SMARTCHAIN

SMART SOLUTIONS IN SHORT FOOD SUPPLY CHAINS

D6.2 Design of the inventory of innovations and related interactive tools

Work Package 6

IPB

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 773785

Document Identification

Project Acronym	SMARTCHAIN
Project Full Title	Towards Innovation - driven and smart solutions in short food supply chains
Project ID	773785
Starting Date	01.09.2018
Duration	36 months
H2020 Call ID & Topic	SFS-34-2017 - Innovative agri-food chains: unlocking the potential for competitiveness and sustainability
Project Website	http://www.smartchain-h2020.eu/
Project Coordinator	University of Hohenheim (UHOH)
Work Package No. & Title	WP6 Innovation platform
Work Package Leader	ISEKI-Food Association (IFA)
Deliverable No. & Title	D6.2 Design of the inventory of innovations and related interactive tools
Responsible Partner	Institute of Physics Belgrade (IPB)
Author (s)	Dušan Vudragović, Petar Jovanović, Antun Balaž
Review & Edit	Dimitrios Argyropoulos (UHOH)
Type	Report
Dissemination Level	PU – Public
Date	16.01.2019
Version	1.0 Dušan Vudragović (IPB)
	1.1 Danilo Christen (WBF)
	1.2 Dimitrios Argyropoulos (UHOH)
Status	Final

Executive Summary

The focus of this deliverable, as defined in the project's description of action, is to propose an initial architecture of the inventory of innovations and related interactive tools. This is done based on the demands identified during the project preparation stage, the requirements collected from actors and stakeholders at the multi-actor workshops, and currently available technology solutions. The inventory of innovations together with the SMARTCHAIN platform will collect all data generated within the project and create a virtual environment that will support the identification of key parameters that influence sustainable food production and rural development. The contents of this deliverable forms a basis for the development of the SMARTCHAIN platform, the inventory of innovations and related interactive tools.

In this document, we have outlined the main functional requirements of the innovation inventory and designed the top-level architecture diagram, which is comprised of three layers: the front-end, the back-end, and the underlying infrastructure. For the central part of the back-end, technology database and indexing and search, we have identified Elasticsearch as the implementation technology that best fits the system requirements. The document analyzer component will be based around Apache Tika library. The metadata store will be a relational database, based on MySQL or PostgreSQL. The system will integrate all components listed in the initial design of the inventory section and provide REST APIs for the front-end to consume the services. The front-end library for components that face users (technology feed, message board, and quick assessment tool) will be chosen with consideration for the implementation platform used in WP6.1.

After a brief introduction in Section 1, Section 2 of this deliverable gives an overview of the actors' and stakeholders' expectations from the project and their needs. These initial technical (ICT) requirements relevant for the development of the inventory of innovations were gathered during the first set of the multi-actor workshops organized in the nine partner countries. The identified expectations and needs highlight the necessity of an accurate analysis of the state-of-the-art, knowledge transfer and sharing of experiences. An overview of available technologies is presented in Section 3. We have recognized several high-quality open source solutions for implementation of the inventory and related tools. These are presented in the form of libraries, services, and engines, and for each of them, the main advantages and limitations are emphasized. Based on this, in Section 4 we have listed functional requirements and proposed an initial design of the inventory. Section 5 links functional requirements with the most suitable technologies and presents the deliverable conclusions.

Table of Contents

Glossary

API	Application program interface
CSV	Comma separated values
HTML	Hypertext markup language
ICT	Information and communications technology
IP	Intellectual property
OCR	Optical character recognition
PDF	Portable document format
PM	Project month
REST	Representational state transfer
SFSC	Short food supply chain
SQL	Structured query language
TF-IDF	Term frequency - inverse document frequency
WP	Work package

1. Introduction

As it is described in the SMARTCHAIN DoA document [1], the project aims to foster and accelerate the shift towards collaborative short food supply chains (SFSC) and to introduce new robust business models and innovative practical solutions that enhance the competitiveness and sustainability of the European agri-food system. This will be done by the analysis of the technological and non-technological (WP2), social (WP3), consumer (WP4), environmental (WP5), and business and policy (WP7) specific factors related to short food supply chains, which will result in identification of the key parameters that influence sustainable food production and rural development.

