



FNS – Cloud

Food Nutrition Security

Food Nutrition Security Cloud

Deliverable 3.5

Extended FNS Services integrated with the FNS-Cloud

Due Date:	31.07.2023
Submission Date:	31.07.2023
Dissemination Level:	Public
Lead beneficiary:	Scalefocus
Main contact:	Daniela Yaneva (daniela.yaneva@scalefocus.com)

Project acronym: FNS-Cloud	Project Number: 863059
Start date of project: 01.10.2019	Project duration: October 2019 – September 2023



Food Nutrition Security Cloud (FNS-Cloud) has received funding from the European Union's Horizon 2020 Research and Innovation programme (H2020-EU.3.2.2.3. – A sustainable and competitive agri-food industry) under Grant Agreement No. 863059 – www.fns-cloud.eu

Document Control Information			
Title	<i>D3.5 Extended FNS Services integrated with the FNS-Cloud</i>		
Editor	<i>Ina Popova (SF), Daria Aladzhem (SF)</i>		
Contributors	<i>Anton Bivolarov (SF), Panče Panov (JSI)</i>		
Reviewer/s	<i>Barbara Korousic Seljak (JSI), Paul Finglas (QIB), Javier de la Cueva</i>		
Dissemination Level	<input checked="" type="checkbox"/> PU - Public		
Approved by	<input checked="" type="checkbox"/> RTDS (COO) <input checked="" type="checkbox"/> QIB (SCO) <input checked="" type="checkbox"/> JSI <input checked="" type="checkbox"/> UCD <input checked="" type="checkbox"/> PMT <input checked="" type="checkbox"/> JDLC <input checked="" type="checkbox"/> EuroFIR <input checked="" type="checkbox"/> UWTSO <input checked="" type="checkbox"/> DTU <input checked="" type="checkbox"/> ENEA <input checked="" type="checkbox"/> HYVE <input checked="" type="checkbox"/> HYLO	<input checked="" type="checkbox"/> UM <input checked="" type="checkbox"/> NUTRIS <input checked="" type="checkbox"/> RIVM <input checked="" type="checkbox"/> WUR <input checked="" type="checkbox"/> UGent <input checked="" type="checkbox"/> IMDEA <input checked="" type="checkbox"/> HUA <input checked="" type="checkbox"/> TUM <input checked="" type="checkbox"/> GS1 <input checked="" type="checkbox"/> SF <input checked="" type="checkbox"/> UoR <input checked="" type="checkbox"/> IFA	<input checked="" type="checkbox"/> ILSI <input checked="" type="checkbox"/> BfR <input checked="" type="checkbox"/> AUTH <input checked="" type="checkbox"/> FEM <input checked="" type="checkbox"/> CNR <input checked="" type="checkbox"/> APRE <input checked="" type="checkbox"/> CAP <input checked="" type="checkbox"/> UNIFI <input checked="" type="checkbox"/> LIFE <input checked="" type="checkbox"/> Nutritics <input checked="" type="checkbox"/> EFFoST
Relevant IPRs	<i>Not applicable</i>		
Underlying Datasets	<i>Not applicable</i>		

Version/Date	Change/Comment
<i>V0.1 2021-05-15</i>	<i>Draft outline prepared by SF</i>
<i>V0.2 2023-05-17</i>	<i>Version 1</i>
<i>V0.3 2023-07-06</i>	<i>Version 2</i>
<i>V0.4 2023-07-28</i>	<i>Version 3</i>
<i>C1.0 2023-07-31</i>	<i>Final Version</i>

Table of Contents

1	Summary.....	3
2	Introduction	4
2.1	Background	4
2.2	Objectives	4
2.3	Target audience	4
3	FNS Harmony Ontology	5
3.1	Background	5
3.2	Ontology Design	5
3.3	Ontology Structure	6
3.4	Ontology Versions and Implementation	10
3.5	Availability	10
4	FNS Services	11
4.1	Background	11
4.2	Related Work	11
4.3	Introduction to the FNS Services and the Back-end Infrastructure	12
4.3.1	FNS-Cloud Back-End Architecture	12
4.3.2	FNS-Cloud Metadata API.....	14
4.3.3	FNS-Cloud Services.....	15
5	Integration of the implemented FNS Services	18
5.1	Background	18
5.2	Extended API Specification	18
5.3	Integration within FNS-Cloud	18
5.3.1	FNS-Cloud Dataset Assessment Tool.....	19
5.3.2	FNS-Cloud Metadata Submission Form.....	20
5.3.3	Metadata enrichment tools	21
6	Conclusion.....	25

1 Summary

Long-term sustainability of the Food Nutrition Security (FNS) Cloud project is based on the implementation of different services for standardisation and interoperability of data on food, nutrition, and security, that have the capacity to link with new resources and enable crosstalk amongst them. The implementation and integration of those services, driven by FAIR principles (findable, accessible, interoperable and reusable), was followed by usability testing of WP4 Use Cases and WP5 Demonstrators, performed amongst user communities from within the consortium. As a result, from this testing, additional needs were identified, considered and implemented in the form of two additional web applications to support users of the FNS-Cloud.

In this deliverable, we first introduce the FNS-Harmony ontology, which was developed to support and interconnect data between different services implemented throughout the project. Next, we introduce a reliable back-end infrastructure which allows implementation and integration of loosely coupled services based on the microservices architecture. External access to the back-end infrastructure was abstracted to the usage of an application programming interface (API), i.e. Metadata API. FNS services for curation and annotation of research data were developed and integrated in the back-end infrastructure as microservices. All tools implemented within the FNS-Cloud, including those identified throughout the exploration of WP4 and WP5 were developed as standalone web applications and could be found in the FNS-Cloud Catalogues¹.

¹ <https://fnscloud.eu/catalogues>

2 Introduction

2.1 Background

In the context of the Food Nutrition Security (FNS) Cloud, data are heterogeneous in data types and formats and encompass different research domains, e.g. Agri-Food, Food Intake and Lifestyle and Nutrition & Health. Therefore, different services for the processing, analysing and visualisation of FNS data are required. Many services developed in the FNS-Cloud project are implemented independently since they rely on different domain data. However, a proper integration of the different services overall into a central entity is required to achieve a user-friendly environment for the FNS end-users.

In this deliverable, we first introduce the FNS-Harmony ontology, which was developed to support different services. Next, we introduce a reliable back-end infrastructure which allows implementation and integration of loosely coupled services based on the microservices architecture. Then, a short overview of the application programming interface (API), i.e. Metadata API, is given. Key FNS services for pre-processing, curation and annotation of research data that were developed throughout the project are also presented. Finally, a short examination of the additional tools implemented within the FNS-Cloud as a result of the exploration of WP4 and WP5 is presented.

2.2 Objectives

The main objectives of this document are to:

- Present the developed FNS-Harmony ontology.
- Give a brief overview of the back-end infrastructure of the FNS-Cloud.
- Introduction of the Metadata Application Programming Interface (API).
- Describe the main applications/services within the FNS-Cloud and
- Describe the identified by WP4 Use Cases and WP5 Demonstrators possibilities for improvements/ modifications.

2.3 Target audience

The target audience include:

- Consortium members in the FNS-Cloud project
- Researcher
- Large enterprises and SMEs
- Policy makers
- Public audience.

3 FNS Harmony Ontology

3.1 Background

In order to provide harmonisation of data within the project as well as to support the development of different software services within WP3 together with the help of partners involved in the applications (WP4) and demonstrators (WP5), we developed the FNS Harmony Ontology (FNS-H). The motivation for developing an ontology for harmonisation of data in food, nutrition, and security stems from the urgent need to address complex challenges and foster collaboration in these critical domains.

An ontology is a structured representation of knowledge that defines and organises the concepts, relationships, and properties within a specific domain. An ontology serves as a standardised framework that enables seamless integration and interpretation of diverse data sources, ranging from agricultural production and food processing to dietary patterns and nutritional outcomes. By establishing a common language and shared understanding, this ontology promotes interoperability, facilitates data exchange, and empowers researchers, policymakers, and stakeholders to gain comprehensive insights into the intricate relationships between food systems, nutrition, and security.

