Food Nutrition Security Cloud

Deliverable 3.2

Methodology for FNS data standardisation and interoperability

Due Date:	30.09.2020
Submission Date:	30.09.2020
Dissemination Level:	Public (PU)
Lead beneficiary:	JSI
Main contact:	Barbara Koroušić Seljak, barbara.korousic@ijs.si

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

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.

Document Control Information	
Title	*D3.2 Methodology for FNS data standardisation and interoperability*
Editor	*Prof. dr. Barbara Koroušić Seljak (JSI), Adela Nacu (RTDS)*
Contributors	*Tome Eftimov (JSI), Panče Panov(JSI), Blaž Škrlj (JSI), Duccio Cavalieri (Unifi), Francesco Vitali (Unifi), Giovanni Bacci (Unifi), Karl Presser (PMT), Agnieszka Matuszczak (PMT), Julia Kurps (HYVE), Chris Evelo (UM), Susan Coort (UM)*
Reviewer(s)	*Julia Kurps (HYVE), Irina Stoyanova (SF)*
Dissemination Level	☐ **CO** Confidential ☒ **PU** Public
Approved by	☒ RTDS (COO) ☒ QIB (SCO) ☒ JSI ☒ UCD ☒ PMT ☒ JDLC ☒ EuroFIR ☒ UWTSD ☒ DTU ☒ ENEA ☒ HYVE ☒ HYLO ☒ UM ☒ NUTRIS ☒ RIVM ☒ WUR ☒ UGent ☒ IMDEA ☒ HUA ☒ TUM ☒ GS1 ☒ SF ☒ UoR ☒ IFA ☒ ILSI ☒ BfR ☒ AUTH ☒ FEM ☒ CNR ☒ APRE ☒ CAP ☒ UNIFI ☒ LIFE ☒ Nutritics ☒ EFF
IPRs underlined	Not applicable
Datasets underlined	Not applicable

Version/Date	***Change/Comment***
0.1_2020-05-12	*Draft outline prepared by JSI, Unifi, UM, PMT*
0.2_2020-07-01	*First version prepared by JSI, Unifi, UM*
0.3_2020-08-05	*Second version prepared by JSI, Unifi, UM, PMT*
0.4_2020-09-01	*Third version prepared by JSI, Unifi, UM, PMT, HYVE*
0.5_2020-09-25	*Final version prepared by JSI, Unifi, UM, PMT, HYVE*
1.0_2020-09-30	*COO edits, final version submitted to EC*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

1. Publishable Summary

FNS data is complex because it is heterogeneous in both data types and data formats. In order to enable standardisation and interoperability of such data, advanced services based on natural language processing and machine learning are required. In D3.2, methodology for data normalisation, annotation, matching, analysis and visualisation is described from the theoretical perspectives. All the methods have been evaluated and have the technological readiness level of at least TRL4. Results of the methods' evaluations have been published in peer-reviewed papers, which proves their reliability. In the selection and development of the methodology, the needs of the FNS project, especially of the use cases and demonstrators, were considered. Moreover, existing resources and tools have been identified and included.

One of the main outputs of the FNS project is the FNS ontology that is discussed in D3.2 as well. The FNS-Harmony ontology is introduced as a new resource that will allow us to harmonize and integrate different reference vocabularies and ontologies from different sub areas of food and nutrition, as well as ontologies representing the domain of data analysis. This is also very important in designing new FNS-oriented studies, where selection of metadata should be based on such an ontology.

Finally, FAIR assessment frameworks are reviewed and guidelines for the selection or development of a framework most suitable for FNS Cloud are provided.

Abbreviations

API	Application Programming Interface
BFO	Basic Formal Ontology
CDK	Chemistry Development Kit
CPCat	Chemical and Product Categories
CV	Cross Validation
DCAT	Data Catalog Vocabulary
DL	Deep Learning
DNN	Deep Neural Networks
EFSA	European Food Safety Agency
FAIR	Findable, Accessible, Interoperable, Reusable
FAO	Food and Agriculture Organization
FNS	Food, nutrition, security
IAO	Information Artefact Ontology
IE	Information Extraction
ML	Machine Learning
NCBO	National Center for Biomedical Ontology
NER	Named-Entity Recognition
NLP	Natural Language Processing
OBI	Ontology of Biomedical Investigations
OGTT	Oral Glucose Tolerance Test
ONS	Ontology for Nutritional Studies
OxO	Ontology Xref Service
RDA	Research Data Alliance
SCFA	Short-Chain Fatty Acids
UMLS	Unified Medical Language System

2. Introduction

In FNS-Cloud, services for supporting standardisation and interoperability of data on food, nutrition and security (FNS data), that is being gathered from different sources, are developed. These services include FNS data pre-processing (to be developed in Task 3.2), curation and annotation (Task 3.3), matching (Task 3.4), analysis (Task 3.5), and visualisation (Task 3.6). In this deliverable, each service is formally specified to be implemented later in Task 3.7 and integrated with the FNS Cloud in Task T3.8.

2.1. Background

FNS data are heterogeneous in data types and formats. In Figure 1, the diversity of FNS *data types* is presented (Figure from D2.1).

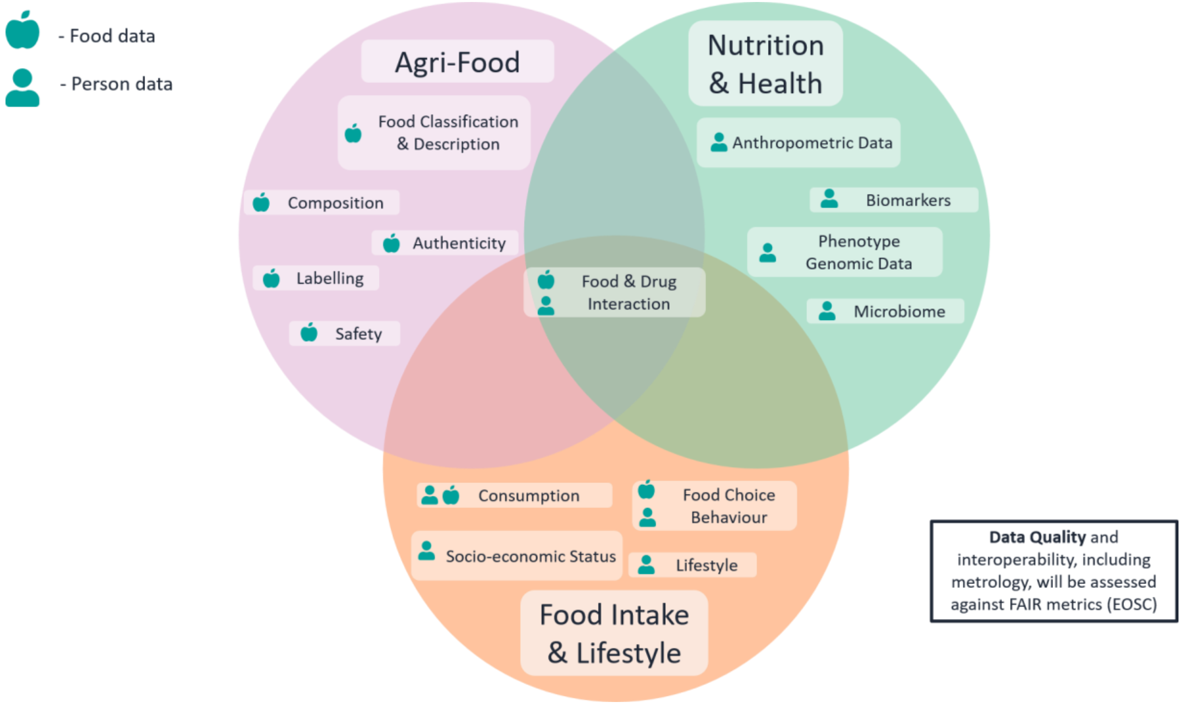


Figure 1. The FNS data types.

The inventory of existing FNS data (FNS inventory, 2020) showed that FNS data elements from both public and proprietary data sources are stored in heterogeneous *data formats*, ranging from simple files to structured databases that are often application-specific. For example, scientific literature, images and other textual data are stored in unstructured or semi-structured formats (plain text files, HTML or XML files, binary files). Genomic and other "-omic" data are stored in spreadsheets as well as in structured relational databases.

Since FNS data is so complex, it is difficult to be integrated. A lot of research and development is still needed to achieve an efficient integration and FAIRness of FNS data. Once FNS data is standardised and interoperable, it will be easier to be found, accessed and reused. Moreover, automation of many data lifecycle processes (such as data quality, data analysis etc.) will be enabled.

FNS data is more or less useless without its *metadata*. According to the Dublin Core definition (Dublin Core, 2020), metadata (literally "data about data") is structured data about anything that can be named, such as scientific literature (e.g. books, journal articles), images, songs, products, processes, people and their activities, research data, concepts, and services. Metadata following specific rules for its creation and publication, such as the Dublin Core principles (Dublin Core principles, 2020), enables data interoperability on the basis of Semantic Web or Linked Data principles.

In this deliverable, first, we define its objectives and describe the target audience. Then, we provide a description of the aims, related work, methodology and technical specification for each type of the FNS services. Since some semantic resources and tools for selected FNS data lifecycle processes have already been developed, their review is part of this deliverable as well. Finally, we conclude presenting an approach for measuring the FAIRness of a digital object.

2.2. Objectives

The main objectives of this document are to

- describe the tasks of FNS data pre-processing, curation and annotation, matching, analysis, and visualisation;
- provide technical specifications for the FNS Cloud developers.

2.3. Target audience

The target audience include

- developers of the FNS Cloud (from PMT, HYVE and SF), and
- end users of the FNS Cloud:
    - data providers
    - data users
- trainers (WP6)

3. FNS data pre-processing

3.1. Task definition

The analysis performed in Task 3.1 showed that FNS data is heterogeneous, meaning that the FNS Cloud needs to deal with high variability of data types and formats. In general, the FNS data types include structured, semi-structured and unstructured data (for more details please see the deliverables D2.1 and D3.1).

Whenever, a new data set will be provided to the FNS Cloud for the integration with other FNS data, first, FNS data will be classified as quantitative/ numerical, qualitative/ normative, structured/ unstructured, etc. Then, unstructured data will be structured using Natural Language Processing (NLP) and Machine Learning (ML) methods. For example, textual data can be structured using NLP-based methods, while image data can be processed using Deep Learning (DL) methods.

Each data set to be integrated with other FNS data will be pre-processed with respect to the selected methodology for data handling (e.g. for data analysis or data visualization). In general, data preprocessing includes cleaning, instance selection, normalization, transformation, feature extraction and selection, etc. Details related to each methodology will be provided in the following sections, however, in this section our approaches focused on structuring textual data and images are presented.