Due to the very varied nature and practice of short supply chains, dependencies on different territorial conditions (culture, climate, resources, governing structures, available infrastructure, market, etc.), the consortium will primarily focus on 18 preselected case studies, existing short food supply chains, from 9 countries (2 case studies per country). Up to this point (PM03), we have successfully established 9 innovation/collaboration hubs and stakeholder groups including further short food supply chain mapping in these countries (Deliverable 1.1) that will be integrated into the virtual innovation platform. Each of them is coordinated by the corresponding hub manager, which is responsible for the facilitation of data acquisition and information flow within case studies and project's actors.

Analysis of the case studies will recognize innovative and practical solutions relevant to short food supply chain scale up and place them within a virtual environment that is under development within the WP6 as the SMARTCHAIN innovation platform. The platform will enable knowledge transfer, innovation, and cooperation between all the stakeholders of the studied short food supply chains. It will allow storing, generating, sharing and utilizing information on innovations, as well as facilitation of communication between the innovation hubs. The front-end of the SMARTCHAIN innovation platform is an interactive online portal oriented towards all the stakeholders and actors, and its back-end is the inventory (database) of available innovations, solutions, and recommendations. The development of the front-end (interactive platform) will be done as a part of task WP6.1, while we are developing its back-end (inventory), which is focus of this deliverable, within the task WP6.2.

The development within the WP6 relies on the demands identified during the project preparation stage, requirements collected from actors and stakeholders at the multi-actor workshops, and currently available technology solutions. Therefore, in the User requirements section (Section 2) of this document we present outcomes of feedbacks analysis from the multi-actor workshops organized within the WP1. In the Overview of available technologies and solutions section (Section 3), we give a brief overview of available technologies and solutions relevant to project needs, and based on this, in the Initial design of the inventory section (Section 4) we propose a design of the project's inventory of innovations and related interactive tools.

2. Overview of user requirements

During the multi-actor workshops conducted in the nine partner countries and coordinated by the WP1 team, the consortium asked participants (actors and stakeholders) about expectations from the project and their needs. Following a predefined questionnaire, in discussion with participants, hub managers identified and collected in total 109 feedbacks about expectations and 68 feedbacks on user needs. The deliverables D1.2 [2] and D1.3 [3], give details about those expectations and needs, while here we list identified technical (ICT) requirements relevant for the development of the SMARTCHAIN innovation platform.

Participants' expectations are grouped into 20 main categories, which allows us to identify three most important ones: analysis of the current situation, knowledge of actors and stakeholders about the short food supply chains, and feedback and sharing of experiences. **Error! Reference source not found.** lists concrete expectations and their frequencies per each category. These expectations highlight the necessity of an accurate analysis of the state-of-the-art, as well as a need for the knowledge transfer and sharing of experiences.

Table 1. Multi-actor workshop participants' expectations and their frequencies per each category

Categories	Frequency	Description of Expectations
Feedback and sharing of experiences	10	Get feedback from actors working with high-end products (truffles for example)
		Be able to have elements of comparisons with other structures
		Experience share to get inspired
		Receiving information on other experiences of short food supply chains
		Sharing experiences.
		Sharing experiences about short supply chain, with detailed information and data, especially with reference to social innovation in order to identify a short supply chain model to be compared with others in the Mediterranean region.
		Increase participation on a daily basis for sharing ideas etc. and for financially supporting ALLOTROPON
		Collaboration of short food supply chains, mapping of short food supply chains, better communication with short food supply chains from Europe, technological improvement of production process and better financial status, more fairs or other manifestations which gathering short food supply chains
		Get feedback from actors working with high-end products (truffles for example)
		Creating the database and opportunity for exchange of practices with cases form Western Europe will have strong impact for Serbian short food supply chain market.
Knowledge of actors and stakeholders on short food supply chains	12	Improvement of knowledge of stakeholders involved in short food supply chains
		Knowledge of different models of short food supply chains
		Make information easily available and comprehensible
		Improve the knowledge of different short food supply chain models
		Production of knowledge and innovative solutions so that can be more accessible to the farmers, food producers, consumers and other stakeholders involved in short food supply chains.
		To obtain more knowledge about short food supply chains as well as extended short food supply chains, and possibilities of utilizing this model in distribution of local developed food products
		What are the definitions of short food supply chain?
		Definitions of goals > where place review(s)
		From Serbian perspective it would be good to have an analysis on individual level of the owners of the enterprises in the cases. Analysis should include the better understanding of the individual factors and motivators for the owners of the producing companies