3.2 Ontology Design

Ontology design is a crucial aspect of knowledge representation and organisation, aiming to create a structured framework that captures the concepts, relationships, and properties within a specific domain. Developing ontologies from scratch can be time-consuming and resource-intensive. Hence, the importance of reusing already developed resources becomes evident. By leveraging existing ontologies and reusing well-established concepts and relationships, researchers and practitioners can save valuable time, effort, and resources. Reuse enables the integration of knowledge from multiple domains, promotes interoperability, and facilitates data sharing and collaboration across different disciplines. Moreover, it fosters consistency, standardisation, and semantic clarity, ensuring that information is accurately interpreted and effectively communicated.

In the development of FNS-H, we aimed at reusing and linking to already established reference ontologies that tackle the domains of food, nutrition and security from different viewpoints and are needed by the different services or applications. These include the following popular ontologies: OBI (Ontology of biomedical investigations), DOID (Human Disease ontology), ONS (Ontology for Nutritional Studies), UO (Units of measurements ontology), ENVO (The environment ontology), UBERON (Uber-anatomy ontology), ChEBI (Chemical Entities of Biological Interest) and others. In the cases where existing terms were not defined in other ontologies or the definition of terms were not adequate for our use cases, we defined the term directly in our ontology.

In FNS-H, we used the MIREOT principle when reusing terms from other ontologies. Extracting external ontology terms with the MIREOT (Minimum Information to Reference an External Ontology Term) principle involves selectively importing or referencing relevant terms from external ontologies into a target ontology. The MIREOT principle aims to avoid unnecessary duplication of

terms and ensure a modular and interoperable ontology design. By following this principle, ontology developers can extract only the necessary subset of terms from external ontologies that are required to describe specific domain concepts in the target ontology.

One of the most important aspects to consider when defining terms in an ontology is to establish a format for the identification of each new term. Term identifiers in ontologies are unique identifiers assigned to individual terms or concepts within the ontology. These identifiers serve as stable references to ensure consistency, precision, and interoperability in knowledge representation. Best practices for term identifiers include using globally unique identifiers (such as URIs) to avoid conflicts with other ontologies or vocabularies. The identifiers should be persistent, meaning they remain unchanged even if the term's label or definition is modified.

For the FNS-H each term has been identified with:

- Unique ID in form of an IRI (Internationalized Resource Identifier), composed of the base ontology IRI (<http://www.fns-cloud.eu/>) and term ID composed of the prefix FNS_H and a ten digits number; and
- Term label, which is the language-based representation of the term understandable by humans.

In order to provide more information about the represented terms in the ontology, we reused metadata annotation properties defined by the Dublin Core metadata standards (<https://www.dublincore.org/>) as well as ontology metadata defined by the OBO foundry (<https://github.com/information-artifact-ontology/ontology-metadata>). Annotation properties in ontologies are used to annotate metadata about terms, providing additional descriptive information that helps to enrich their meaning and context. These properties serve as a means to add informative annotations that are not directly related to the formal semantics of the ontology but are valuable for documentation, interpretation, and discovery purposes. Annotation properties can be used to provide textual definitions, synonyms, examples, version information, authors, and other relevant metadata about terms.

3.3 Ontology Structure

The development of the FNS-Harmony ontology started with the need of standardisation across different databases that contain microbiome datasets. The idea was to harmonise the metadata about microbiome studies in order to enable meta-analysis as well as allow researchers to enhance the metadata about their studies. This conceptualisation formed the first ontology module. Further on in the ontology development, we also included other ontology modules. Ontology modularisation brings numerous benefits that make it highly advantageous in knowledge representation and management. Firstly, modularisation allows for the decomposition of complex ontologies into smaller, more manageable parts. Furthermore, ontology modularisation promotes reusability by creating a repository of standardised, self-contained modules that can be easily shared and integrated into different applications and domains. Another advantage of ontology modularisation is its ability to enhance scalability and performance.

The current development version of FNS-Harmony is composed of six modules. Each model contains the conceptualisation of a certain part of the food, nutrition and security domain. The ontology

modules can be used separately or combined depending on the use-case needs. Here we briefly describe the different modules.

Microbiome data module

In the initial phase we started with collection of metadata terms from ENA database checklists (<https://www.ebi.ac.uk/ena/browser/checklists>). For the representation of the metadata terms we imported and mapped to terms from existing ontologies. If a term does not exist in any relevant external ontology, we modelled it inside FNS-Harmony. This ontology module was used by ScaleFocus as a base for the metadata submission form and service that was designed to support study metadata enrichment. In Figure 1, we present the overall structure of this ontology module.

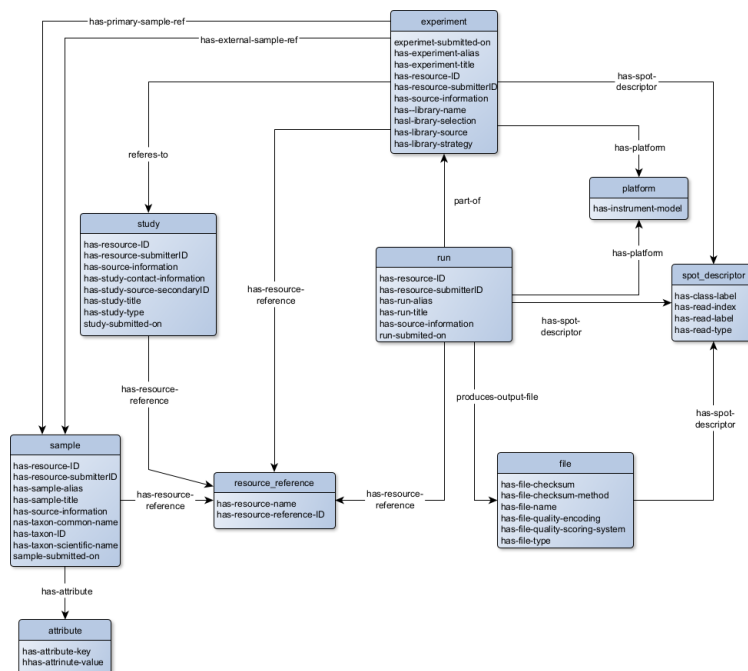


Figure 1. Overall structure of the Microbiome ontology module.

In Figure 1, we present the overall structure of the microbiome ontology module. The core FNS-Harmony entities in this module are study, sample, attribute, experiment and run. Supporting entities are platform, file, spot descriptor and resource reference.

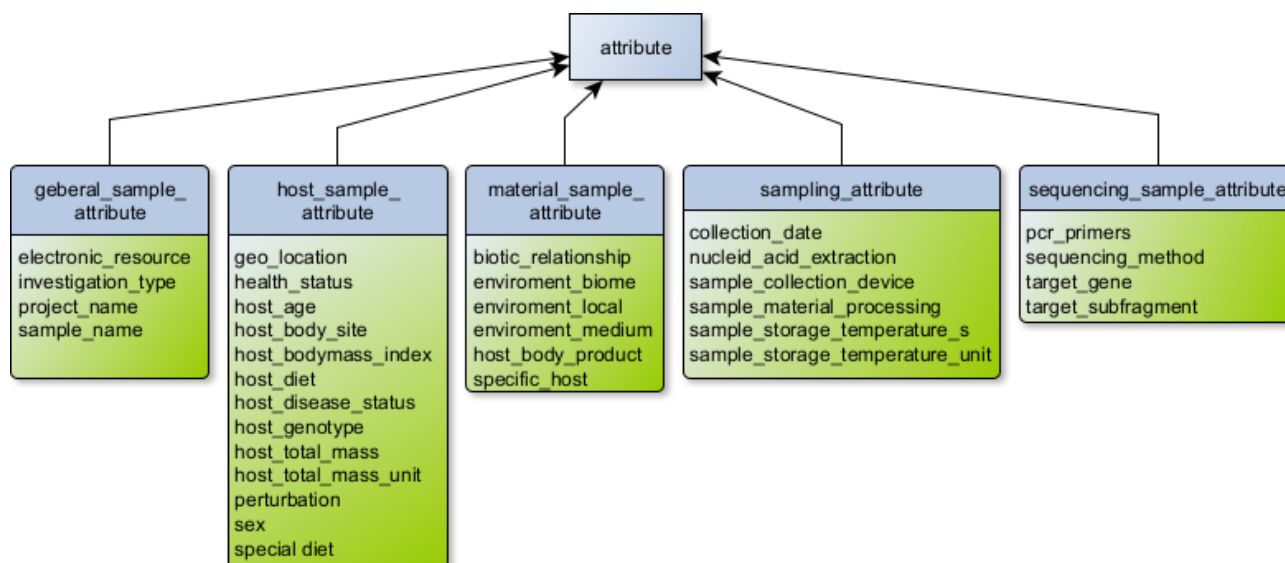


Figure 2. Representation of biological metadata as attributes.