3.2. Related work

As great amounts of FNS-related information is presented in the form of heterogeneous textual data, computer-based methods are useful to automatically extract such information. One way to do this is to utilize Named-Entity Recognition (NER) methods (Nadeau and Sekine, 2007) that are broadly used in computer science for information extraction (IE) (Sarawagi, 2008) Despite the existence of numerous and well-versed NER methods in the biomedical domain, the domain of food science still remains scarcely resourced.

Extracting food entities from textual data is a challenging task that can be used for filling in the gaps in many practical applications. Automatically extracting food entities as well as all other biomedical entities (i.e. drugs, diseases, treatments, etc.) from scientific papers can help us to follow the knowledge that comes with each new publication. Additionally, this is useful for analyzing relations that exist between these entities.

Moreover, machine learning (ML) algorithms can be used to find some hidden (i.e. unknown) relations that exist between food entities and disease entities. This is especially important for food allergy studies. Additionally, automated extraction of food information can be used to fill in missing values that appear in food-related databases (e.g., food composition databases). Another interesting application is where information extraction is used to extract food entities from dietary records for individuals (i.e. written as free-form text), and then map them on a nutrient level. This information can be combined and used by recommender systems.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863059.

IE is the task of automatically extracting information from unstructured data and, in most cases, is concerned with the processing of natural language texts. The goal of IE is to provide a structured representation of the extracted information captured from the analysed text.

The information to be extracted is contained within the texts themselves. It consists of predefined entities of interest, as well as relationships between the entities that are typical for some domain. For example, users may be interested in extracting information about food entities, nutrient entities, quantity/unit entities, population groups, together with the relations between them. Let us assume that we have the dietary recommendation "*Babies need about 10g protein a day*". Using an IE method, the extracted information should be "*Babies*" as a population group, "*protein"* as a nutrient entity, "*10g a day*" as a quantity/unit entity, and "*need*" as the relation between the population group and the nutrient entity.

One well known IE task is named-entity recognition (NER), which addresses the problem of identification and classification of predefined concepts in a given domain. It aims to identify words or phrases from the text and then label them into predefined classes (labels) that describe concepts of interest in a given domain.

Various NER methods exist: terminology-driven, rule-based, corpus-based, methods based on active learning (AL), and methods based on deep neural networks (DNNs). Additionally, hybrid approaches for named-entity recognition from unstructured textual data exist.

In recent years, numerous NER methods have been developed for the biomedical domain (Boag et al, 2015), which are available together with comparison studies on different benchmark data sets (Jagannatha et al, 2019). For example, QuickUMLS (Soldaini and Goharian, 2016) is a fast-unsupervised technique for medical concept extraction. Clinical Named Entity Recognition system (CliNER) is a named entity recognition method that can be used for extracting clinical entities from electronic health records. However, to the best of our knowledge, there are very few food-named entity recognition methods and no comprehensive comparative studies between them. It is important to mention that it is very challenging to transfer any existing NER method to a different domain. This is the case because many of them are trained on data from a specific domain (i.e. in the case of corpus-based NERs), or they are based on rules that are crucial for one domain, but not important to other domains.

3.3. FNS-Cloud methodology

At JSI, we developed a new food NER system - FoodIE (Popovski et al, 2019), which consists of a rule engine. These rules are based on computational linguistics and semantic information which describe each food concept. Unlike USAS, FoodIE considers word chunking when extracting and annotating the food concepts, i.e. multiple words (tokens) can be grouped into a single food concept. The evaluation of this method has been performed using two independent benchmark data sets. The first one consists of 200 recipes extracted from Allrecipes (Groves, 2013) including recipes from five categories: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks. From each recipe category 40 recipes were included in the first benchmark data set. The second benchmark data set consists of 1000 new recipes also extracted from the Allrecipes website and consists of 200 recipes from each recipe category. After extracting the food entities, one human expert manually selected what should be extracted while another human expert compared the results with what was extracted using FoodIE. The evaluation that was done

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

using the two different data sets showed that FoodIE's behaviour is consistent, as well as that it achieves very promising evaluation results.

Further, we have compared the FoodIE results using the NCBO annotator, which is a web framework that extracts and annotates food concepts from free-form text provided by the user. The domain of the concepts and the performance of the annotator depend on the ontology that is chosen to be used in the background. Accordingly, it is able to extract and annotate only the concepts that are present in the ontology. For this reason, any combination of NCBO and an ontology can be assumed as a different NER method. Its annotation workflow is centered around a highly efficient syntactic concept recognition engine and a set of semantic expansion algorithms. The NCBO annotator (Noy et al, 2009) is available within the BioPortal software services and is able to use ontologies that are available there as well. In our case, we used the NCBO annotator in a combination with three ontologies FoodOn (Dooley et al, 2018), OntoFood, and SNOMED CT (Donnelly et al, 2006).

To evaluate the results of the NER methods we used the recently published FoodBase corpus (Popovski et al, 2019), which consists of 1000 recipes annotated with food concepts. For each evaluation metric (precision, recall, and F1 score), FoodIE outperforms the remaining three NER methods. Specifically, F1 Scores are: FoodIE (96.05%), SNOMED CT (63.75%), OntoFood (32.62%), and FoodON (63.90%). The metric value is on a scale from 0 to 1, we express it as percentages, as is standard. It is apparent that FoodIE has quite a notable advantage over the other three NER methods, as the absolute differences for the F1 Score between FoodIE and the remaining three NER methods are: 32.30%, 63.43%, and 32.15%, respectively. The NER method with the worst evaluation metrics is NCBO (OntoFood), which also gives us an indication that the OntoFood ontology does not cover the food domain adequately, i.e., many food entities are not present in it.

Evaluating four different NER methods in the food domain: FoodIE, NCBO (SNOMED CT), NCBO (OntoFood), and NCBO (FoodON) on a data set of 1,000 manually annotated recipes, it is evident that FoodIE provides more promising results for each individual evaluation metric, as well as the best overall result (Popovski et al, 2020).

Additionally, extracting food entities can further be linked with entities from other domains, such as health, bioinformatics, consumer and social sciences etc. This can help in reducing knowledge gaps that inhibit public health goals as well as the optimal development of scientific, agricultural and industrial policies.

Another important but also challenging source of FNS-related information presents images of food and drinks. In order to support information extraction from this kind of unstructured data, we developed a method for automated recognition of food and drinks from images. The method is based on a novel deep learning architecture for food image recognition, named NutriNet. To train a model using this architecture, freely available images were downloaded from the Internet and organised into appropriate food classes. At the 13th European Nutrition Conference (FENS 2019), this work was selected as one of three best Assessment and novel techniques. The pixel-level recognition approach of NutriNet seems promising for recognising food images, as was shown in recent Food Recognition Challenge (https://www.aicrowd.com/challenges/food-recognition-challenge), where the approach achieved the second place.

3.4. Technical specifications

FoodIE will be embedded in other FNS services (see Sections 4.3 and 5.3 for further details).

4. FNS data curation and annotation

4.1. Task definition

To support FNS data discovery and reuse we will consolidate existing methodologies and open frameworks for facilitating curation and annotation with community standards (e.g. ISA) establishing metadata services to annotate FNS data according to FAIR principles. Metadata provenance also needs to be considered, meaning data and descriptors must be packed into 'containers' (pre-processed) for T3.4-T3.6 (processing). Pre-processing not only structures data, but also extracts data semantics for subsequent annotation. Information about entities and concepts recognised in the processed data will be structured hierarchically and stored in an ontology, together with their relationships, using a semi-automatic method developed by JSI in QuaLiFY (FP7) and refined in RICHFIELDS (H2020). Extracted knowledge will be reviewed systematically by experts using Protege (Stanford University, US) and user communities engaged via a feedback loop to consensus. The FNS-Cloud ontology will also link with other FNS domain ontologies (e.g. ENPADASI taxonomy and the ontology, ONS [Ontology for Nutritional Studies]; Pathway; Gene Ontology; Disease Ontology; etc) and WP3 will explore linking with reference networks for co-expression, different kinds of regulations (transcription factors, miRNAs), protein-protein and cell-cell interactions, as well as, with existing open-source platforms (e.g. Disqover).

4.2. Related work

Many research questions from different domains involve combining different data sets in order to explore a research hypothesis. One of the main problems that arises here is that different data sets are structured with respect to different domain standards and ensuring their interoperability is a time-consuming task. In the biomedical domain, the Unified Medical Language System supports interoperability between biomedical data sets by providing semantic resources and Natural Language Processing tools for automatic annotation. This allows users to understand the links between different biomedical standards. While there are extensive resources available for the biomedical domain, the food and nutrition domain is relatively low-resourced.

Data normalization is a crucial task that allows interoperability between data sets that are described using different standards. By applying data normalization, we are mapping entities between different vocabularies, standards, or semantic resources.

In the food domain, there are several resources that can be used for food data normalization. Some of them are:

- FoodEx2 - a description and classification system, proposed by the European Food Safety Agency (EFSA) (Ioannidou et al, 2019)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

- FoodOn - provides semantics for food safety, food security, agricultural and animal husbandry practices linked to food production, culinary, nutritional and chemical ingredients and processes.
- OntoFood - a nutrition ontology for diabetes.
- SNOMED CT - a standardized, multilingual vocabulary of clinical terminology that is used by physicians and other health care providers for the electronic health records. It also consists of a Food concept.
- The Hansard corpus - a collection of text and concepts created as a part of the SAMUELS project (Alexander and Anderson, 2012). It consists of 37 higher level semantic groups; one of them is Food and Drink.

Hand in hand with the development of such resources, the evidence that a universal and comprehensive data model for coherent metadata annotation of scientific study arose. Among those, the open source ISA framework (Sansone et al, 2012) is one of the most widely accepted and known standards for metadata annotation. The framework includes three main blocks that are used to provide a full description of any biological experiment. The first block (called "Investigation") describes the overall context of the experiment (such as: title, metadata, and a general description) and links all related study blocks that are used to characterize the biological entities used in the experiment. The second block is indeed called "Study" and records the characteristics of biological samples used in the experiment, starting from raw material (also called "source material") and describing all the collection and preparation steps performed before the analysis. Steps are reported as directed acyclic graphs with source material nodes followed by sample collection (and/or processing) nodes which in turn are connected to sample material nodes. The third block is called "Assay" and is used to describe all procedures which led to the production of quantitative and/or qualitative data used in the final work. The subject of an assay could be a single sample but also a group of samples that have been used to produce some kind of data as a whole. Assayes are recorded in a similar fashion as Studies: a directed acyclic graph is produced, starting from a sample material node (the last node in the Study block) and ending to the production of raw data (data nodes). Central nodes in an assay block are used to record all the steps that led to the production of the final data, including machines and platforms used to process samples (e.g. sequencing platforms but also DNA quantification methods and so on). The ISA framework also provides more general terms that can be used to describe an experiment. The source material nodes, for example, should be annotated using ontology annotations that can help researchers to unequivocally understand the context of the experiment and the type of material collected.

