



FNS - Cloud

Food Nutrition Security

Food Nutrition Security Cloud

Deliverable 2.1

Definition of data models and APIs

Due Date:	31.03.2020
Submission Date:	31.05.2020
Revision Date	13.08.2021
Dissemination Level:	Public
Lead beneficiary:	PMT & EuroFIR
Main contact:	Karl Presser (karl.presser@premotec.ch) Mark Roe (mr@eurofir.org) Agnieszka Matuszczak (agnieszka.matuszczak@premotec.ch) Paul Finglas (pf@eurofir.org)

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	<i>D2.1 Definition of data models and APIs</i>																																				
Editor(s)	<i>Karl Presser, Agnieszka Matuszczak & Wiktor Kapela (PMT), Mark Roe & Sian Astley, Barbara Korousic Seljak (JSI), Eileen Gibney (UCD), Enrique Carrillo (IMDEA), Rachel Berry & Maria Traka (QIB), Paul Finglas (QIB & EuroFIR)</i>																																				
Reviewer(s)	<i>Marinel Cavelaars & Tess Korthout (HYVE), Irina Stoyanova (SF)</i>																																				
Dissemination Level	<input type="checkbox"/> CO Confidential <input checked="" type="checkbox"/> PU Public																																				
Approved by	<table border="0"> <tr> <td><input checked="" type="checkbox"/> RTDS (COO)</td> <td><input checked="" type="checkbox"/> UM</td> <td><input checked="" type="checkbox"/> ILSI</td> </tr> <tr> <td><input checked="" type="checkbox"/> QIB (SCO)</td> <td><input checked="" type="checkbox"/> NUTRIS</td> <td><input checked="" type="checkbox"/> BfR</td> </tr> <tr> <td><input checked="" type="checkbox"/> JSI</td> <td><input checked="" type="checkbox"/> RIVM</td> <td><input checked="" type="checkbox"/> AUTH</td> </tr> <tr> <td><input checked="" type="checkbox"/> UCD</td> <td><input checked="" type="checkbox"/> WUR</td> <td><input checked="" type="checkbox"/> FEM</td> </tr> <tr> <td><input checked="" type="checkbox"/> PMT</td> <td><input checked="" type="checkbox"/> UGent</td> <td><input checked="" type="checkbox"/> CNR</td> </tr> <tr> <td><input checked="" type="checkbox"/> JDLC</td> <td><input checked="" type="checkbox"/> IMDEA</td> <td><input checked="" type="checkbox"/> APRE</td> </tr> <tr> <td><input checked="" type="checkbox"/> EuroFIR</td> <td><input checked="" type="checkbox"/> HUA</td> <td><input checked="" type="checkbox"/> CAP</td> </tr> <tr> <td><input checked="" type="checkbox"/> UWTSO</td> <td><input checked="" type="checkbox"/> TUM</td> <td><input checked="" type="checkbox"/> UNIFI</td> </tr> <tr> <td><input checked="" type="checkbox"/> DTU</td> <td><input checked="" type="checkbox"/> GS1</td> <td><input checked="" type="checkbox"/> LIFE</td> </tr> <tr> <td><input checked="" type="checkbox"/> ENEA</td> <td><input checked="" type="checkbox"/> SF</td> <td><input checked="" type="checkbox"/> Nutritics</td> </tr> <tr> <td><input checked="" type="checkbox"/> HYVE</td> <td><input checked="" type="checkbox"/> UoR</td> <td><input checked="" type="checkbox"/> EFF</td> </tr> <tr> <td><input checked="" type="checkbox"/> HYLO</td> <td><input checked="" type="checkbox"/> IFA</td> <td></td> </tr> </table>	<input checked="" type="checkbox"/> RTDS (COO)	<input checked="" type="checkbox"/> UM	<input checked="" type="checkbox"/> ILSI	<input checked="" type="checkbox"/> QIB (SCO)	<input checked="" type="checkbox"/> NUTRIS	<input checked="" type="checkbox"/> BfR	<input checked="" type="checkbox"/> JSI	<input checked="" type="checkbox"/> RIVM	<input checked="" type="checkbox"/> AUTH	<input checked="" type="checkbox"/> UCD	<input checked="" type="checkbox"/> WUR	<input checked="" type="checkbox"/> FEM	<input checked="" type="checkbox"/> PMT	<input checked="" type="checkbox"/> UGent	<input checked="" type="checkbox"/> CNR	<input checked="" type="checkbox"/> JDLC	<input checked="" type="checkbox"/> IMDEA	<input checked="" type="checkbox"/> APRE	<input checked="" type="checkbox"/> EuroFIR	<input checked="" type="checkbox"/> HUA	<input checked="" type="checkbox"/> CAP	<input checked="" type="checkbox"/> UWTSO	<input checked="" type="checkbox"/> TUM	<input checked="" type="checkbox"/> UNIFI	<input checked="" type="checkbox"/> DTU	<input checked="" type="checkbox"/> GS1	<input checked="" type="checkbox"/> LIFE	<input checked="" type="checkbox"/> ENEA	<input checked="" type="checkbox"/> SF	<input checked="" type="checkbox"/> Nutritics	<input checked="" type="checkbox"/> HYVE	<input checked="" type="checkbox"/> UoR	<input checked="" type="checkbox"/> EFF	<input checked="" type="checkbox"/> HYLO	<input checked="" type="checkbox"/> IFA	
<input checked="" type="checkbox"/> RTDS (COO)	<input checked="" type="checkbox"/> UM	<input checked="" type="checkbox"/> ILSI																																			
<input checked="" type="checkbox"/> QIB (SCO)	<input checked="" type="checkbox"/> NUTRIS	<input checked="" type="checkbox"/> BfR																																			
<input checked="" type="checkbox"/> JSI	<input checked="" type="checkbox"/> RIVM	<input checked="" type="checkbox"/> AUTH																																			
<input checked="" type="checkbox"/> UCD	<input checked="" type="checkbox"/> WUR	<input checked="" type="checkbox"/> FEM																																			
<input checked="" type="checkbox"/> PMT	<input checked="" type="checkbox"/> UGent	<input checked="" type="checkbox"/> CNR																																			
<input checked="" type="checkbox"/> JDLC	<input checked="" type="checkbox"/> IMDEA	<input checked="" type="checkbox"/> APRE																																			
<input checked="" type="checkbox"/> EuroFIR	<input checked="" type="checkbox"/> HUA	<input checked="" type="checkbox"/> CAP																																			
<input checked="" type="checkbox"/> UWTSO	<input checked="" type="checkbox"/> TUM	<input checked="" type="checkbox"/> UNIFI																																			
<input checked="" type="checkbox"/> DTU	<input checked="" type="checkbox"/> GS1	<input checked="" type="checkbox"/> LIFE																																			
<input checked="" type="checkbox"/> ENEA	<input checked="" type="checkbox"/> SF	<input checked="" type="checkbox"/> Nutritics																																			
<input checked="" type="checkbox"/> HYVE	<input checked="" type="checkbox"/> UoR	<input checked="" type="checkbox"/> EFF																																			
<input checked="" type="checkbox"/> HYLO	<input checked="" type="checkbox"/> IFA																																				
IPRs underlined	None identified																																				
Datasets underlined	To be addressed																																				