In Figure 2, we represent the different biological metadata that characterise the samples. The representation is done by employing the attribute class. FNS-Harmony includes several different types of attributes: general sample attributes, host sample attributes, material sample attributes, sampling attributes and sequencing sample attributes. It is important to note that the list of attributes can be easily extended with new attributes.

To have reusable and interoperable metadata we align the FNS attribute terms to terms from state-of-the-art domain ontologies. Currently, this module imports terms from: OBI (Ontology of biomedical investigations), DOID (Human Disease ontology), ONS (Ontology for Nutritional Studies), UO (Units of measurements ontology), ENVO (The environment ontology), and UBERON (Uber-anatomy ontology).

Food components module

The food components module contains conceptualisation and taxonomy of food components that are standardised according to EuroFIR. Each food component contains a mapping to the appropriate EuroFIR code, has component names in English and Slovenian, has label mappings to different national food components databases (France, UK, Australia, Danmark, Netherlands and US). In addition, we also represented the taxonomy of food components. For some food components, we also provide a mapping to the ChEBI ontology using the ChEBI ID. The first level of the taxonomy distinguishes between ten major categories, which include:

- Carbohydrate components;
- Food additives;
- Food properties and measures;
- Lipid components;
- Minerals and inorganic components;
- Nitrogen components;
- Organic acids;
- Phenolic components;
- Proximates; and

- Vitamins.

Bioactives module

The bioactives module contains representations of different groups of bioactive components that are especially important for the project demonstrators. The bioactive components were extracted from the e-BASIS (Bioactive Substances in Food Information Systems) EuroFIR (<https://www.eurofir.org/our-tools/ebasis/>) resource. For terms representing bioactive components, we provide a mapping to adequate ChEBI terms where available. The terms are organised in five major groups:

- Polyphenols;
- Sulphur compounds;
- Phytosterols;
- Antioxidant compound group; and
- Others.

Dietary interventions module

The dietary intervention module contains terms that denote different dietary interventions that are used in nutritional intervention studies. For this purpose, we extracted a list of interventions from studies represented in the Phenotype Database² (dBNP). The preliminary list from dBNP was cleaned from ambiguous entries and duplicates were removed. Where possible, we provide mappings of intervention terms from other ontologies. The terms in this module are used by ScaleFocus in the metadata submission form.

Branded food module

The branded food module contains the initial conceptualisation of branded food products. The conceptualisation is based on the processes of collecting and organising data about branded food products such as the CLAS database³ (Composition and Labelling Information System), developed by the project partner NUTRIS. We believe that this ontology module has large potential of further development and consequently to be reused by various institutions and companies working with such data.

Product identifiers module

The product identifiers module contains the conceptualisation of the different GS1 identifiers (<https://www.gs1.org/>) which are used for identification of products. Here, in collaboration with the project partner GS1 Slovenia we performed initial conceptualisation of the different identifiers such as GTIN (GTIN8, GTIN12, GTIN13), and others. We also represented part of the Global Product Classification (GPC) that contains the Food/Beverage/Tobacco segment in an ontological form. We believe that this module can be used together with the branded food module to provide more in-depth representation of branded food products and their characteristics.

² <https://dashin.eu/interventionstudies/>

³ <https://www.frontiersin.org/articles/10.3389/fnut.2021.798576/full>

3.4 Ontology Versions and Implementation

The FNS-H ontology has been developed by using the Protégé ontology editor (<https://protege.stanford.edu/>). Classes from external ontologies were extracted using the OntoFox tool (<https://ontofox.hegroup.org/>) using the MIREOT principle (Minimal Informa). For the transformation of information from spreadsheets to the ontology, we used the Celfie plug-in for Protégé (<https://github.com/protegeproject/celfie-plugin>) as well as the ROBOT OBO tool (<http://robot.obolibrary.org/>). The ontology is expressed in the OWL2 language (<https://www.w3.org/TR/owl2-overview/>), which is a standard language for knowledge representation recommended by the W3C (<https://www.w3.org/>).

The ontology has been released in three versions (1.0, 1.1 and 1.2) and the final version 1.3 will be released before the project ends. Versions 1.0 – 1.2 are focused mostly on the microbiome use case, while version 1.3 contains also other modules: food components, bio actives, dietary intervention, branded food and product identifiers.

3.5 Availability

The developed ontology is available via several sources:

- GitHub repository: <https://github.com/panovp/FNS-Harmony>
- Ontology PURL: <https://purl.org/fns-h>
- BioPortal Ontology Repository: <https://bioportal.bioontology.org/ontologies/FNS-H>

The GitHub repository contains the raw ontology source files and allows collaborative development of the ontology. The repository contains all released versions of the ontology. In addition, the git repository contains an issue tracker where collaborators can post requests for new terms to be added to the ontology as well as to post questions and open discussion points. The ontology permanent identifier PURL uses <https://purl.archive.org/> service to always resolve and redirect to the last published version of the ontology. This PURL can be used from any ontology editor to directly load, explore and use the last version of the ontology. Finally, the ontology is also indexed and available through the BioPortal ontology repository (<https://bioportal.bioontology.org/>) which is one of the largest repositories of ontologies for the domain of biomedicine and wider. With this we are increasing the possibility of potential users to find the ontology and reuse it for their use cases.

4 FNS Services

4.1 Background

The Food Nutrition Security Cloud is a multi-component system which provides a framework to store and analyse dataset repositories from the food, nutrition, and security fields in order to gain new insights on healthy diet and healthy lifestyle of individuals. On that account, data created in research studies needs to be stored and managed in such a way that consortium members can share data sets with one another and collaborate closely to gain knowledge from the data.

The concept of describing scientific data in the form of metadata provides strong insights of research information and it is especially useful when comparing data from different domains. By definition, metadata is a set of data that provides information about data objects. In the context of the project, metadata plays a crucial role for the description of scientific research and studies data. Therefore, the need for a reliable infrastructure for the extraction of metadata from external sources and its transformation in a suitable form for further analysis was distinctly perceived at the very beginning of the project. Besides the curation of metadata from external sources e.g. European Nucleotide Archive (ENA), such infrastructure should also provide an efficient mechanism for the direct submission of metadata for annotation of scientific research data to the FNS-Cloud, also by considering the newly developed FNS Harmony ontology.

Analysis and visualisation of various data sets can further help FNS-Cloud users and researchers perceive new insights on the research data. Automated data classification and matching to other data sets with the aid of Ontologies provide users of the FNS-Cloud with an insight on similarities and dissimilarities between different data sets and can be found useful in many research cases.

4.2 Related Work

Microservices is a relatively modern way of developing back-end applications. When talking about microservices, the functionality of the whole software system is split up into multiple small applications or services, where each service can be developed and deployed independently from one another.

Usually, communication between the services is accomplished over a network using common protocols such as HTTP (RESTful APIs). Lately, with the growth in popularity of the microservices architecture the means of communication between microservices have changed from protocol-oriented communication to event-based communication.

A versatile basis for building distributed event-based infrastructures are publish/subscribe systems⁴. Publish/Subscribe is a messaging pattern where components communicate by publishing notifications about occurred events to specific topics and other components subscribe for the notifications to the same topics. The result of this software architecture is loose coupling between participating components.

⁴ <https://ably.com/topic/pub-sub>

4.3 Introduction to the FNS Services and the Back-end Infrastructure

In the scope of the FNS-Cloud project, a wide range of datasets from different domains are integrated and analysed in order for researchers to gain new perspectives and knowledge about healthy diet, healthy lifestyle and longevity. The mechanism for curation and annotation of scientific data from the food, nutrition and security fields and its automatic processing, analysis and visualisation is at the core of the developed functionalities. The FNS-Cloud Domain consists of heterogeneous data and as already stated above its processing must be supported by a reliable infrastructure which is highly available and ensures data consistency.