In the domain of data analysis, at JSI we have been working for a long time on the development of ontologies used for representation of the core entities in the process of data analysis. **In this context, we have developed a generic ontology of data types (OntoDT) (Panov et al, 2016), which is based on ISO standards for data types frequently occuring in computer systems.** The ontology has been developed in a generic fashion to allow easy extension in any domain of interest and hence representation of domain specific data types. For the domain of data mining, in the OntoDT ontology, we provided a set of domain specific data types needed for semantic annotation of data mining datasets. In addition, we have developed an ontology of core data mining entities (OntoDM-core) (Pance et al, 2014) that allows us to semantically annotate datasets, algorithms, implementations of algorithms, executions of algorithms, as well as different types of outputs of algorithms such as patterns and models. Finally, we have developed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

an ontology for describing data mining investigations (OntoDM-KDD) (Pance et al, 2013) that allows us to semantically annotate different data analysis scenarios, belonging to different parts of the KDD process such as domain understanding, data understanding, data preprocessing, data analysis, evaluation and deployment.

The developed resources were produced using ontology best practices based on OBO Foundry principles (Smith et al, 2007), as well as using the upper-level Basic Formal Ontology (BFO) (Arp et al, 2015) as a template and set of formally defined relations from Relations Ontology (RO) (Smith et al, 2005). In addition, all developed resources heavily reused classes and relations from Ontology of Biomedical Investigations (OBI) (Bandrowski et al, 2016) and Information Artefact Ontology (IAO) (IAO, 2020). All this allows us to extend and integrate the already developed resources with other domain resources (e.g., from the domain of food and nutrition) built on the same design principles, such as the FoodOn ontology, as well as Ontology for Nutrition Studies. This integration would allow us to semantically represent domain specific data types, datasets, analysis pipelines and knowledge discovery scenarios that are performed in the specific domains.

4.3. FNS-Cloud methodology

As stated in paragraph 4.2, the ISA framework represents an already widely accepted and very solid standard for metadata annotation. Nonetheless, at UNIFI, we have started an extensive bibliographic research to build a catalog of other standards for annotating metadata. The catalogue will be finally refined with reasoned annotations on general characteristics of each standards (i.e. accessibility, language, development team), number of citations of the resource, international resources/initiatives adopting the standard, suitability for cloud applications with or without further development by the FNS Consortium, and a general judgment of adequacy for the purposes of FNS Cloud. We foresee that the ISA framework would still come out as the most suitable resource, and further considerations included in this document will be centered on that standard. Anyhow, if this bibliographic research identifies another standard as better for our purpose (e.g. less development needed) the technical specification of FNS Cloud resource for metadata annotation and querying presented in paragraph 4.4, would remain the same.

The ISA framework standard has two major limitations:

1. *It is not cloud-ready.* Various tools and implementation fall under the ISA framework, but none of those seems really suited or ready for a completely cloud environment. The main application is in fact developed in Java, and we foresee that a web application would be more suited for the need of FNS Cloud.
2. *It is not easy or of immediate use.* We think that the ISA framework, to be used to the fullest, requires some form of staff training. Again, in its present form, the ISA framework seems not to be ideal for the FNS Cloud application.

In the face of these disadvantages, the ISA framework represents a data model that already covers most of the *desiderata* for a metadata annotation tool in FNS Cloud.

- Has a JSON format and schema specification defined and available (https://isa-specs.readthedocs.io/en/latest/isajson.html)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

- Has semantic application (formerly written in JAVA but now moving to Python) that converts an ISA-TAB to semantic OWL format (linkedISA; https://github.com/ISA-tools/linkedISA) or to RDF (ISA to RDF; https://github.com/ToxBank/isa2rdf)
- Has a form of web application and DB (https://github.com/ISA-tools/BioInvIndex)
- The complete suite is licenced under CC BY-SA 4.0

At JSI, we started with the design and implementation of FNS-Harmony ontology (FNS-H) (see Figure 2). We envisioned FNS-H as mid-level ontology having BFO as a template at the upper level. The FNS-H ontology would allow us to harmonize and integrate the different reference vocabularies and ontologies from different sub areas of food and nutrition, as well as ontologies representing the domain of data analysis. The design of FNS-H follows principles for building application ontologies and reuses as much as possible already developed reference ontology resources. In the first phase, we will integrate the FoodOn ontology and the ONS ontology, representing the domain of food and nutrition, with the OntoDM suite of ontologies, representing the domain of data analysis. With this integration, we will be able to (1) define domain specific data types for the domain of food and nutrition by extending OntoDT generic data types; (2) define food and nutrition analysis pipelines for the domain of food and nutrition by extending OntoDM-core; and (3) define food and nutrition knowledge discovery scenarios by extending OntoDM-KDD ontology.

The development of the FNS-H ontology started in a top-down fashion. The ontology is expressed in OWL2 ontology language (OWL2, 2020) and developed in the Protege ontology development tool (Protege, 2020). In order to implement the ontology first, we performed in depth analysis of two deliverables, which are outputs of the FNS-Cloud project: D2.1 Definition of data models and APIs (v1.0) and D3.1 Data requirements and applicability criteria (v1.0). From these two deliverables, we extracted the initial set of higher-level terms describing different FNS data domains, domain data types, data formats, data provenance metadata, lists of external ontologies and vocabularies used for describing the domain of food and nutrition, as well as listings of FNS dataset instances identified by project partners (See Figure 3). The FNS-H ontology will follow well established naming conventions for all defined classes, relations and instances, and will be available for download via the project internal/public web pages (in the development phase), a GitHub repository (https://github.com/panovp/FNS-Harmony), as well as via BioPortal.

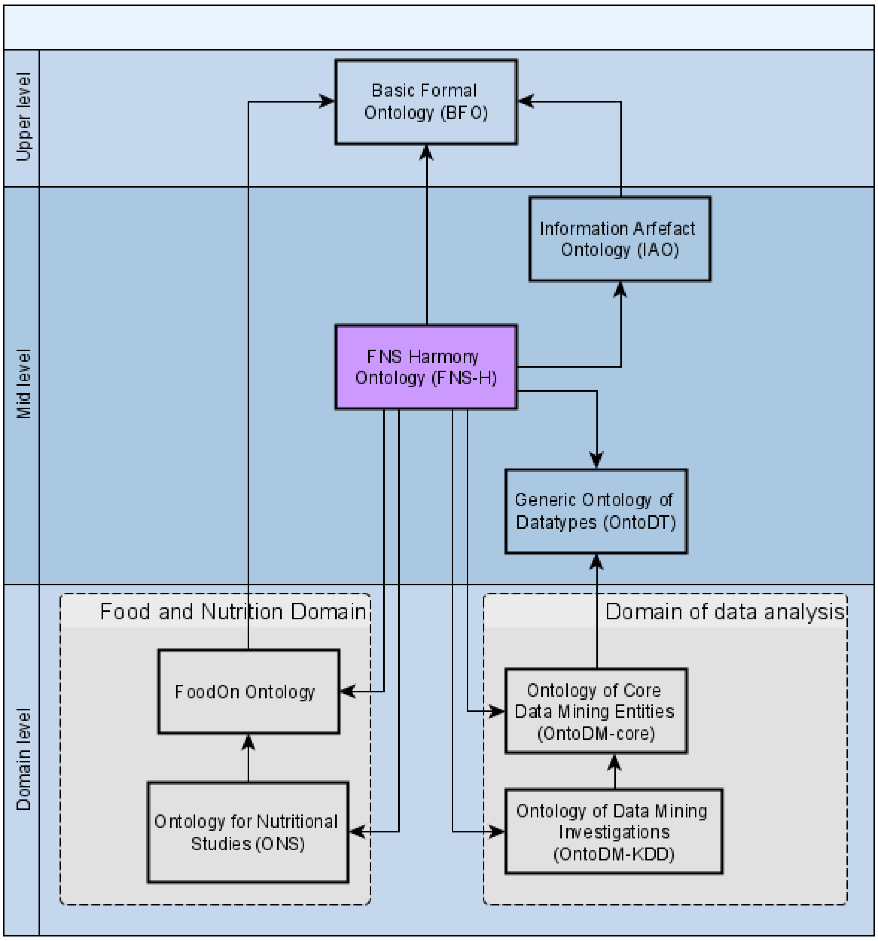
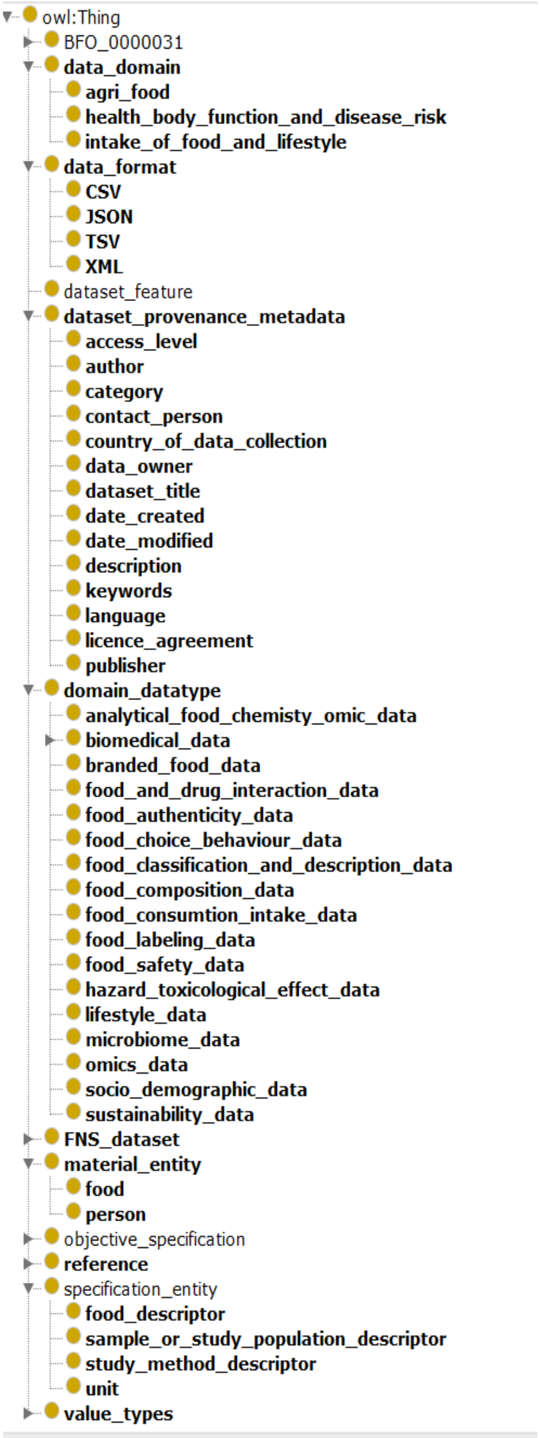

Upper level
Mid level
Domain level
Basic Formal Ontology (BFO)
Information Arfefact Ontology (IAO)
FNS Harmony Ontology (FNS-H)
Generic Ontology of Datatypes (OntoDT)
Food and Nutrition Domain
Domain of data analysis
FoodOn Ontology
Ontology for Nutritional Studies (ONS)
Ontology of Core Data Mining Entities (OntoDM-core)
Ontology of Data Mining Investigations (OntoDM-KDD)

Figure 2. An initial design of FNS-Harmony (FNS-H) ontology

In the next steps, we will first align the extracted terms with the BFO ontology and then integrate them with domain terms from the domain ontologies based on BFO, such asFoodOn and ONS, at the first instance, as well as with OntoDM set of ontologies. Other potentially relevant ontologies include the Ontology for Biomedical Investigations (OBI), Ontology of Biological and Clinical Statistics (OBSC), Ontology of Chemical Entities of Biological Interest (ChEBI), Ontology of Statistical Methods (STATO) and others. This integration is possible because all the above-mentioned ontologies are developed following the same set of design principles and best practices (defined by the OBO foundry principes) and use the same top-level classes originating from the BFO ontology. The aim is to reuse as much already developed resources as possible, and with this reduce duplication of modeling efforts. To achieve integration of different ontological resources, we will use the ROBOT (Jackson et al, 2019; ROBOT, 2020) tool that supports automation of a large number of ontology development tasks and helps developers to efficiently create and release high-quality ontologies. The further development of the ontology will be guided by the needs of the project, more specifically the cloud platform, as well as the project use-cases and demonstrators (e.g. Use Case on Gut microbiome).