		Necessary knowledge when starting or further developing short food supply chains
		Development of official definition of short food supply chain, and recognition of short food supply chains into multiple policy areas
		Find objective criteria about what should the short food chains answer
Analysis of current situation	12	Linking to practice taking time to look at the situation in Swiss Romandie and at national level
		Customer journey per type of short chain company
		Linking to existing innovation support in Switzerland (InnoSuisse, Innovaud, Fri-up), PHR project "Local consumption in the Lemanic city" and PA ZZ+
		Categorizing 3CF's, tailored (size/type) approach
		Analysis of the current situation of short supply chains and structured exchange of information flow in order to qualify the work and the role on the market, underlying their added value in terms of quality (better quality of products) environmental (impacts) social (better quality) of life)
		Diagnosis of competition (with existing companies)
		Turnover
		Costs
		Farmer's share
		Operational/set-up
		Labor
		Recognition of organizational patterns for short food supply chains

Participants' needs are grouped into 17 main categories. After our analysis, seven of these are considered as most important: diffusion and awareness, education of stakeholders, feedback, methods and data collection, networking, policy support, and short food supply chain dynamics. **Error! Reference source not found.** lists concrete needs and their frequencies, as mentioned by the participants. Majority of them are linked to knowledge transfer, sharing, and networking.

Table 2. Multi-actor workshops participants needs and their frequencies per each category

Categories	**Frequency**	**Description of the needs**
Diffusion and awareness	7	How to reach the farmers with the project results?
		Private and research institutions or MMSs initiatives that promote and sustain an eco-ideology at the local level. Diffusion of knowledge to farmers, consumers and citizens in general
		Consumers' and citizens' awareness
		Improve organic production, implementation of processing technology, increase consumer awareness on the quality of organic products
		Establishing a higher degree of confidence to the domestic products
		Support the spreading of information on practical short food supply chain implementation through the EU funded projects
		Increase awareness of consumers on the quality of short food supply chain products
Education of stakeholders	1	Better education of short food supply chain stakeholders (technological, non-technological and social innovations)
Feedback	4	Getting feedback after every participatory workshop or Work Package in order to have an overview and understanding of the project and not only a national
		Taking lessons from PHR project "Local consumption in the Lemanic city" to provide insights for other case studies
		Experience related to rules of hygiene, food health, standards, taxation, certification, trading, which is often tailored to industrial companies

		Support structure for information exchange among and about short food supply chains
Methods and data collection	7	Tool for data collection (consumers' survey)
		Tool for energy and ecological balance analysis
		Validation and veracity of the data is a challenge
		Collection of data for short food supply chain analysis
		Comparable data collected through simple methodologies that are comparable to all categories of actors involved
		Data that represent in the most exhaustive way possible from a technical, social and economic point of view, short supply chain experience
		Providing support, tools and reflections that can be used by farmers, entrepreneurs, project leaders
Networking	3	Enhance marketing and advertising mechanisms in order to expand the network between producers and consumers
		Identification and networking of short food supply chains
		Better local and regional networking among all stakeholders
Policy support	10	To be better represented with policy makers
		More clarity in terms of available funding
		Info about legal obstacles
		Funding
		Change subsidy structures
		Change ratio of price/sustainability
		Better policy framework for short food supply chains
		Recommendations on possible policy tools to support short food supply chains
		Better delivered and policy recommendations for short food supply chains
		Better policy framework for short food supply chains
Short food supply chain dynamics	12	Understand the classical bottlenecks
		How to attract young workers
		Do the short food chains bring value (which value?) to the consumers and the producers?
		Great interest in social innovation, therefore need to understand strategies, models and dynamics in short food supply chains regarding SI
		Understanding new short food supply chain models
		To rely on a most complete as possible set of information about the short food supply chain reality, with particular refer to the data fluxes and relations among actors
		Visibility
		Information and information flux
		Transparency
		Monitoring system
		Clear definitions
		Practical implementation of short food supply chains - organization, marketing, promotional activities, overcoming administrative obstacles, strategic planning, modernization of both production and sales capacities of the manufacturers