4.3.1 FNS-Cloud Back-End Architecture

With the microservices architecture in mind, we have divided the identified data manipulation and transformation procedures required for classification and analysis of metadata into microservices that can be vertically scaled on high demand.

To ensure high throughput we have adopted Apache Cassandra⁵ as a storage solution since it is designed to handle large amounts of data across many commodity servers, providing high availability with no single point of failure. Writes procedures to Cassandra are achieved through Sink Connectors from the distributed event streaming platform Apache Kafka⁶ which is positioned at the centre of the backend infrastructure to ensure proper communication between API functions and the developed services.

By using the service-oriented computing paradigm we aim to take advantage of serverless computing using container technology while implementing best DevOps practices such as continuous deployment and integration pipelines. The figure below depicts the develops infrastructure:

⁵ <https://cassandra.apache.org/>

⁶ <https://kafka.apache.org/>

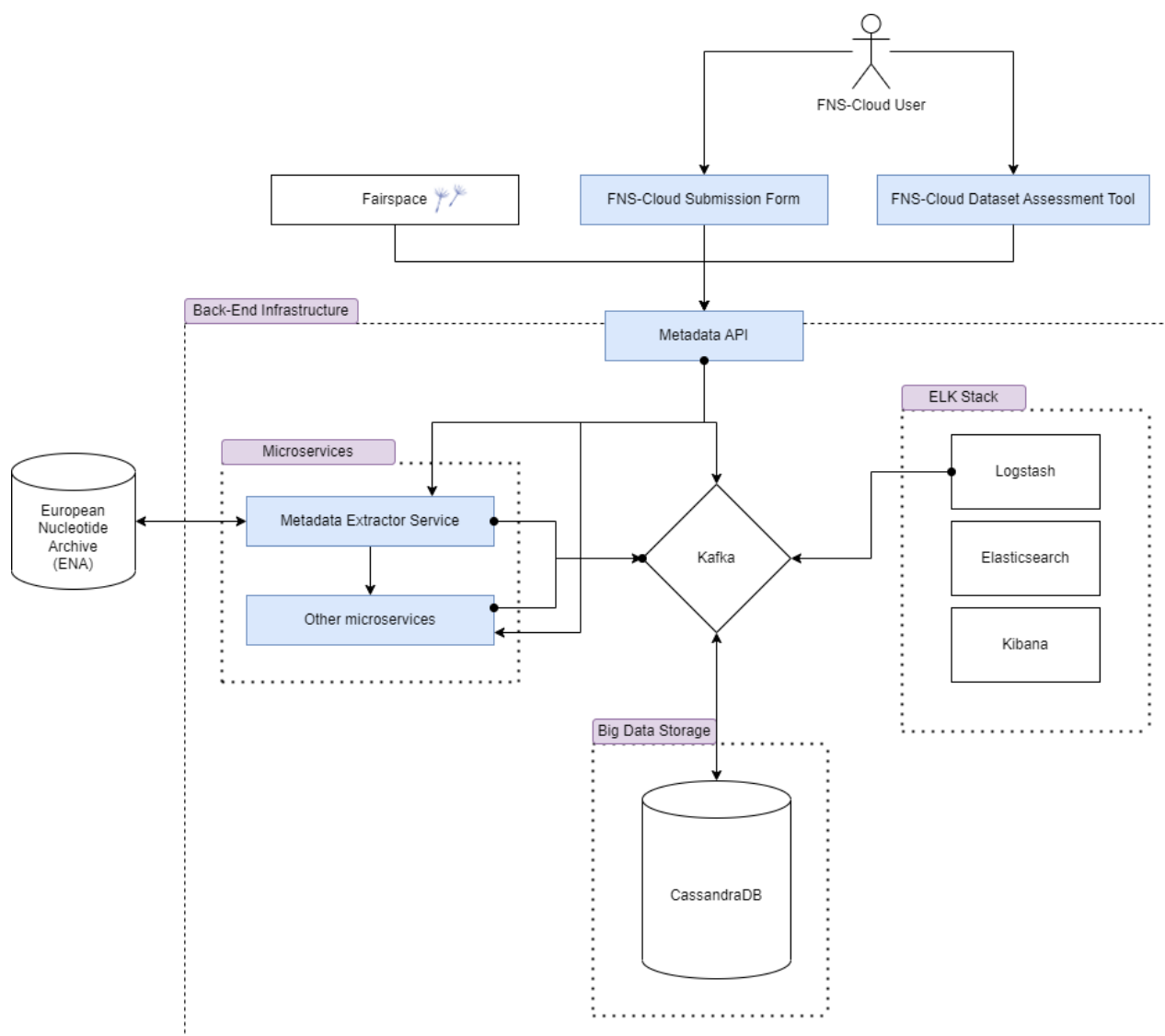


Figure 3. FNS-Cloud Back-end infrastructure

As depicted in the figure above, search capabilities are achieved using the ELK Stack⁷ which consists of Elasticsearch, Logstash and Kibana. Elasticsearch is a full-text search and analytics engine based on another search engine known as Apache Lucene. Elasticsearch is also a high-performance database which allows searching and exploring huge volumes of data in almost real time manner. Except for search functionality, in the backend infrastructure Elasticsearch is also used to retrieve each data set as a catalogue entry consisting of different collections such as samples, runs, etc. We have identified that closely coupling Elasticsearch with Cassandra provides better querying capabilities than executing queries directly to Cassandra.

Since the streaming platform at the heart of the infrastructure is event-based for each event occurring in the system logs are produced. These logs are then streamed to Logstash, which is a data collection tool capable of reading and normalising log messages from various sources. The inputs from the sources are often unstructured or have their specific characteristic structure and Logstash

⁷ <https://www.elastic.co/what-is/elk-stack>

brings them in a consistent form, usually a JSON-like event containing semi-structured document sets.

Visualisation of the data in the back-end storage is achieved through Kibana which is a highly configurable open-source visualisation tool for Elasticsearch data. It provides the functionality of importing logs from Logstash and Elasticsearch and can give insights on the performance in an understandable and detailed way. It offers different kinds of graphical representation like bar charts, tables and diagrams which makes it easy to comprehend large volumes of data.

4.3.2 FNS-Cloud Metadata API

With the intention of web services to easily take advantage of the infrastructure described above, an Application Programming Interface (API) was implemented. An API is an intermediary between two or more applications which clearly defines how these applications would talk to each other. An API can as well be seen as an open portal to another application or a component. It provides high level abstraction and flexibility to other developers when programming and helps saving time and financial resources. The Metadata API developed in WP3 of the FNS-Cloud provides common means for submission and querying research metadata stored in back-end infrastructure described above.

In order to access the data stored in back-end FNS infrastructure and to submit new data REST was chosen as a software architecture for the implementation of the Metadata API. By adopting the REST architectural style, we introduced flexibility in the infrastructure when accessing the metadata stored at the backend. Using REST, Uniform Resource Identifiers (URIs) were defined for each identified collection resource stored in the back end, i.e. research project, sample, ontology, etc. Using URIs, the API complies with the FAIR principle for data i.e., findable, accessible, interoperable and reusable. The figure below shows the documentation page of the Metadata API:

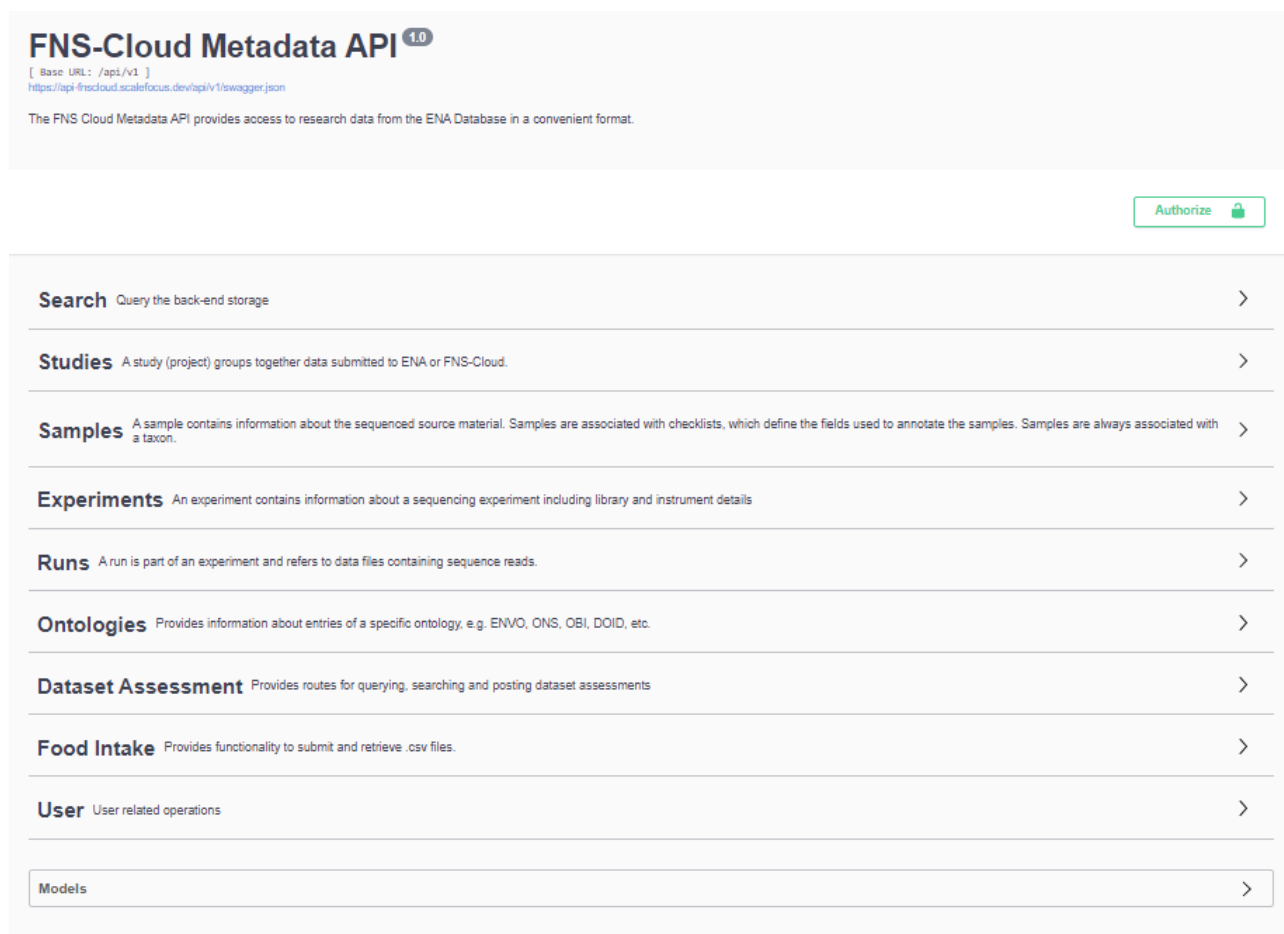


Figure 4. FNS-Cloud Metadata API

4.3.3 FNS-Cloud Services

Based on the initial work on WP3 tasks several FNS Services were distinctly identified. All of the developed services are standalone web applications or services that can be accessed through the FNS-Cloud Catalogues under Tools or services, respectively. The tools developed by ScaleFocus are strongly integrated with the back-end infrastructure and the Metadata API described above. Additionally, Scalefocus has prepared guidelines (https://drive.google.com/file/d/1WON0JBcQuhW7ppU_SD3oL0bwstoZyvKL/) to be used after the end of the project for the potential integration of new services. Below we give a brief overview of the FNS services developed in WP3:

4.3.3.1 Food matching service

The food matching service developed by JSI in WP3 comprises three services, which support the following functionality: data pre-processing, food matching, and calculation of ingredient quantities and missing nutrient values. These services can be integrated in any information system requiring matching of food-related data of different types (e.g., food consumption data with food matching data, or data on food ingredients specified on the food product label with food composition data) and calculation of missing nutrient values (e.g., for branded foods having information about their ingredients and basic nutrient profile). The services are annotated with metadata specified by the

FNS-Cloud Catalogues. JSI developed web services for food matching are accessible via the FNS-Cloud Catalogues.

4.3.3.2 FoodMagnifier Mobile Application

The FoodMagnifier app is a low-end mobile application developed by PMT and BfR and its main purpose is to educate the general public about the content data of contaminants and residues in foods. Information is presented to the user in visual format, utilising graphs and charts comparing the amounts of substances in different types of foods. Background information such as a description of the chemical, exposure data, health-based and toxicological reference values, and the role that various food categories play in dietary exposure are also included in the app.

4.3.3.3 Metadata Extraction Service

The Metadata Extraction Service developed by Scalefocus is used to query data from external bioinformatics data repositories. Currently the service supports querying the European Nucleotide Archive (ENA) and retrieving metadata about existent research data. The ENA queries are based on a given accession key of a study/project collection. The information retrieved about a research project and its sub-collections i.e. samples, runs, etc is transformed into suitable format that can be easily parsed by machines and understood by humans at the same time. Based on the data retrieved from ENA, we were able to identify which information and its structures were most important for the data annotation, data matching and analysis.

4.3.3.4 Microbiome Analysis Pipeline

Another software solution used for data analysis is a pipeline developed by JSI in collaboration with Quadram to support the microbiome demonstrator (described in detail in D5.7). Once applied for the analysis on data collected in the DIME study, the pipeline will be published in the FNS-Cloud Catalogues.

4.3.3.5 Analysis and Visualisation Pipelines

UM has developed two integrative data analysis pipelines to functionally analyse and visualise metabolomic, transcriptomic and metagenomic data. Data from the Metabolights and MGnify databases will be selected by the user with the metadata capture framework in FairSpace. Thereafter, the data can be analysed with the dedicated analysis pipeline (more detail provided in D5.7).

4.3.3.6 Descriptive and predictive analysis demonstration

A lightweight Python notebook was developed by JSI to demonstrate the possibility of using descriptive and predictive machine learning tools in FairSpace with the use of data that is available in FairSpace through links in meta-data. The notebook focuses on the DIME study data (performed by Quadram) by querying the biome expression (bacterial relative abundance) data using the MGnify API. Several variants of the datasets are prepared using the Pandas data analysis python library, which are then analysed using the Orange3 machine learning toolkit, using descriptive (feature ranking) and predictive (decision trees) approaches. This notebook closely follows the

analysis presented in Section 4.3.4 of D5.7. Code is available at <https://github.com/aosojnik/fnscloud-jsi-ml-analysis>.



5 Integration of the implemented FNS Services

5.1 Background

For the purposes of the Food Nutrition Security (FNS) Cloud project, different services for standardisation and interoperability of data on food, nutrition, and security have been developed.

Although many of the FNS services are implemented independently, integration into a central entity is required to improve FNS-Cloud end-user experience. However, merging multiple software applications designed for individual purposes together is a challenging task. Therefore, PMT developed the FNS-Cloud Catalogues⁸ which contain information about datasets related to food, nutrition and security as well as e-tools used for management and analysis of data. Applications developed throughout the implementation phase of WP3 as well as tools identified after the exploration of WP4 and WP5 can be found in the Catalogues.

In this chapter, we introduce the tools and services developed as a result of the exploration of WP4 and WP5 and give a brief overview of functionalities.

5.2 Extended API Specification

Since the Metadata API plays the role of a gateway to the back-end infrastructure, it supports several web applications developed in WP3. However, several changes of the API were needed in order to introduce additional operations on FNS-Cloud data and fit the needs of WP4 and WP5 specifications. On that account, new functionalities were necessary in order to improve the user experience at the front-end applications. As depicted in Figure 4, the metadata doesn't provide only operations on the main microbiome collections i.e., Studies, Samples, Runs, etc., but also provides functionalities about Ontologies and Food Intake. Finally, we implemented a mechanism for evaluation of FNS-Catalogues which serves as a foundation for the Dataset Assessment Framework described below.