- owl:Thing
  - BFO_0000031
  - data_domain
    - agri_food
    - health_body_function_and_disease_risk
    - intake_of_food_and_lifestyle
  - data_format
    - CSV
    - JSON
    - TSV
    - XML
  - dataset_feature
  - dataset_provenance_metadata
    - access_level
    - author
    - category
    - contact_person
    - country_of_data_collection
    - data_owner
    - dataset_title
    - date_created
    - date_modified
    - description
    - keywords
    - language
    - licence_agreement
    - publisher
  - domain_datatype
    - analytical_food_chemisty_omic_data
    - biomedical_data
    - branded_food_data
    - food_and_drug_interaction_data
    - food_authenticity_data
    - food_choice_behaviour_data
    - food_classification_and_description_data
    - food_composition_data
    - food_consumtion_intake_data
    - food_labeling_data
    - food_safety_data
    - hazard_toxicological_effect_data
    - lifestyle_data
    - microbiome_data
    - omics_data
    - socio_demographic_data
    - sustainability_data
  - FNS_dataset
  - material_entity
    - food
    - person
  - objective_specification
  - reference
  - specification_entity
    - food_descriptor
    - sample_or_study_population_descriptor
    - study_method_descriptor
    - unit
  - value_types

Figure 3. An initial structure of the FNS-Harmony (FNS-H) ontology. The initial ontology contains sets of terms collected from D2.1 and D3.1.

To make the links between different food standards understandable by food subject matter experts and to make them familiar with the interoperability process using different standards, we developed FoodViz (Stoyanov et al, 2020), which is a web-based framework used to present food annotation results from existing Natural Language Processing and Machine Learning pipelines in conjunction with different food semantic data resources. Currently, a lot of work

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

can already be done in an automatic way, but it is very important that the results are presented to experts in a concise way so that they can check and approve (or disapprove) the results. To show the utility of FoodViz, we visualize the results that are already published in the FoodOntoMap resource. The results consist of recipes that are coming from the curated and uncurated version of FoodBase, which was constructed by using the food NER method FoodIE.

There exist three different parts that can be explored: NER tagger, Documents, and Test custom document. The Documents part displays the curated and uncurated recipes of the FoodBase corpus. There are 1000 curated recipes, 200 per each recipe category, and more than 22.000 uncurated recipes available in FoodViz. The curated version is a ground truth data set, because in the process of developing it, the missing food entities were manually included, while the false positive entities were manually excluded from the corpus (Popovski et al, 2019).

FoodViz allows users to filter the recipes by name, by the recipe category and between the curated and uncurated recipes. Next, the user can select a recipe, for which the semantic annotations are shown. For each extracted food entity the synonyms are presented, which are the food names available in different food semantic resources, followed by the semantic tags from Hansard corpus, FoodOn, SNOMED CT, and OntoFood. Additionally, users can further explore the semantic tags from the FoodOn, SNOMED CT and OntoFood, which are linked to their original semantic definitions.

The uncurated version of FoodBase does not include the false negatives entities and does not exclude the false positive food entities, since it is created from a collection of around 22,000 recipes. With this, subject matter experts can help the process of annotations, by removing the false positives, and including the false negatives, or the FoodViz tool can be also used as an annotation tool. By applying this, we will be able to create a much bigger annotated corpus that will allow training on more robust NER based on deep neural networks. Therefore, FoodViz allows manual removal of the false positives, and adding of the false negatives.

In addition, to allow automatic annotations, at JSI, we introduced FoodNER, which is the first corpus-based food NERmethod. It consists of 15 different models obtained by fine-tuning the three pre-trained BERT models on five groups of semantic resources. The models are aimed at predicting the following: food entity distinction, two subsets of Hansard semantic tags, FoodOn semantic tags, or SNOMED CT semantic tags. All models provide very promising results obtaining 95-98% weighted F1 score, which represents the new state-of-the-art in food Information Extraction. To bring our work closer to subject matter experts from the food domain, we provide have integrated all the fine-tuned models in the FoodViz platform (http://foodviz.ds4food.ijs.si/fbw/\#/predict). This interface radically simplifies the usage of the state-of-the-art models for subject matter experts in the food domain, without their knowledge of the underlying details, such as Machine Learning or IOB format understanding. Additionally, the current architecture of the FoodViz application allows integration of new prediction models only with their upload at the corresponding location at the server.

4.4. Technical specifications

At the moment, no resource was developed for metadata annotation and query inside the FNSCloud. At UNIFI we are developing the basic idea of a resource for metadata annotation and query (to be implemented by the help of SF, in Task 3.7, see Figure 4). The main identified needs are:

1. Ease of use for all the FNS-Cloud consortium. We aim at developing resources which need no or minimal training to be used for metadata annotation and for metadata query. The FNSCloud user base is, in fact, very varied as a background and we want the potential to reach everyone
2. Semantic annotation of metadata. The developed resource will certainly make extensive use of metadata fields that are not free text, but bound to come from an ontology. The user should be forced to choose metadata from classes defined in a set of well-determined ontologies (i.e. OBI, OntoDM, ONS, FOODON) as many times as possible. This should be possible leveraging on the BIOPORTAL APIs (https://bioportal.bioontology.org/ and http://data.bioontology.org/documentation) or the Ontology Lookup Service APIs (https://www.ebi.ac.uk/ols/index and https://www.ebi.ac.uk/ols/docs/api)
3. Lightweight annotation. We imagine a resource that, using the user input as above depicted, is capable of producing a single metadata file (likely JSON or XML format) which accompany the raw data at each step of the FNSCloud infrastructure.

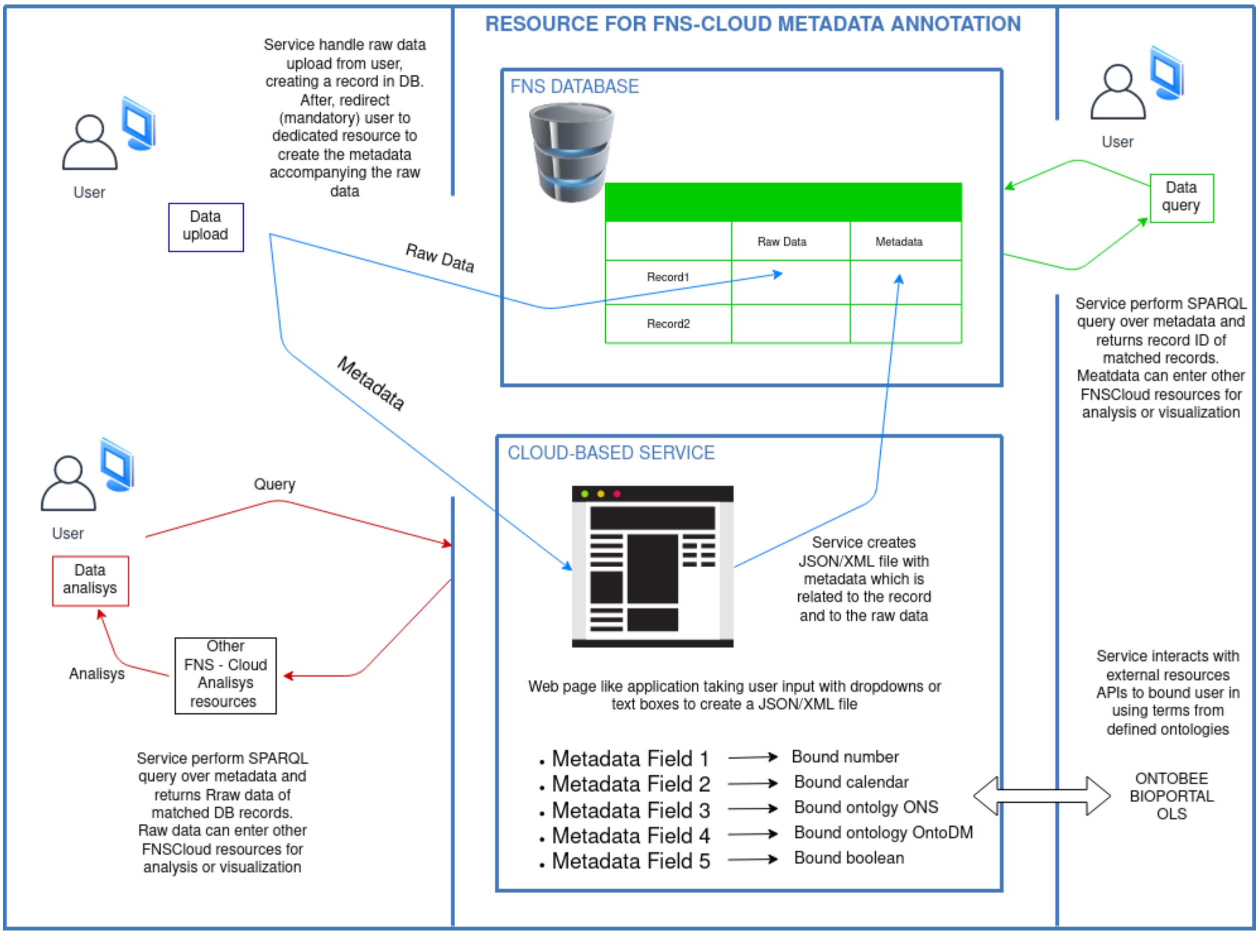

RESOURCE FOR FNS-CLOUD METADATA ANNOTATION
Service handle raw data upload from user, creating a record in DB. After, redirect (mandatory) user to dedicated resource to create the metadata accompanying the raw data
User
Data upload
Raw Data
Metadata
FNS DATABASE
Raw Data
Metadata
Record1
Record2
User
Data query
Service perform SPARQL query over metadata and returns record ID of matched records. Meatdata can enter other FNSCloud resources for analysis or visualization
User
Query
Data analisys
Analisys
Other FNS - Cloud Analisys resources
Service perform SPARQL query over metadata and returns Rraw data of matched DB records. Raw data can enter other FNSCloud resources for analysis or visualization
CLOUD-BASED SERVICE
Service creates JSON/XML file with metadata which is related to the record and to the raw data
Web page like application taking user input with dropdowns or text boxes to create a JSON/XML file
• Metadata Field 1
• Metadata Field 2
• Metadata Field 3
• Metadata Field 4
• Metadata Field 5
Bound number
Bound calendar
Bound ontolgy ONS
Bound ontology OntoDM
Bound boolean
Service interacts with external resources APIs to bound user in using terms from defined ontologies
ONTOBEE
BIOPORTAL
OLS

Figure 4. An initial design for FNS resource for metadata annotation and query

The FoodViz (http://foodviz.ds4food.ijs.si/fbw/\#/recipes) (developed by JSI) is a single page application using React (https://reactjs.org/), served by a back-end application programming interface (API) developed in Flask (https://flask.palletsprojects.com/en/1.1.x/). The back-end API serves pre-processed recipes annotated in our previous work FoodBase and the annotation mappings from FoodIE.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863059.