The development and maintenance of the SMARTCHAIN interactive innovation platform will follow the above listed expectations and needs. The architecture, data structure and the content of the platform will achieve those expectations and needs and will be validated by end-users during the project lifetime. The platform is designed to share information and experiences, difficulties and benefits, to help and inspire other chains, and to figure legacy data. In addition to this, it will provide search and blog functionalities.

The first development phase of the innovation platform will aim to support mostly text-based data, such as text documents (e.g., scientific papers, patents, case studies, product descriptions, farm descriptions, and technology descriptions), scanned paper documents, presentations, and spreadsheets. The second phase will cover support for other data types, such as data from web or electronic sources, audio/video data, geospatial data (maps, satellite images), numerical data (statistics, sensor readings), etc. The volume of initial data is expected to be low, considering most of it will come from documents (e.g., PDF). All collected data will be available for download, and it will be continuously collected during the project lifetime through the technology feed component of the architecture.

Concerns around privacy and anonymization of published data will be left for relevant actors to decide on (WP10 Ethics). However, the SMARTCHAIN interactive platform will allow the restriction to project members only, and actors will have control over what information gets published under the restrictions and will be able to change these at later times. An added value of the project is the involvement of many stakeholders that will be interested in accessing the information on the platform. Therefore, licenses for the stakeholders will be integrated within the platform as well.

Beside a free-form text search, the system will support a parameterized search in order to achieve better accuracy. The filters will be value-restricted on specific fields and properties of the documents. The following filters are proposed by actors and stakeholders:

- Language;

- Short food supply chain type:
    - On farm sales (such as farm shops or pick-your-own schemes);
    - Off farm sales (such as farmers' markets, vending machines, sales to hospitals, schools);
    - Farm direct deliveries (delivery schemes, internet sales, specialty retailers);
    - Community-Supported Agriculture.

- Innovation types according to the project WPs:
    - Business;
    - Technological innovation (product, service, process);
    - Non-technological innovation (marketing, organization);
    - Social innovation;
    - Consumers;
    - Environmental impact;
    - Policy requirements.

3. Overview of available technologies

At this project phase, we expect a lot of unstructured information to be stored within the SMARTCHAIN inventory system. To enable an efficient search for large quantities of unstructured data, the system needs to build a data structure called index. The index holds searchable information extracted from documents and is organized to support full-text search on every available piece of information. The process of creating such a data structure is known as indexing. It goes through all data records, extracts a list of terms which appear in them, and for each term makes a note of which records hold it. Such approach enables quick lookup of records which hold the terms of a search query.

Before the indexing can take place, the document analysis step must extract all the data that will be searchable. The extraction transforms raw information into a suitable format (usually a semi-structured document). This processing step is heavily dependent on the type of raw data being ingested. For example, the extraction will differ significantly in the case of HTML pages and scanned documents. While the former is directly parse-able, the latter has to go through an image processing step and optical character recognition (OCR) to extract textual content. In addition to this, in the case of external data sources, an additional crawling process has to pull and inject such data into the system, making them available to a document analyzer.

The search processing consists of two components. First, the search query that the user has entered in natural language or via an optional formal syntax is translated into the actual set of terms and constraints that will be looked for in the index. This process can include word stemming, lemmatization, removal of stop words (such as "the"), and other language-specific transformations to make the query more flexible for matching in the index. The second component of the search is assigning the rank of relevance to each result retrieved. This can be controlled by tweaking the weights of fields in the index records. The relevance is calculated also using the predetermined metrics, such as TF-IDF (term frequency - inverse document frequency). This component of the system can support search suggestions as well.