5.3 Integration within FNS-Cloud

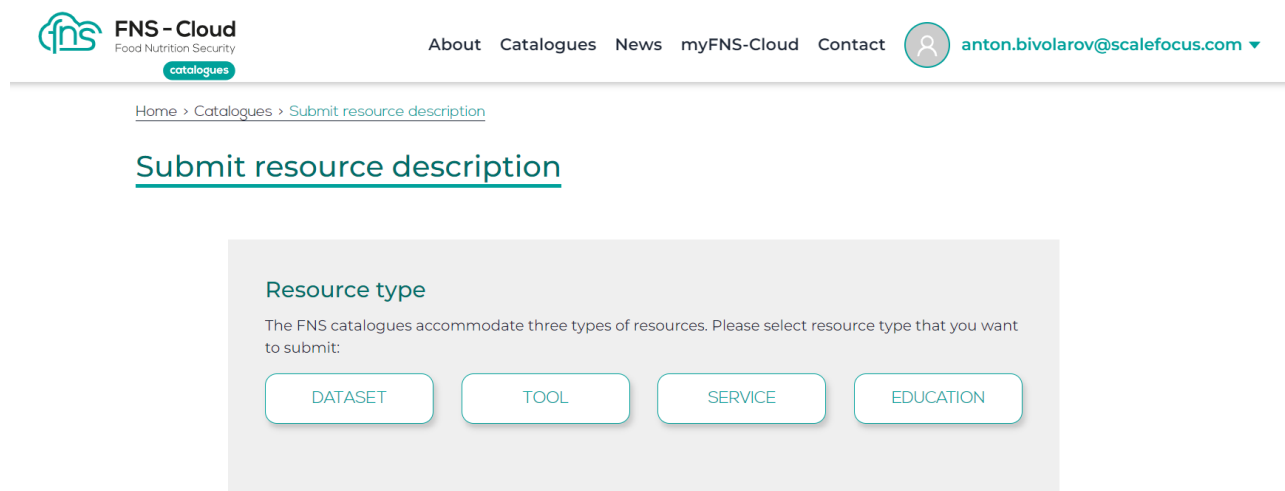
Most of the tools and services introduced above are already integrated or will be integrated in the FNS-Cloud Catalogues by the end of the project. Those resources can be queried under the Tools and Services section of the FNS-Cloud Catalogues.

As stated above, FNS-Cloud Catalogues were created by PMT in WP2 and refined according to the needs of WP4 and WP5. They contain information about datasets related to the topics of food, nutrition and security, tools like apps and software to manage and analyse data, services that are provided by FNS-Cloud or our verified partners, as well as training and education materials about food nutrition security. They allow the user to search using queries and FNS-Cloud Topics, as well as other metadata of the available resources. Results are visualised as a list or a two-dimensional table, depending on the chosen search option.

Resources can be submitted by users after setting up an account with the FNS-Cloud Authentication and Authorisation Infrastructure. This is also an entry point to many FNS-Cloud services, like FNS-

⁸ <https://fnscloud.eu/catalogues>

Cloud Catalogues, Fairspace, Dataset Assessment Framework, Metadata API and the Metadata Submission Form:



The screenshot shows the 'Submit resource description' page of the FNS-Cloud Catalogues. The page header includes the FNS-Cloud logo, navigation links (About, Catalogues, News, myFNS-Cloud, Contact), and a user profile for anton.bivolarov@scalefocus.com. The breadcrumb trail is 'Home > Catalogues > Submit resource description'. The main heading is 'Submit resource description'. Below this, a section titled 'Resource type' explains that the FNS catalogues accommodate three types of resources and asks the user to select the type they want to submit. There are four buttons: DATASET, TOOL, SERVICE, and EDUCATION.

Figure 5. Submitting new resource to the FNS Catalogues (<https://fnsccloud.eu/catalogues/submit>)

In addition, after exploring the services which were described in the previous section and were developed throughout the WP3 development phase of the project, it was evident that additional development of tools is required in order to validate and further aid the user-experience in the FNS-Cloud. Below we give an overview of the additional tools and services developed as a result of the exploration of the WP3 services by WP4 and WP5.

5.3.1 FNS-Cloud Dataset Assessment Tool

The dataset quality assessment framework supports researchers in the process of evaluating if a given dataset in the FNS-Cloud catalogues can be used for answering a specific research question in a researching project. It guides the user through several question-based topics and stores each assessment the user provides. Users also have the ability to get an overview of all of the assessments they have submitted and continue an assessment if they haven't finished. By using this framework users of the FNS-Cloud can validate how useful a dataset can be for their research. The Dataset Assessment Tool is integrated in the FNS-Cloud Catalogues. The figures below depict the landing page and the overview an assessment of the web application:

Click to go back, hold to see history **Assessment** Overview anton.bivolarov@scalegfocus.com **Logout**

Introductory message

This tool is intended to support you, the data user in making the decision of whether a pre-existing dietary intake dataset is appropriate to re-use to answer your specific research question. You will be asked a series of questions related to how the data was collected, the underlying databases used to generate the data, how the data was handled / coded, and associated data which may also be available within the dataset e.g., demographic, anthropometric, consumer behaviour and lifestyle indices. At the end you will get a personalised report complete with messages containing some information on each parameter and factors you should consider based on your answers.

You may begin by entering a research question you would like to answer below and selecting the dataset you are considering using to answer this question. If you are considering using more than one dataset you may complete multiple assessments on the same research question.

As you progress through the questions, there is a facility to skip questions you are unsure of and come back to them at a later point. Saved assessments can be found in the 'Overview' tab at the top of this screen. Before you begin an assessment, you may view the metadata available for the dataset selected if it is present in the FNS-Cloud catalogue. Alternatively, you may complete a quality assessment on a dataset that is not in the FNS-Cloud catalogue by selecting 'Other' from the dropdown menu under 'Select dataset'.

New assessment

Enter research question(s)*

Select dataset*

Assess


 Food Nutrition Security Cloud (FNS-Cloud) has received funding from the European Union's Horizon 2020 Research and Innovation programme (H2020-EU.3.2.2.3. – A sustainable and competitive agri-food industry) under Grant Agreement No. 863059. Information and views set out across this website are those of the Consortium and do not necessarily reflect the official opinion or position of the European Union. Neither European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use that may be made of the information contained herein.