Development of the ISA framework and its integration with the FNS Cloud as well as integration of FoodViz with the FNS Cloud will be implemented in the following project period (to be described in D3.3 and D3.4).

5. FNS data matching

5.1. Background

FNS data is not only of different types but is also described and classified in various ways, using different standardised systems (e.g. LanguaL, FoodEx2, etc.) or local systems. One of the tasks is to find a way of matching data that is described and classified using different systems.

5.2. Related work

In this subsection, results of a review of existing computational methods for matching heterogeneous data (e.g. optimisation methods, ontology mapping, ML, etc.) are presented:

- **StandFood (Eftimov et al, 2017) - Standardization of Foods Using a Semi-Automatic System for Classifying and Describing Foods According to FoodEx2:** In recent years, food concepts normalization has become an open research question that is highly researched by the food and nutrition science community, calling it food matching. For this reason, StandFood was recently introduced, which is a semi-automatic system for classifying and describing foods according to a description and classification system, such as FoodEx2, proposed by the European Food Safety Agency (EFSA). It consists of three parts. The first involves a machine learning approach and classifies foods into four categories, with two for single foods: raw (r) and derivatives (d), and two for composite foods: simple (s) and aggregated (c). The second uses a natural language processing approach and probability theory to perform food concepts normalization. The third combines the result from the first and the second part by defining post-processing rules in order to improve the result for the classification part. However, the food normalization process was based only on lexical similarity between the food concepts names, avoiding the semantic similarity between them.
- **EMBL-EBI Ontology Xref Service (OxO) for cross-ontology mappings:** OxO is a service for finding mappings (or cross-references) between terms from ontologies, vocabularies and coding standards. OxO imports mappings from a variety of sources including the Ontology Lookup Service and a subset of mappings provided by the UMLS. We're still developing the service so please get in touch if you have any feedback. https://www.ebi.ac.uk/spot/oxo/.
- **BridgeDb** is a framework to map identifiers between various biological databases. These mappings are provided for genes, proteins, genetic variants, metabolites, and metabolic reactions. BridgeDb includes a Java library that provides an API for programmatic access. More information on how to use the BridgeDb framework for the identifier mapping can be found here: https://bridgedb.github.io/. At the moment several plugins are available see, https://bridgedb.github.io/pages/plugins.html

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

- **The Chemistry Development Kit (CDK)** is a collection of modular Java libraries for processing chemical information (Cheminformatics). The modules are free and open-source and are easy to integrate with other open-source or in-house projects. More information on how to implement the CDk can be found here: https://cdk.github.io/
- **AGROVOC** (AGROVOC, 2020) is a controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc. It is published by FAO and edited by a community of experts.
- **CPCat (Chemical and Product Categories)** is a database containing information mapping more than 43,000 chemicals to a set of terms categorizing their usage or function.

5.3. FNS-Cloud methodology

To provide a data set in which food concepts are normalized by semantic tags from different food ontologies, we selected 22,000 recipes from Allrecipes, which is the largest food-focused social network where people share their recipes and provide information about recipes in a non-standardized manner. The recipes were selected from five recipe categories: Appetizers and snacks, Breakfast and Lunch, Dessert, Dinner, and Drinks.

For this reason, at JSI, we developed the FoodOntoMap (Popovski eta al, 2019) data set consisting of food concepts extracted from the recipes and normalized to different food ontologies. It also provides a link between the food ontologies. For this reason, the food concepts are matched and for each food concept the semantic information from each dataset is assigned. The concept matching is done by iterating through each food concept that is extracted by the NER method FoodIE. If the concept is also recognized wholly or partially by the NCBO annotator in combination with any of the selected ontologies, the semantic tags from that ontology are also assigned to the food concept. However, it is not uncommon while using the NCBO annotator for it to provide semantic tags on a token level instead of on a concept level. Such an example would be when an ontology returns two outputs for the food concept "*salad dressing*", instead of classifying it as a single food concept consisting of two tokens. In these cases, each incomplete food concept extracted by the NCBO needs to be matched to its corresponding superset food concept extracted by FoodIE. This was done by checking if the location metrics from the NCBO annotator are in accordance (more specifically, whether they are a subrange) with the location metrics of a food concept provided by FoodIE. If such a match exists, the NCBO food concept is added to the mapping set of the FoodIE concept. By doing this, the mapping sets aggregate all the corresponding food concepts that map to a food concept, along with their semantic information, even if the NCBO annotator does not classify some food concepts wholly.

The results from the FoodOntoMap are four different datasets and one mapping. Each dataset consists of an artificial id for each unique food concept that is extracted using each approach, the name of the extracted food concept, and the semantic information assigned to it. Each dataset corresponds to one of the four semantic resources: Hansard corpus, FoodOn, OntoFood, and SNOMED CT. At the end there is one data set mapping, called FoodOntoMap, which, for each concept that appears at least in two datasets, provides the links between them by listing the artificial id of the concepts from each of the datasets in which it is mentioned. The datasets consist of 13,205; 1,069; 111; and 582 unique food concepts, obtained using Hansard corpus, FoodOn,

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

OntoFood, and SNOMED CT, respectively. The FoodOntoMap data set consists of 1,459 food concepts that are found in at least two food semantic resources.

To the best of our knowledge, FoodOntoMap is the first resource that provides normalization of food concepts to different food ontologies, additionally providing a link between them. The motivation for building such a resource in the food domain comes from the existence of the UMLS, which is extensively used in the biomedical domain. For example, the MRCONSO.RRF table that is a part of UMLS is used in a lot of semantic web applications because it can map the medical concepts to different biomedical standards and vocabularies. To make progress in analysing the large amount of data that is available in order to find these relations, resources for food concepts normalization are extremely valuable and welcome.

The main benefit of using FoodOntoMap is that the food concepts can be normalized by mapping them to a unified system. Furthermore, the semantic tags can be reused to find the non-linear relations that exist between the concepts in the vector space, by learning the embedding space. This can also be done together with some medical and environment concepts. Once the embedding space is learned, the embeddings can be used for predictive studies in order to explain the relations between human health, food systems, and the environment.

FoodOntoMap can be used as a resource that represents a normalized dataset of food concepts. Additionally, users can also follow the described pipeline of steps used to create the FoodOntoMap mapping in order to create their own new resource where the food concepts will be normalized. Both approaches, FoodIE and NCBO, used for food concepts extraction have already been well documented and evaluated. The ontologies that are used by the NCBO annotator have also been well documented and are easy to utilize. Furthermore, FoodOntoMap can be easily extended on additional recipes, as well as a wide variety of different ontologies. With this it can provide an ever-wider coverage of food concepts. Additionally, the FoodOntoMap pipeline can be used to normalize food concepts that exist in food consumption and food composition databases. Additionally, FoodViz annotation tool is used as a visualization tool for the FoodOntoMap resource.

5.4. Technical specifications

The resource FoodOntoMap is published and publicly available for download at https://doi.org/10.5281/zenodo.2635437 under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) licence. With this, we encourage users to further contribute to this resource and modify it as need be. ZENODO was the platform of choice, as it provided all the framework tools needed for such a dataset. Hosted along with the resource's datasets is a DCAT specification file, which briefly describes the structure of the resource.

Furthermore, the resource will be actively maintained and extended as new ontologies, annotators, NER methods and NLP methods become available. The goal is to keep the resource relevant and contemporary while also improving its domain coverage with the ever-improving NLP tools.

As there may be more datasets available at ZENODO relevant for the FNS community, we have decided to explore ways of connecting the FNS Cloud platform with ZENODO.

5.5. Integration of Zenodo in the FNS Cloud

Zenodo is a widely used web application funded by CERN. Core functionalities are file storage and metadata annotation to support open science through publication of research articles and data sets. In this regard Zenodo's functionalities are comparable with the core functionalities provided by Fairspace. Besides these similarities, there are a number of differences between both applications, primarily on the following aspects: (1) data model, (2) metadata customization and (3) access management/authorization. In Fairspace, a single, holistic data model is implemented to integrate all available data sets in multiple formats from multiple research fields. Metadata fields are customizable and therefore more flexible than the simpler and more limited set of available metadata fields in Zenodo. While metadata in Zenodo is centered around publication metadata (such as author, journal, research organisation), the metadata fields in Fairspace can be more focused on the content of the data sets. Since Zenodo's main aim is supporting open science and data sharing, both metadata and data are openly accessible. Both can also be marked as private, but the ultimate aim is open access. In comparison Fairspace provides a complex access and authorization system which enables granular access permissions, both based on data sets and users. The selection of each individual application or an integrative approach is highly dependent on the use case specified by key users and members of the FNS Cloud consortium.

Because of the overlapping functionalities, we suggest exploring possibilities to integrate Zenodo with Fairspace and/or other applications used by organisations that are participating in the FNS Cloud consortium. Such a setup would enable use cases in which researchers publish open access articles in Zenodo and link the closed-access underlying data sets in Fairspace. The alternative setup could be local deployment of a Zenodo instance at individual research organisations with an integration to existing data storage systems. Or even setting up a customized version utilizing the flexibility of Invenio, the platform that Zenodo is based on. Besides we can explore the integration of the existing FNS Cloud catalogue to directly upload data sets, that are specified in the catalogue, to Zenodo or to provide a direct link to Zenodo in the data catalogue entry for a specific data set. Those integrations can help improve the FNS Cloud compliance to the FAIR guidelines.