There are several high-quality open source solutions for implementation of the search facility available today in the form of libraries, services, and engines. Libraries provide the greatest flexibility to developers, giving them just the parts needed to implement the entire system from scratch. Services make libraries more encapsulated and easier to use and scale, while still giving enough flexibility to design the rest of the system that uses the search service. Engines are complete solutions for document management, whose flexibility is limited by their established design and implementation. They usually support limited customization through configuration or extensions of some specific components.

In this section, we present frequently used open source libraries, services, and engines relevant to the development of the inventory of innovations and related interactive tools. For each of them, the main advantages and limitations are emphasized.

3.1 Libraries

There are several open source libraries for efficient search of large quantities of unstructured data. By far the most developed and popular library is Apache Lucene [4]. The library is implemented in Java programming language and forms the base of many other systems, such as Elasticsearch [4], Solr [6], etc. The other popular search library is Xapian [7], natively implemented in C++ programming language and thus able to work without Java platform, usually with smaller memory footprint. Comparing these two libraries, Lucene has the following advantages:

- It is more popular and therefore better supported;
- It has a reasonably complete and tested set of features for implementing search engines at any scale;

- There is an abundance of documentation and a large community to ask for help.

Xapian, on the other hand, offers a reasonably complete feature set and has no dependency on Java. A disadvantage of Xapian is that it is less popular than Lucene, and thus has sparser documentation and smaller community.

3.2 Services

Direct usage of search libraries, such as Lucene or Xapian, is quite an involved process. Based on our experience, many parts of the system need to be developed and configured to get to a usable state where it can index data and respond to search queries. Also, Lucene's dependency on Java imposes that implementation platform for the rest of the system. To avoid these constraints, services that encapsulate these libraries can be used instead.

The services such as Elasticsearch and Solr run as a server with an API that the rest of the system can use, isolated from the worries of how to configure and run a Lucene-based search efficiently. They can automatically handle scaling so that, if the data magnitude or traffic exceed capabilities of a single server, a search can be distributed among many machines.

Solr and Elasticsearch provide a similar set of features, and both have extensive documentation, but Elasticsearch has had more focus on scalability throughout its development. They provide a REST API and handle search, indexing, and crawling processes. In indexing, a data schema can be configured, and crawling can be extended to extract information from new types of input documents.

Sphinx is another search service, which mimics the MySQL API [8] with query language syntax extensions. It also has quite an active community, but it is the newest among the compared services and is less thoroughly covered by the documentation. Its advantages over Solr and Elasticsearch are that it does not depend on Java programming language and that its query syntax and API closely mimic MySQL, so the existing database drivers can be used to perform searches on it without additional development. This can be a plus for existing projects and teams that already have tight integration and experience with MySQL database.

3.3 Engines

Ambar [9] is an open source document search engine with automated crawling, OCR, tagging and instant full-text search. It is based on Elasticsearch and supports upload and search of documents in various formats including scanned documents (which are processed by OCR). The advantages of Ambar are that it comes in an easily deployable package (docker image [10]) and that it covers most use cases for a document management system. However, it does not provide a recommendation engine and building on top of it would require familiarity with its extensive mix of implementation platforms and dependencies. Another option would be OpenSemanticSearch [11], which is even more feature-packed engine than Ambar. Its advantage is an integrated recommendation facility.

The major downside of both of these engines is that their many components are often implemented in completely different languages and platforms. For instance, Ambar's web UI relies on JavaScript and Node.js, a language processing pipeline is written in Python, information extraction of the crawl phase relies on Java (Apache Tica library), data is stored in MongoDB, Redis is used for cashing purposes, and RabbitMQ manages operations in the system. The situation is similar with OpenSemanticSearch. This complicates the process of introducing new features or modifying existing ones because developers need to be familiar with many technology stacks and the elaborate system architecture.