Figure 6. Dataset Assessment Landing Page

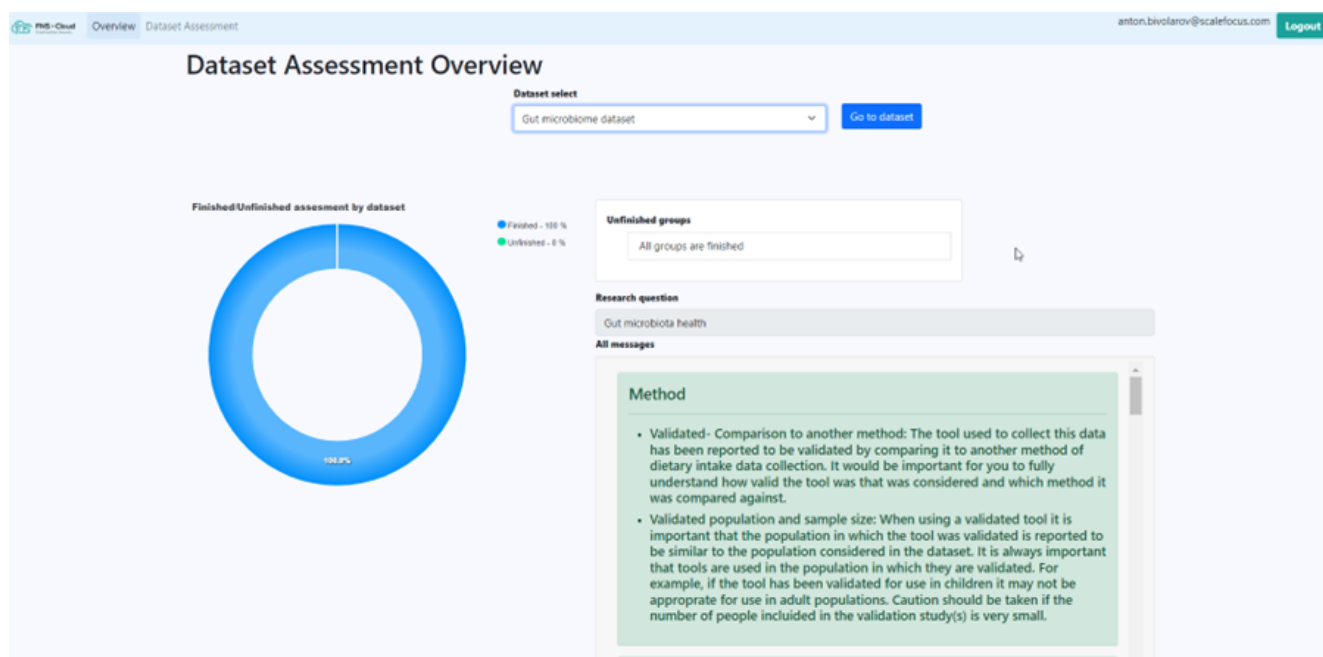
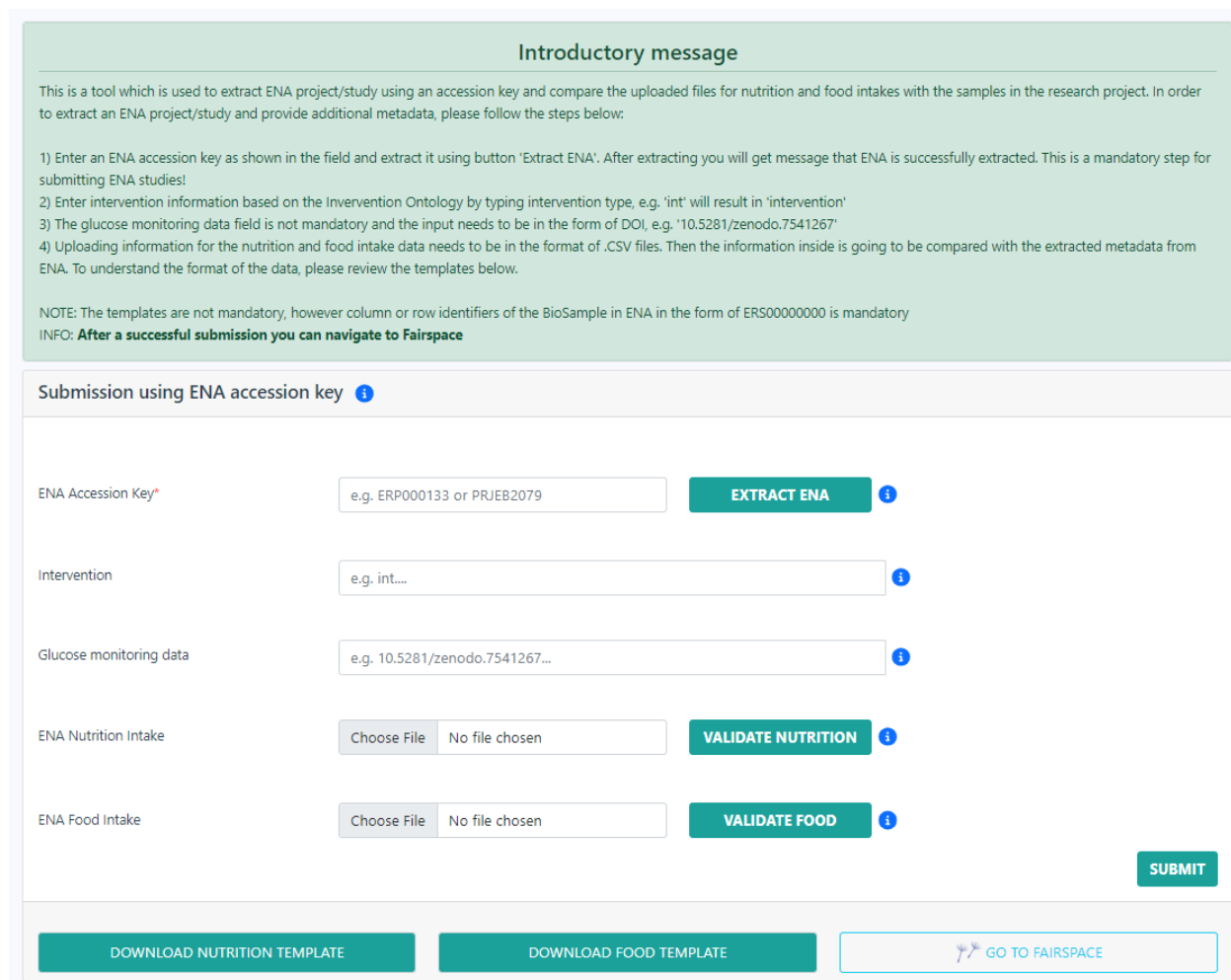


Figure 7. Dataset Assessment Overview page

5.3.2 FNS-Cloud Metadata Submission Form

The Metadata Submission Form is a web-based tool which is used to enhance microbiome studies from the European Nucleotide Archive (ENA) with food and nutritional intake data. The tool also provides the ability to submit intervention information about the study based on the intervention subdomain of the FNS Harmony Ontology. Moreover, the web application can also be used for the submission of metadata to a microbiome study given the raw data is stored in a server which is not

part of the ENA servers stack. This type of submission supports Ontology based input on many metadata fields and it is used mainly to constrain the inconsistent free-text input of users. The Submission Form plays a central role in the DEMO3 pipeline and Fairspace queries data which was submitted by users using the form (See Figure 3). The tool can also be used independently by querying enhanced data directly from the Metadata API (See Section 4.3.2.). The figure below depicts the landing page of the web application:



Introductory message

This is a tool which is used to extract ENA project/study using an accession key and compare the uploaded files for nutrition and food intakes with the samples in the research project. In order to extract an ENA project/study and provide additional metadata, please follow the steps below:

- 1) Enter an ENA accession key as shown in the field and extract it using button 'Extract ENA'. After extracting you will get message that ENA is successfully extracted. This is a mandatory step for submitting ENA studies!
- 2) Enter intervention information based on the Intervention Ontology by typing intervention type, e.g. 'int' will result in 'intervention'
- 3) The glucose monitoring data field is not mandatory and the input needs to be in the form of DOI, e.g. '10.5281/zenodo.7541267'
- 4) Uploading information for the nutrition and food intake data needs to be in the format of .CSV files. Then the information inside is going to be compared with the extracted metadata from ENA. To understand the format of the data, please review the templates below.

NOTE: The templates are not mandatory, however column or row identifiers of the BioSample in ENA in the form of ERS000000000 is mandatory
INFO: **After a successful submission you can navigate to Fairspace**

Submission using ENA accession key

ENA Accession Key* **EXTRACT ENA**

Intervention **VALIDATE NUTRITION**

Glucose monitoring data **VALIDATE FOOD**

ENA Nutrition Intake **VALIDATE NUTRITION**

ENA Food Intake **VALIDATE FOOD**

SUBMIT

DOWNLOAD NUTRITION TEMPLATE **DOWNLOAD FOOD TEMPLATE** [GO TO FAIRSPACE](#)

Figure 8. Enhanced metadata submission form

5.3.3 Metadata enrichment tools

The metadata on studies in Fairspace and other databases is commonly annotated to a varied degree with some studies having abundant metadata, while for most of them the metadata is very limited. One of the main reasons for lacking metadata is probably the laboriousness and the corresponding cost of this process. The use of metadata for searching and filtering is further hampered by open-ended categories used in annotations and human errors in manual annotation (see Deliverable D5.7). While categorization can be resolved with well-designed ontologies, which also serve a number of other purposes (see Section 3), human errors and missing annotations demand other solutions. For this purpose, we have developed a collection of software tools for Fairspace metadata enrichment. Roughly these are of two kinds: (I) based on automated data scraping: the **meta-data Web scraping tool**, and (II) based on machine-learning and large language

models: the **topic modelling** and **keyword extraction**. While the metadata Web scraping tool provides the missing annotations to the studies, the other two approaches provide additional new annotations that can be useful for searching and filtering. In the following we describe them in more detail.

5.3.3.1 Metadata Web scraping

Metadata in exemplary studies from Fairspace, which we received as representatives and startpoints for meta-data enrichment, consists of a unique label identifier, title, short text description, and a URL to the study in the EBI repository. Additionally, studies may include a PubMed identifier, authors of the study, and the publishing year, if a corresponding publication exists. We mined additional information for these studies by multimodal data scraping from multiple sources.