Specifications about the potential implementation of Zenodo in the ecosystem of the FNS Cloud will be further explored in WP2 (D2.3). Aspects that will be discussed are the general role of Zenodo in the FNS Cloud infrastructure, the integration of Zenodo with other FNS Cloud applications through APIs and advantages of centralized vs decentralized deployment.

5.6. Validation

Our food matching approach has been validated on FNS data collected for the aims of the WP4 Use Case 2 on Food labelling data and reformulation tools, where compositional data on branded foods was collected without metadata describing food groups. Applying the methods StandFood (described in Section 5.3) and FoodNER (described in Section 4.3), we matched food composition data about branded foods on FoodEx2 food groups (level 2). The food matching approach and the results will be presented in a joint paper submitted for publication in a peer-reviewed journal.

6. FNS data analysis

6.1. Background

To enable analysis of heterogeneous FNS data prepared and integrated in T3.2-T3.5, advanced methods will be provided by or through FNS-Cloud Services. T3.5 will explore existing methods from ML (i.e. ensembles of methods, meta-learning, supervised learning, DL), statistics, bioinformatics (e.g., pathway and network analysis) and information theory, and select the most suitable for analysis of multiple data sources. Using information fusion and analysis, synergies amongst the methods can be achieved, producing more accurate and useful information than an individual source. Deciphering the underlying biological processes can be performed by applying pathway and network analysis methods (e.g. metagenomics and metabolite data, WP4&5). These methods will be developed as Services (T3.7) or integrated through FNS-Cloud-linked tools and apps (e.g., Galaxy for sequencing, R for arrays).

6.2. Related work

There are two fields of interest: i) integration of data on metagenomics and metabolomics, and ii) metabolic prediction.

Integration of data on metagenomics and metabolomics (UM, Unifi, QIB, HYVE):

- **Data:** IBD data, QIB data
- **Methods:**
    - At UM, work on applying a workflow for metagenomics, metabolomics and proteomics data is performed (Microbiome analysis, 2020):
        - **Integration of the microbiome and host metabolome at pathway level:** Multispecies pathway creation and analysis in PathVisio (PathVisio, 2020). This will allow the identification of important biological processes. PathVisio is a software tool built by the bioinformatics groups at Maastricht University, for more information see https://pathvisio.github.io/.
        - **Network analysis in Cytoscape to create co-occurrence networks** using the CoNet app (CoNet, 2020). This will allow the identification of co-occurrence of microbial populations at the genus level and metabolites. In addition several Cyoscape apps, (co-)created at Maastricht University, can be used to extend the network with additional information, like pathway-gen interactions from WikiPathways.
        - UM: An overview of what has already been done: https://projects.bigcat.unimaas.nl/projects/.

Metabolic prediction (JSI, QIB, Unifi): With regard to the ML predictive tools, we will investigate the accessibility of data available from previous published work in metabolic prediction using ML-based algorithms, and using data provided by QIB we will develop new models (see section 6.3). Previous published work includes a randomized crossover trial with 20 healthy subjects

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

comparing the effects of traditionally milled and prepared whole-grain sourdough bread and industrial white bread made from refined wheat on multiple clinical and disease markers and on the composition and function of the gut microbiome was researched (Korem et al, 2017). The main research question was: "Can we predict from baseline measures in advance whether it is white or sourdough bread that will induce lower glycemic responses for each individual?" For this reason, using linear mixed models, the authors compared the treatment effects of a week-long consumption of white bread to that of sourdough bread on 20 clinical variables, using measurements taken as baseline (days 0, 21) and outcome (days 7, 28) of each clinical variable. **They found no significant difference** between the two treatments both for the primary outcome measure of this trial, glycemic control, quantified using the response to an oral glucose tolerance test (OGTT) and wake up glucose levels, and for 18 secondary outcome measures.

For this study, the authors employed a stochastic gradient boosting regression algorithm which was trained on the difference between the **average responses to white and sourdough bread divided by the average OGTT response**, and tested against the true classification to two categories of lower PPGRs to white or sourdough bread using leave-one-out cross validation (CV). The prediction used: (a) metagenomics-derived species (capped at 105), pathways and module abundances, of which, 6 non-sparse features were selected based on Pearson correlation to the target value in the training set of each CV fold; (b) four principal components of composition of genes, with principal component analysis calculated only on the training set of each CV fold; (c) the number of said genes present in the samples; and (d) mapping percentage to the human genome, the gene catalog, and a database of complete bacterial genomes."

6.3. FNS-Cloud methodology

We aim to integrate with the FNS Cloud existing tools relevant for the FNS community (specified by UM).

In addition, at JSI we are developing new **metabolic predictive tools (ML pipelines)** that will be trained on data provided by QIB. The ML tools are based on unsupervised and supervised learning, trying to explore the relations between microbiome and short-chain fatty acids (SCFAs). Further, representational learning techniques (Eftimov et al, 2020) are used to develop proper representation of the data (dimensionality reduction), to avoid overfitting of the ML tools. Transfer learning can be also applied in order to transfer the model to some other semantically similar data. The main issue for developing such ML pipelines is limited access to the data, which is mandatory for training ML models.

Further, non-synthetic knowledge graphs are becoming available throughout many disciplines of science, ranging from biology, physics, engineering to nutrition science. In recent years, the methodology capable of knowledge graph embedding is becoming a prevalent way to handle and process large knowledge graphs (Ji et al, 2020). As part of the proposed work we will explore how the existing, state-of-the-art knowledge graph methodology can be adopted to address tasks of potential relevance to food science. We identified two main tasks which were previously not addressed, namely the assessment of noisiness of the existing knowledge graphs, as well as their completion (Gesese et al, 2020). The two tasks will address the issue of inferring potentially interesting relations between existing entities, such as for example food types, food and other domain entities and assess whether the artefacts of knowledge (human errors) can be automatically identified and suggested to curators. The data that is explored is the FoodOn, which is a harmonized food ontology to increase global food traceability, quality control and data integration (Dooley et al, 2018).

6.4. Technical specifications

Existing and newly-developed tools will be integrated with the FNS-Cloud (to be presented in D3.3 and D3.4).

7. FNS data visualisation

7.1. Background

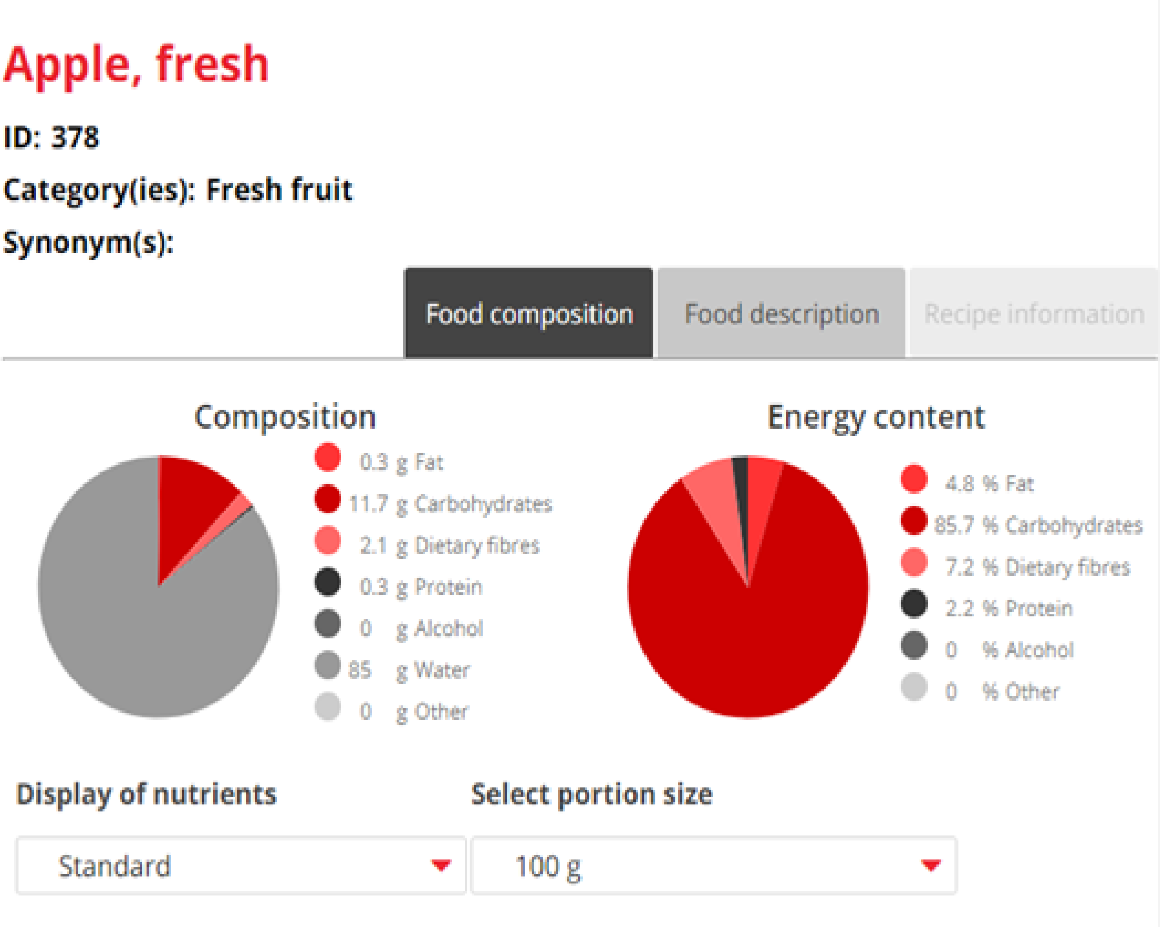
Data visualisation is the process to transform raw data into graphics, plots or figures to support scientists and users on visual literacy. Many different visualisation forms exist for different purposes. Studies show that users have different preferences when using visualisation forms and therefore a personal component is also involved. In some use cases in WP4, data visualisations are planned and appropriate solutions will be investigated.

7.2. Related work

Data visualisation is a complex and interconnected topic, with the complexity growing with different types of data and their size. The approach to visualising data can be either exploratory or explanatory (Data visualisation, 2020; EEA, 2020). Exploratory data visualisation can be a part of data analysis, when the user is not specifically familiar with the data(set) and would like to gain more knowledge or discover new dependencies by visualising the data and observing trends. The implementation of an exploratory data vis tool should allow flexibility to analyse the datasets from different angles and not miss any possible connections. Explanatory data visualisation might be easier in the development or actual visualising phase, but requires previous knowledge and understanding of the data. While designing the tool, it’s necessary to know what is to be shown and explained through the visualisation. The exploration of data(sets) should proceed it’s explanation and feed into it. Ideally the two should work in conjunction, where a few charts or diagrams generated in the exploratory phase can be used for later explanation of discovered trends or as input for a design of a visualisation tool or infographic for less experienced users with lower visual literacy. Visual literacy (as defined by the International Visual Literacy Association (IVLA, 2020) is a set of abilities that enables an individual to effectively find, interpret, evaluate, use, and create images and visual media. Visual literacy skills equip a learner to understand and analyze the contextual, cultural, ethical, aesthetic, intellectual, and technical components involved in the production and use of visual materials. A visually literate individual is both a critical consumer of visual media and a competent contributor to a body of shared knowledge and culture (Literacy Standards Task Force, 2011).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863059.