4. Initial design of the inventory

SMARTCHAIN innovation inventory is designed as a document organization and retrieval system which supports quick finding and discovery of information related to short food supply chains. Through it, the users will be able to upload, share and discover innovations, patents, IPs, and other material related to food supply chains. The target user group are farmers and agricultural organizations looking to optimize their operations, as well as innovation donors, i.e., researchers, technology providers, etc., who wish to raise the visibility of their innovations within a highly interested audience.

Main user-facing components of the system will be:

- Quick assessment component for searching innovations/solutions based on a free-form or parametrized search as well as their user profile and preferences;
- Component for the rating of innovations/solutions by the users, which will power the recommendation engine;
- Component for upload of the innovations/solutions by the innovation donors;
- Crawler which will handle import of the documents and other submitted data into the system, and also collect certain information automatically;
- Message (blog-like) board which will enable farmers, technology providers, researchers, entrepreneurs, SMEs, consumers, and others to ask questions about innovations/solutions, search partners for collaborations, announce news, events, etc.

4.1 Functional requirements

The system should support free-form and parameterized search through the database of documents, including the discovery of relevant information using personalized recommendations. Document formats will vary from text documents such as Word, PDF, plain text, to tabular formats such as spreadsheets, CSV, databases, to image data such as document scans. Innovation donors should be able to upload their documents (innovations) and provide metadata using a specialized web form for that purpose. Metadata should help to better categorize documents uploaded into the system. Also, the users browsing the system will be able to enter more information about their interests to their profiles, so the search and recommendations can be better tailored to their needs.

A rating system will also give users the ability to rate the documents system has found for them. This will include a rating of training activities and materials in the project (task WP6.3). These ratings will be fed as another source of data into the recommendation engine and enable higher quality recommendations through collaborative filtering. Another type of user-contributed data will be a definition of data sources, such as news feeds, from which to automatically scrape new data as they arrive and to import them into the system. The system will also include a message board where users will be able to start discussions regarding short food supply chains.

One non-functional design goal is to reduce the complexity of platform requirements as much as possible. Similar systems are usually implemented by loosely coupled services, each implemented in a different language on a different platform with different dependencies. The design of implementation platform for SMARTCHAIN is made with ease of deployment and maintainability in mind.

4.2 System Architecture

To support the functional requirements, derived by analysis of demands identified during the project preparation stage and collected from actors and stakeholders at the multi-actor workshops, which can be implemented using currently available technology solutions, the system is decomposed into the following components:

- Technology feed;
- Quick assessment tool;
- Message board;
- Computing and storage;
- Technology database;
- Metadata storage;
- Indexing and search engine;
- Document analyzer.