In the data enrichment process we first focus on the exploration and extraction of the information from the EBI repository (<https://www.ebi.ac.uk/>) via the provided link. We used the BeautifulSoup Python library (<https://www.crummy.com/software/BeautifulSoup/bs4/doc/>) for web scraping and parsing the scraped results. From the scraped data, we enriched the initial data with information about the samples-counts used in experiments, the origin of data used in the experiments, the name of the research institution where experiments took place, and the URLs of related studies, links, and publications. From the data origin links, we further extracted information about the number of experimental samples, the number of spots, instruments, strategy, source tissue, selection method, layout, and the sequencing type, if present.

Finally, for the studies with PubMed entries, we scraped text information also from the PubMed repository. We utilised the PubMed-s endpoint via JSON packaging to obtain the files in a machine-understandable way. If present, we extracted the title, abstract, authors, and keywords of the publication.

5.3.3.2 Topic modelling

Topic modelling based on textual contents of the studies' meta-data was conducted in order to gain additional annotations, in this case bottom-up data-based ones.

In our approach, we utilised BERTopic (<https://maartengr.github.io/BERTopic/index.html>) as the chosen topic modelling technique. The process began by generating potential topics through word embedding using BERT-based embeddings. Subsequently, UMAP (<https://umap-learn.readthedocs.io/en/latest/>) was employed to project the embeddings into a lower-dimensional latent space. Next, CountVectorizer and c-TF-IDF were utilised to rank the potential topics, selecting only those statistically significant to the given metrics. c-TF-IDF is defined as:

where the frequency of each word t is extracted for each class i and divided by the total number of words w . This action can be seen as a form of regularisation of frequent words in the class. Next, the total, unjoined, number of documents m is divided by the total frequency of word t across all

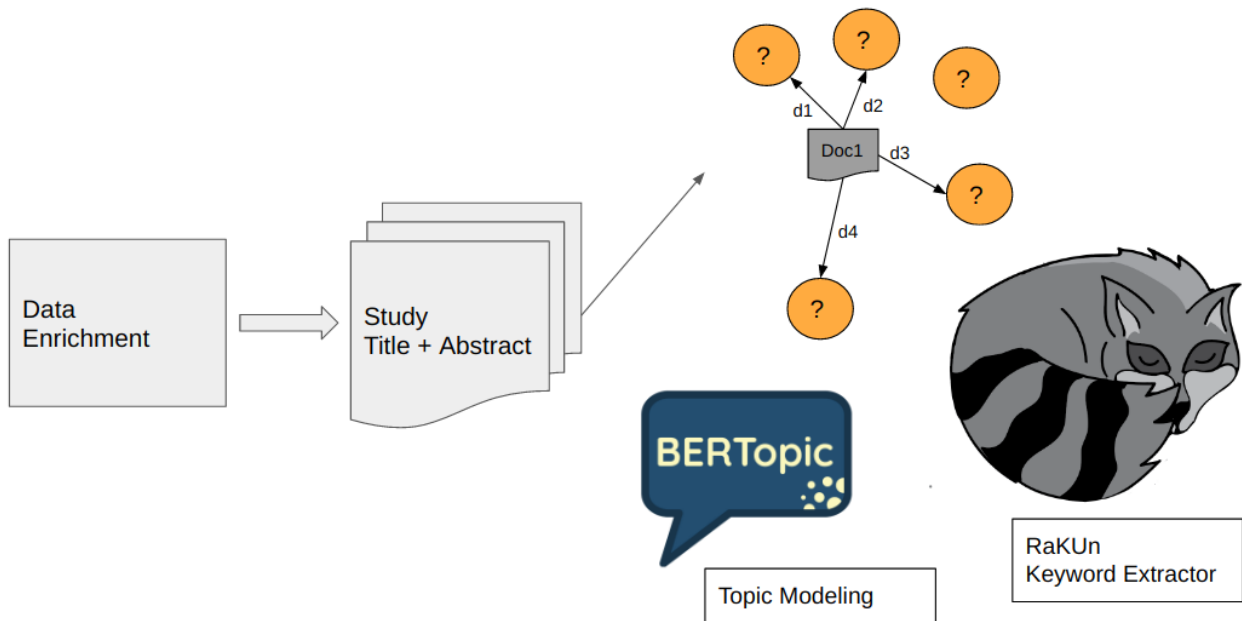
$$c\text{-TF-IDF} = \frac{t_i}{w_i} \cdot \frac{m}{\sum_j^n t_j}$$

classes n . Finally, HDBScan clustering was applied, assigning topic labels to documents based on the topic distribution within the clusters.

5.3.3.3 Keyword extraction

In our keyword extraction process, we employed the RaKUn method, a technique that transforms text into a graph and ranks its nodes. To implement this method, we used the version obtained from GitHub (<https://github.com/SkBlaz/rakun>). We specified several parameters to fine-tune the extraction process. The "num_keywords" parameter was set to 10 to focus on the most significant keywords. We set the "merge_threshold" to 1.1, merging similar nodes in the graph. The "alpha" parameter, set at 0.3, influenced the ranking algorithm. Lastly, the "token_prune_len" parameter was set to 3, excluding very short tokens. These parameter configurations helped us extract meaningful keywords from the text, providing valuable insights for further analysis.

5.3.3.4 Summary



Example of a final result (taken from Deliverable D5.7) is in the figure below [CAPTION: Example of how initial meta-data of a study (framed in black) was enriched with annotations gained with Web-scraping (framed in orange) and newly proposed semantic annotations (framed in blue)].

Initial data	id	MGYS00005607
	links	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
Scraped data	accession	MGYS00005607
	bioproject	PRJEB39305
	samples-count	57.0
	is-private	False
	secondary-accession	ERP122806
	centre-name	VETERINARY MEDICINE UNIVERSITY
	public-release-date	NaN
	study-abstract	Dietary n-3 PUFA fed to ruminants are extensiv...
	study-name	New strategies for the protection of PUFA from...
	data-origination	SUBMITTED
	last-update	2020-09-22T15:21:08
	samples	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	downloads	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	analyses	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	geocoordinates	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	publications	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	studies	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	biomes	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
	links.1	https://www.ebi.ac.uk/metagenomics/api/v1/stud...
ML annotations	abstract texts	527
	extracted_keywords	['extensively hydrogenated', 'microbiome compo...]
	Topic	10
	Name	10_rumen_manure_dairy_cow
	Top_n_words	rumen - manure - dairy - cow - efficiency - pu...
	Probability	1.0
	Representative_document	False

Figure 11. Final result of metadata enrichment.

5.3.3.5 Further work

To optimise our workflow and integrate it with Fairspace or similar tools, we plan to use Docker⁹ in order to create containers for the three crucial components: the data extraction, topic modelling, and keyword extraction models. By containerizing these models along with their dependencies, we aim to streamline their deployment process, improve portability, and ensure consistent performance across different environments. Containerisation (Docker) will enable us to package each model as a self-contained unit, including the necessary libraries and configurations, avoiding potential compatibility issues. With Docker (<https://www.docker.com/>), we can manage and deploy these models as isolated containers, simplifying version control, updates, and maintenance. By using Docker for the data extraction, topic modelling, and keyword extraction models, we will enhance their versatility, scalability, and integration capabilities, enabling us to efficiently utilise them in diverse applications and environments while ensuring reproducibility and ease of deployment. The task would be done by the end of the project.

⁹ <https://www.docker.com/>

6 Conclusion

In this deliverable, we introduced the FNS-Harmony Ontology used to provide harmonisation of data within the FNS-Cloud project and to support the development of different software services. Additionally, we introduced a reliable infrastructure that serves the necessities of several front-end applications while promoting high availability and fault tolerance. It supports high data throughput based on Apache Cassandra database solution. The infrastructure adopts the popular microservice architecture which allows for flexible design of small applications to fulfil different requirements. Some of the identified functionality requirements in WP3 were implemented as loosely coupled services in the infrastructure. Others are developed as standalone web applications that are already integrated with FNS-Cloud Catalogues developed in WP2. Some of the standalone web tools rely on the Metadata API for many functionalities. Tools developed as a result of the exploration of WP4 and WP5 were also introduced and could already be found in the FNS Catalogues.