Visulab is an application to visualise multidimensional datasets. The idea behind Visulab is to allow users to explore multidimensional datasets with different graphs [Schmid, 1994, 1999] [Bürgi, 2004]. Scheuner used the tool to educate students and investigated how visualisation was used by students [Scheuner, 2014]. Inspired by these efforts, visualisation concepts for multi-dimensional datasets were developed and implemented in the food data management tool FoodCASE [Presser, 2018] and evolved in different tools around FoodCASE.



Apple, fresh

ID: 378

Category(ies): Fresh fruit

Synonym(s):

Composition

🔴	0.3 g	Fat
🔴	11.7 g	Carbohydrates
🔴	2.1 g	Dietary fibres
●	0.3 g	Protein
●	0 g	Alcohol
●	85 g	Water
●	0 g	Other

Energy content

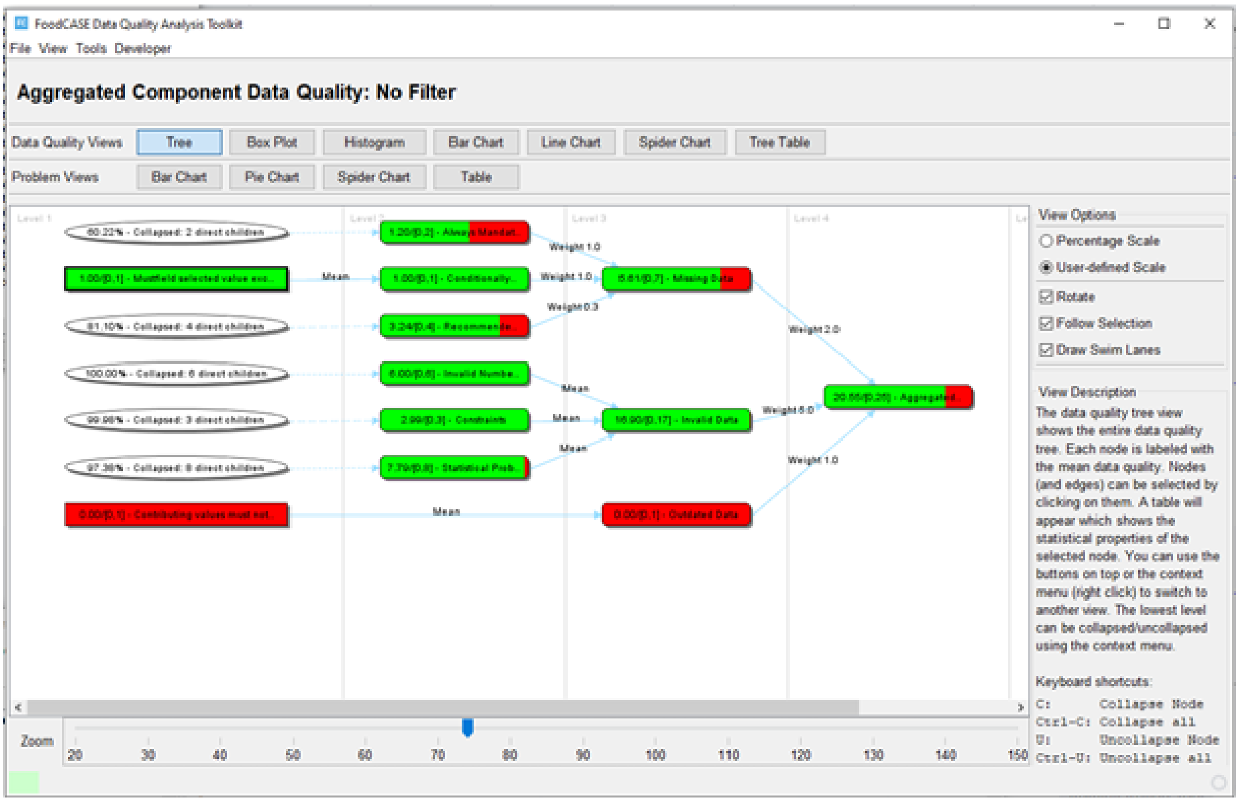
🔴	4.8 %	Fat
🔴	85.7 %	Carbohydrates
🔴	7.2 %	Dietary fibres
●	2.2 %	Protein
●	0 %	Alcohol
●	0 %	Other

Display of nutrients: Standard

Select portion size: 100 g

Figure 5. Food composition visualisation for BLV (FCD, 2020; BLV, 2020).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.



FoodCASE Data Quality Analysis Toolkit
File View Tools Developer
Data Quality Views
Tree
Box Plot
Histogram
Bar Chart
Line Chart
Spider Chart
Tree Table
Problem Views
Bar Chart
Pie Chart
Spider Chart
Table
Level 1
Level 2
Level 3
Level 4
60.22% - Collapsed: 2 direct children
1.00/[0,1] - Multfield selected value exc...
81.10% - Collapsed: 4 direct children
100.00% - Collapsed: 6 direct children
99.98% - Collapsed: 3 direct children
97.36% - Collapsed: 8 direct children
0.00/[0,1] - Contributing values must not...
1.20/[0,2] - Always Mandat...
1.00/[0,1] - Conditionally...
3.24/[0,4] - Recommende...
6.00/[0,6] - Invalid Numbe...
2.99/[0,3] - Constraints
7.70/[0,8] - Statistical Prob...
5.61/[0,7] - Missing Data
16.90/[0,17] - Invalid Data
0.00/[0,1] - Outdated Data
20.55/[0,20] - Aggregated...
Mean
Weight 1.0
Weight 1.0
Weight 0.3
Mean
Mean
Mean
Mean
Weight 2.0
Weight 5.0
Weight 1.0
Zoom
20
30
40
50
60
70
80
90
100
110
120
130
140
150

Aggregated Component Data Quality: No Filter

View Options

○ Percentage Scale

● User-defined Scale

☑ Rotate

☑ Follow Selection

☑ Draw Swim Lanes

View Description

The data quality tree view shows the entire data quality tree. Each node is labeled with the mean data quality. Nodes (and edges) can be selected by clicking on them. A table will appear which shows the statistical properties of the selected node. You can use the buttons on top or the context menu (right click) to switch to another view. The lowest level can be collapsed/uncollapsed using the context menu.

Keyboard shortcuts:

```
C:      Collapse Node
Ctrl-C: Collapse all
U:      Uncollapse Node
Ctrl-U: Uncollapse all
```

Figure 6. Food composition data quality visualisation in FoodCASE (FoodCASE, 2020).

7.3. FNS-Cloud methodology

The FNS-Cloud data map developed in Task 2.1 gives an overview of the types of datasets connected to food nutrition security and an idea how challenging it can be to create an appropriate visualisation tool for each type of data. For that reason, tools will be developed only for selected data types and use cases in WP2 and WP4, considering what tools are already available. To these tools belong for example tools to visualise pathways and to analyse biological processes like PathVisio www.pathvisio.org or WikiPathways http://apps.cytoscape.org/apps/wikipathways.

Data Visualisation concepts will be elaborated and used in the following tasks:

- T2.2 FNS-Cloud catalogues
  The catalogues will represent datasets, tools and services in the FNS-Cloud. The tool provides meta-data about each catalogue entry and data visualisation should be used to give users quick and easy access to catalogue entries.
- T4.1 search engine
  The search functionality of the catalogue in T2.2 should be extended to not only show search results in a list but to provide a graphical overview about datasets in combination with dimensions. Dimensions are several classifications in which datasets can be categorised. Multiple visualisation should be used and allow a user to select the most appropriate presentation to perform exploratory data analysis.
- T4.3 Visualisation of food contamination data
  a dedicated visualisation app for food contamination data will be developed with the added complexity of visualising large amounts of data on a small (smartphone) screen.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 863059.

- T4.5 Visualisation of microbiome data
  In the microbiome use case, some new visualisation will be investigated that can help researchers to explore and analyse their data. Existing tools like WikiPathways will be used and integrated.

In each of the tasks it is necessary to go in depth and understand the data that has to be visualised. In addition, it must be identified what information researchers want to find and how visualisation can support this finding process. The datasets and investigable information will be discussed with data owners and users to ensure full understanding of the datasets as well as user needs and requirements when it comes to visualisation.

7.4. Technical specifications

To ensure the tools are aligned with the FAIR principles and can be used by users with different technical skills and devices available, a decision was made to develop the tools, where possible, as web apps, using Angular (https://angular.io/) and Bootstrap (https://getbootstrap.com/). These frameworks are open source and support web and mobile developments. Java (https://www.java.com/) and NodeJS (https://nodejs.org/) will be used for back-ends. Both are also open source with some restrictions in the case of Java (https://www.java.com/en/download/faq/distribution.xml). No installation of the tools will be required and the users will only need to have access to an internet connection and have one of the commonly used internet browsers installed (i.e. Chrome, Edge/IE, Firefox… (W3Schools, 2020)).

During the design phase of the tools, a mobile-first approach is taken, so that the tools can be used also on smaller screens and mobile devices. The design includes a low-fidelity, conceptual design of the tools and later a high-fidelity prototype, using FNS-Cloud branding, both done in Adobe XD. On each step of the design and development end users and dataset owners will be consulted and also usability tests will be performed in cooperation with WP6 to ensure a high quality and fit-for-purpose solution.

8. FAIRness evaluation

The FNS data interoperability depends on the quality of data and its metadata, therefore, it is important that the state or level of each digital source that is connected to the FNS Cloud is evaluated with regard to the FAIR principles (FAIR, 2020).

Recently, the Research Data Alliance data maturity model Working Group has published a draft specification and guidelines document with a list of indicators for each FAIR principle (findable, accessible, interoperable, reusable) (FAIR data model, 2020). These indicators are aimed to measure the state of a digital object, which may be either data or data-related algorithms, tools or services.

The list of indicators is long and the measuring process requires an extensive amount of work. To ease the work, the RDA community has developed tools for measuring the FAIRness progress of the resource per indicator (FAIRAssist, 2020).

8.1. Review of existing resources

Compliance to the FAIR guidelines should be envisioned as a gradient in a spectrum, rather than a simple question of 'yes' or 'no'. Since it can be challenging to even qualitatively evaluate compliance, multiple initiatives have been developing common standards and frameworks to facilitate this process.