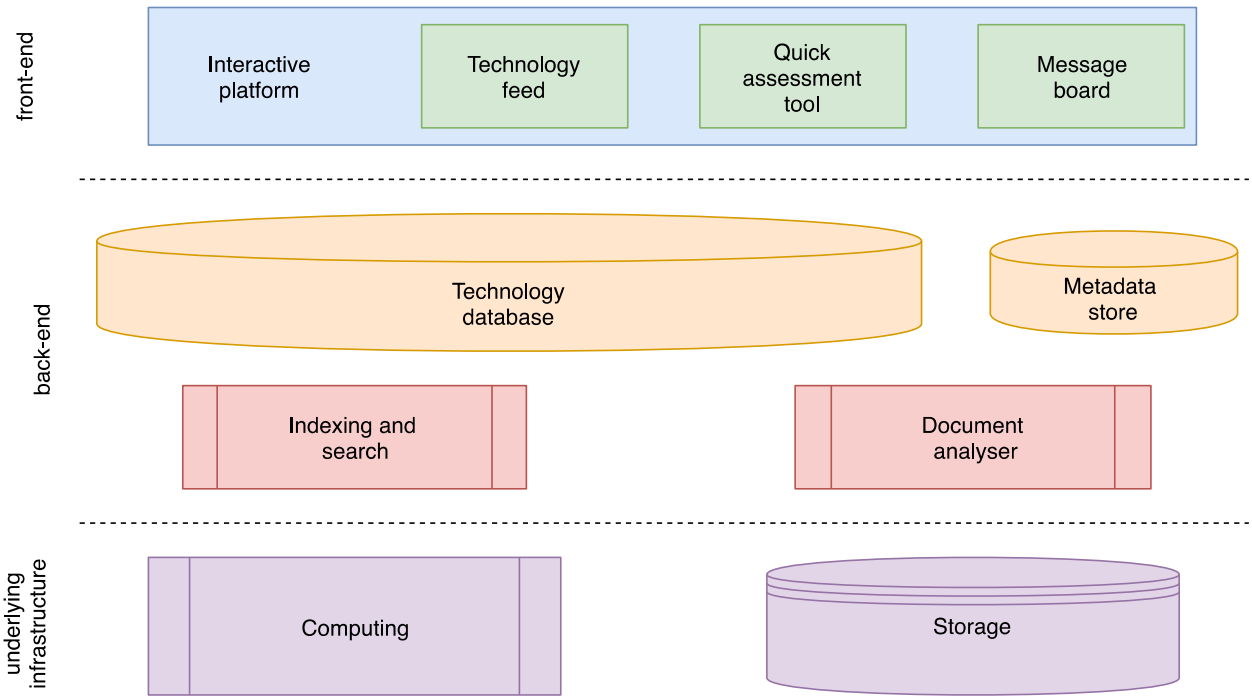

Figure 1. High-level architecture of the SMARTCHAIN inventory

The SMARTCHAIN interactive platform is the user-facing web interface whose design is presented in D6.1 [12], and which will be developed as a part of task WP6.1. In collaboration with WP6.1, to ensure technical compatibility, the technology feed and quick assessment tool will be developed as web-based forms incorporated into the platform. The quick assessment will present users with an interface to describe their interests, while the technology feed is going to handle upload of innovative solutions and related metadata. The message board component will support user questions and announcements.

The technology database will store the documents uploaded into the system, as well as the search index. The metadata store will contain additional data associated with the documents, users and other system entities that support the functioning of other parts of the system, such as the recommendation engine. The indexing and search component will handle building and updating of the index, processing of search queries, and provide additional support to the recommendation engine. The document analyzer will preprocess uploaded documents as needed to import them into the technology database component. This will also include the crawler for automated information extraction from certain data sources.

The underlying infrastructure layer will be based on the computing and storage capacity of the PARADOX cluster [13] at the Institute of Physics Belgrade, which will host the innovation inventory system.

The majority of the components defined in this document will form the backend for the interactive platform. All components will provide REST APIs for interaction with the front-end system.

4.2.1 Front-end components

The main source of data in the innovation inventory system is expected to be the upload of innovation donors. The upload will be structured and handled through a web form defined by the technology feed component. The form will also support the entry of additional data, related to innovation/solution documents and descriptions being uploaded. Such metadata could include geographical location, technology readiness level, potential customers, patent information, etc., which will be used to better gauge the relevance of the innovation to various search queries and users. The recommendation system will use this data in making its prediction of what is relevant to which users.

The message board is a user-facing component which will support posting of questions regarding innovations, searches for partner collaborations, announcements of news and events, and any other topic related to knowledge transfer and innovation in the short food supply chains. System operators will also take active part in motivating local users to raise questions on encountered problems and identified needs.

4.2.2 Back-end components

The technology database component will contain the main data artefacts of the system. Its document storage will accept user uploads. The index will be stored here as well, and this component will be in charge of providing data access to users and system scaling, when necessary. The uploaded artefacts are to be stored in original form, while the index will be built out of semi-structured text data extracted from the documents. The index will be based on reverse index data structure. The metadata storage will contain additional data provided with the uploaded documents, user profile data, as well as user-provided rating data. This metadata will be stored in a relational database whose schema will be defined taking into consideration the needs of the WP6.1 platform.

The creation of index, within the indexing and search component, will be done periodically and incrementally as new documents become available to the system. The search will take into account two primary elements: a search query and available recommendations. In order to support both free-form and parameterized queries, query processing will be performed prior to searching the index with the given query. This will involve word lemmatization, stemming, and further syntax processing that will improve the generality of the query and enable the index to return results with a more flexible match to the query. Recommendations will come from two sources. The first one will be based on the underlying search engine, which will make suggestions based on similarities with user's queries as well as related queries. The second type of suggestions will come from the collaborative filtering approach, based on the ratings data that users have provides on specific documents.

