies	52
MAZI Pilot Studies	6
Making Sense campaigns	6
MAKE IT case studies	10
EMPATIA case studies	23
EMPATIA pilots	6
Citizen science projects	9
Consumer-related projects	10
WEBCOSI: civil society initiatives	6

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

P2P Value case studies	4
Social Investment case studies	5
TRANSIT social innovation networks	22
SI DRIVE case studies: education and lifelong learning	14
Manual	25
SI-drive	1004
Impact Hub Stockholm	33
Marias World Foundation	6
http://coeso.org/	7
http://jakodoma.org/	2
http://capitalriego.innova.unia.es	4
http://kebabplus.ch	2
http://e.org.pl/	15
http://futureeverything.org	3
http://torodev.co.ug/	2
http://www.ushahidi.com	1
http://www.alpine-pearls.com	2
https://www.financite.be	1
NULL	0
Kickstarter	4297

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

Appendix IV. ESID Actors by Subtypes

SubType	Number_of_Actors
Non-profit or Social Enterprise	995
Association	14
Company	8
Partnership	2
None	100
Cooperative	7
Multiple	1
NGO	6
Other groups	2
NULL	299
Network	1
NULL	0
Grassroot/Community network	790
Private for-profit business	305
Academia/Research organisation	188
Government/Public Sector	79
For profit	12
For profit, Charity	4
Charity	3
For profit, Foundation	1
Association, Cooperative	2
Public authority	4
Start up	1
Third sector	3
NPO	8
For profit, NPO	2
NPO, For profit	1
Foundation	7
Association, For profit	1
University	4
civil society	658
public	820
private	978
private and public	7
state	16
municipal	5
urban development	1
international agency	1
environment	1
provision of support for services	1
provision of space	1
all	1
sickness fund	4

This project is funded by the European Union under Horizon2020 Research and Innovation Programme Grant Agreement n°824091

islamic microfinance	1
civil society/private sector	1
international bank	1
international organisation	3
individual	3
network of organisations, companies and institutions	1
tripartite network - NGO, ministry and business	1
religious organisation	2
independent charitable foundation	1
clothing retailer	2
asset management and insurance	1
federaciÃ³n	1
funding agency	1
international	1
solar	1