We propose the assessment of a selection of frameworks used to measure the level of FAIRness of digital resources (e.g. data sets, services, applications). During this process we are reviewing existing FAIR assessment frameworks (FAIR assessment frameworks, 2020) and exploring which of the solutions can be the most beneficial for the FNS Cloud consortium. We began with a comprehensive list of available FAIR evaluators, from which we chose the most promising to test against multiple selection criteria. The suggested criteria will cover a broad range of domains, such as content, performance, governance and usability, including openness, efficiency (enables measuring the FAIRness progress per selected indicators as FNS Cloud will dynamically adapt to the situation; considers all relevant indicators), visualization.

For each FAIR assessment framework, a short description has been provided, followed by a qualitative test result which will illustrate whether or not the selection criteria have been successfully met or not. Based on the evaluation of the FAIR assessment frameworks, a recommendation section will be created to conclude this effort. The choice of the recommender framework will be dependent also on the use cases and needs of the FNS Cloud consortium.

Once the FAIR assessment framework is selected, it will be integrated with the FNS Cloud in Task 3.7 (to be presented in D3.3 and D3.4).

9. Conclusions

In this deliverable, specifications of services for dealing with FNS data are presented. As FNS data is complex, from the type and format points of view, services for data pre-processing and structuring, data normalisation, annotation and information extraction, food matching, data analysis and data visualisation are required. Methodology behind the services are described from the theoretical perspective. All the methods presented in D3.2 have been evaluated and have the technological readiness level of at least TRL4. Results of the methods' evaluations have also been published in peer-reviewed papers, which proves their reliability. In the selection and development of the methodology, we considered the needs of the FNS project, especially of the use cases and demonstrators. Moreover, existing resources and tools have been identified and included. The implementation of the services based on the methodology presented in this deliverable, will be presented in D3.3 and D3.4. Results of the application of the services in the FNS use cases and demonstrators will be presented in WP4 and WP5 deliverables.

10. References

WP2-5 FNS Cloud data inventory FINAL, internal document (2020) Available at: https://docs.google.com/spreadsheets/d/1euXaKo419auDMRVVi3HHqMeKWpIukipKvKHeW3fErhs/edit#gid=2142014720 (accessed on August 2020)

Dublin Core definition (2020) Available at: https://www.dublincore.org/resources/metadata-basics/ (accessed on August 2020)

Dublin Core principles (2020) Available at: https://www.dublincore.org/resources/userguide/ (accessed on August 2020)

Nadeau, D., & Sekine, S. (2007). A survey of named entity recognition and classification. Lingvisticae Investigationes, 30(1), 3-26.

Sarawagi, S. (2008). Information extraction. Now Publishers Inc.

Boag, W., Wacome, K., Naumann, T., & Rumshisky, A. (2015). CliNER: a lightweight tool for clinical named entity recognition. AMIA joint summits on clinical research informatics (poster).

Jagannatha, A., Liu, F., Liu, W., & Yu, H. (2019). Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0). Drug safety, 42(1), 99-111.

Soldaini, L., & Goharian, N. (2016, July). Quickumls: a fast, unsupervised approach for medical concept extraction. In MedIR workshop, sigir (pp. 1-4).

Popovski, G., Kochev, S., Korousic-Seljak, B., & Eftimov, T. (2019). FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction. In ICPRAM (pp. 915-922).

Groves, S. (2013). How Allrecipes.com became the world's largest food/recipe site. ROI of Social Media (blog).

Noy, N. F., Shah, N. H., Whetzel, P. L., Dai, B., Dorf, M., Griffith, N., ... & Musen, M. A. (2009). BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research, 37(suppl_2), W170-W173.

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., ... & Hsiao, W. W. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food, 2(1), 1-10.

Donnelly, K. (2006). SNOMED-CT: The advanced terminology and coding system for eHealth. Studies in health technology and informatics, 121, 279.

Popovski, G., Seljak, B. K., & Eftimov, T. (2019). FoodBase corpus: a new resource of annotated food entities. Database, 2019.

Popovski, G., Seljak, B. K., & Eftimov, T. (2020). A Survey of Named-Entity Recognition Methods for Food Information Extraction. IEEE Access, 8, 31586-31594.

European Food Safety Authority (EFSA), Ioannidou, S., Nikolic, M., & Gibin, D. (2019). FoodEx2 maintenance 2016-2018. EFSA Supporting Publications, 16(2), 1584E.

Alexander, M., & Anderson, J. (2012). The Hansard Corpus, 1803-2003.

Sansone SA, Rocca-Serra P, Field D, Maguire E, Taylor C, Hofmann O, Fang H, Neumann S, Tong W, Amaral-Zettler L, Begley K. Toward interoperable bioscience data. Nature genetics. 2012 Feb;44(2):121-6.

Panov, P., Soldatova, L. N., & Džeroski, S. (2016). Generic ontology of datatypes. Information Sciences, 329, 900-920.

Panov, P., Soldatova, L., & Džeroski, S. (2014). Ontology of core data mining entities. Data Mining and Knowledge Discovery, 28(5-6), 1222-1265.

Panov, P., Soldatova, L., & Džeroski, S. (2013, October). OntoDM-KDD: ontology for representing the knowledge discovery process. In International Conference on Discovery Science (pp. 126-140). Springer, Berlin, Heidelberg.
Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Leontis, N. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology, 25(11), 1251-1255.
Arp, R., Smith, B., & Spear, A. D. (2015). Building ontologies with basic formal ontology. Mit Press.
Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., ... & Rosse, C. (2005). Relations in biomedical ontologies. Genome biology, 6(5), R46.
Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., ... & Fan, L. (2016). The ontology for biomedical investigations. PloS one, 11(4), e0154556.
Information Artefact Ontology (2020) Available at: https://github.com/information-artifact-ontology/IAO (accessed on August 2020)
OWL2 (2020) Available at: https://www.w3.org/TR/2012/REC-owl2-overview-20121211/ (accessed on August 2020)
Protege (2020) Available at: https://protege.stanford.edu/ (accessed on August 2020)
Jackson, R.C., Balhoff, J.P., Douglass, E. et al. ROBOT: A Tool for Automating Ontology Workflows. BMC Bioinformatics 20, 407 (2019). https://doi.org/10.1186/s12859-019-3002-3
ROBOT (2020) Available at: http://robot.obolibrary.org/ (accessed on August 2020)
Stojanov, R., Popovski, G., Nasi, J., Trajnov, D., Koroušić Seljak, B., & Eftimov, T. (2020). FoodViz: Visualization of Food Entities Linked Across Different Standards. In Proceedings of the 6th International Conference on Machine Learning, Optimization and Data Science, In Press.
Popovski, G., Seljak, B. K., & Eftimov, T. (2019). FoodBase corpus: a new resource of annotated food entities. Database, 2019. Available at: https://www.mdpi.com/2072-6643/9/6/542 (accessed on August 2020)
Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2017). StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. Nutrients, 9(6), 542.
AGROVOC (2020) Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc (accessed on August 2020)
Popovski, G.; Seljak, B. and Eftimov, T. (2019). FoodOntoMap: Linking Food Concepts across Different Food Ontologies.In Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 2: KEOD, ISBN 978-989-758-382-7, ISSN 2184-3228, pages 195-202. DOI: 10.5220/0008353201950202
Microbiome analysis (2020) Available at: https://github.com/annaritzen/microbiome-analysis; https://github.com/sabdeolive/Sabrina_BMS_Bachelor-Internship (accessed on August 2020)
PathVisio (2020) Available at: https://pubmed.ncbi.nlm.nih.gov/25706687/ (accessed on August 2020)
CoNet (2020) Available at: https://pubmed.ncbi.nlm.nih.gov/27853510/ (accessed on August 2020)
Korem, T., Zeevi, D., Zmora, N., Weissbrod, O., Bar, N., Lotan-Pompan, M., ... & Suez, J. (2017). Bread affects clinical parameters and induces gut microbiome-associated personal glycemic responses. Cell metabolism, 25(6), 1243-1253.
Eftimov, T., Popovski, G., Valenčič, E., & Seljak, B. K. (2020). FoodEx2vec: New foods' representation for advanced food data analysis. Food and Chemical Toxicology, 138, 111169.

Ji, S., Pan, S., Cambria, E., Marttinen, P., & Yu, P. S. (2020). A survey on knowledge graphs: Representation, acquisition and applications. arXiv preprint arXiv:2002.00388.
Gesese, G. A., Biswas, R., & Sack, H. (2019, June). A Comprehensive Survey of Knowledge Graph Embeddings with Literals: Techniques and Applications. In DL4KG@ ESWC (pp. 31-40).
Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., ... & Hsiao, W. W. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. npj Science of Food, 2(1), 1-10.
Data visualisation (2020) Available at: https://www.oreilly.com/library/view/designing-data-visualizations/9781449314774/ch01.html (accessed on August 2020)
EEA (2020) Available at: http://www.storytellingwithdata.com/blog/2014/04/exploratory-vs-explanatory-analysis (accessed on August 2020)
IVLA (2020) Available at: https://ivla.org/ (accessed on August 2020)
Visual Literacy Standards Task Force, ACRL, 2011
Schmid, C., Hinterberger, H. (1994). Comparative Multivariate Visualization Across Conceptually Different Graphic Displays. Proceedings of the Seventh International Working Conference on Scientific and Statistical Database Management (VII SSDBM). Los Alamitos, USA, pp. 42-51.
Schmid, C. (1999). Active Comparative Visualization: A Novel Way of Exploring Multivariate Data. Dissertation ETH No. 13116.
Buergi, M. (2004). Interaktive, multivariate Visualisierungsmethoden als COMAdd-Ins für Microsoft Excel. Master Thesis, ETH Zürich, Switzerland.
Scheuner B., (2014), Analyse eines technical visual literacy-Unterrichts mit e-Observation, Dissertation, https://doi.org/10.3929/ethz-a-010210532
Presser, K., Weber, D., & Norrie, M. (2018). FoodCASE: A system to manage food composition, consumption and TDS data. Food Chemistry, 238, 166–172. http://dx.doi.org/10.1016/j.foodchem.2016.09.124
FCD (2020) Available at: https://naehrwertdaten.ch/de/ (accessed on August 2020)
BLV (2020) Available at: https://www.blv.admin.ch/blv/de/home.html (accessed on August 2020)
FoodCASE (2020) Available at: https://www.foodcase.org/ (accessed on August 2020)
W3Schools (2020) Available at: https://www.w3schools.com/browsers/ (accessed on August 2020)
FAIR (2020) Available at: https://www.go-fair.org/fair-principles/ (accessed on August 2020)
FAIR data model (2020) Available at: https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/fair-data-maturity-model-specification-and-guidelines (accessed on August 2020)
FAIRAssist (2020) Available at: https://fairassist.org/ (accessed on August 2020)
FAIR assessment frameworks (2020) Available at: https://www.rd-alliance.org/group/fair-data-maturity-model-wg/outcomes/results-analysis-existing-fair-assessment-tools (accessed on August 2020)

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 863059.