Before injection of the raw documents into the search index, they will be transformed into a suitable semi-structured text-based format. The extraction of these text descriptors from documents of various types will be handled by the document analyzer component. The system will support the most common types of documents, ranging from text-based, such as Microsoft Word, plain text, PDF, etc., to image scans of printed documents. The analysis of documents will be performed on a type by type basis, using Apache Tika library [14]. Beside extraction and text processing, this component will include a crawler sub-component that will be able to automatically pull information from external data sources for further injection into the index. As the data sources will be specified by the system operators, the crawler will not need to discover data sources on its own. Depending on the format in which a data source publishes information, specific agents may have to be developed to extract textual content, preparing them for further analysis. The document analyzer will operate semi-autonomously, based on predetermined configuration and inputs by the system operators. It will run on a schedule or in response to new uploads.

All back-end components will provide REST interfaces to be usable by other parts of the system. The specification of the REST APIs will be made in collaboration with the task WP6.1.

4.2.3 Underlying infrastructure

The PARADOX cluster at the Institute of Physics Belgrade with its 106 compute nodes, equipped with NVIDIA Tesla M2090 GPU cards will host production and testing instances of the inventory, provide computing capacity required by the document analyzer component, and store all data collected by the project. The recently upgraded storage system has the total capacity of 400 TB, and a 10 Gbps Internet connection.

5. Conclusions

The SMARTCHAIN innovation platform, together with the inventory of new innovative solutions and interactive tools, is based on available technologies and designed to facilitate the project needs. It is planned as a document organization and retrieval system that supports quick finding and discovery of solutions and information relevant to short food supply chains.

In this document, we have outlined the main functional requirements of the innovation inventory and designed the top-level architecture diagram, which is comprised of three layers: the front-end, the back-end, and the underlying infrastructure. For the central part of the back-end, technology database and indexing and search, we have identified Elasticsearch as the implementation technology that best fits the system requirements. It is capable of providing a mature platform for the index, search, query analysis, and, to some degree, one of the back-ends for the recommendation engine. The document analyzer component will be based around Apache Tika library that is becoming a standard for text extraction from various types of documents and has the broadest support and largest user community among similar solutions. The metadata store will be a relational database, based on MySQL or PostgreSQL [15]. The database schema will be designed and later refined as the implementation phase progresses.

The system will integrate all components listed in the initial design of the inventory section and provide REST APIs for the front-end to consume the services. The front-end library for components that face users (technology feed, message board, and quick assessment tool) will be chosen with consideration for the implementation platform used in WP6.1.

References

[1] Project SMARTCHAIN 773785 – Annex I - Description of the action

[2] SMARTCHAIN D1.2 – Report listing the data to be collected

[3] SMARTCHAIN D1.3 – Report on methods and tools to be used to collect data

[4] O. Gospodnetic, E. Hatcher, Lucene in Action, Manning Publications Co. Greenwich (2010)

[5] Official Elasticsearch webpage, https://www.elastic.co/

[6] L. Lambert, Apache Solr Essentials, CreateSpace Independent Publishing Platform (2016)

[7] Xapian project website, https://xapian.org/

[8] B. Schwartz, P. Zaitsev, V. Tkachenko, High Performance MySQL: Optimization, Backups, and Replication, O'Reilly Media (2012)

[9] Ambar document search engine webpage, https://ambar.cloud/

[10] J. Turnbull, The Docker Book: Containerization is the new virtualization, Docker Community Edition 18.09

[11] Open Semantic Search wegpage, https://www.opensemanticsearch.org/

[12] SMARTCHAIN D6.1 – Design and building of SMARTCHAIN interactive platform

[13] PARADOX cluster user guide, http://www.scl.rs/PARADOXClusterUserGuide/

[14] Apache Tika - a content analysis toolkit, https://tika.apache.org/

[15] C. Chauhan, D. Kumar, PostgreSQL High Performance Cookbook, Packt Publishing (2017)