Version/Date	Change/Comment
V0.1_2020-03-19	<i>Initial merge of documents prepared by Mark and Karl</i>
V0.2_2020-03-25	<i>Second merge of documents prepared by Mark, Karl, Wiktor and Agnes</i>
V0.3_2020-03-26	<i>Merge of Mark, Barbara and Agnes versions, document structure adjustment</i>
V0.4_2020-03-27	<i>Addition of microbiome data information and appendix from Rachel</i>
V0.5_2020-03-30	<i>Comments from Paul. Revision of figure 1, addition of conclusion and summary</i>
V0.6_2020-04-06	<i>Revisions based on group discussion</i>
V0.7_2020-04-14	<i>Added final version of figure 1 (data map). Reformatted to new template.</i>
V0.8_2020-05-05	<i>New text added, comments addressed, restructured to match data map figure</i>
V0.9_2020-05-20	<i>New text added to address remaining comments</i>
V1.0_2020-05-29	<i>Final administrative edits by COO</i>
V2.0_2021-08-13	<i>Updated deliverable as per experts request for revision</i>

Table of Contents

1	Publishable summary	4
2	Introduction	5
3	FNS-Cloud Data Map	7
4	Data transfer models	9
4.1	Application Programming Interfaces (APIs)	11
4.2	Existing data transfer harmonisation	13
4.3	Data transfer recommendations.....	13
4.3.1	Data field (attribute) types	16
5	Agri-Food Data	17
5.1	Food description and classification	17
5.1.1	Food production and processing	18
5.1.2	Source of foods	18
5.2	Composition	19
5.2.1	Nutrients	19
5.2.2	Bioactive components	22
5.2.3	Contaminants.....	22
5.2.4	Data Exchange Model	23
5.2.5	Thesauri	26
5.2.6	APIs	27
5.3	Labelling	28
5.3.1	Branded Food.....	28
5.4	Authenticity.....	32
5.4.1	Data Exchange Model	33
5.4.2	Thesauri	36
5.4.3	API.....	40
5.5	Food safety.....	41
5.5.1	Analytical Data and total diet study (TDS) data	41
5.5.2	Metrology	48
6	Food Intake and Lifestyle Data	57
6.1	Consumption and lifestyle data	57
6.1.1	EFSA EU Menu guidelines	57
6.1.2	Software examples.....	59
6.1.3	Data Exchange Model	62
6.1.4	Thesauri	68
7	Nutrition and Health Data	74
7.1	Biomedical data.....	74
7.1.1	Data Exchange Model	79
7.1.2	Thesauri	81
7.2	Genomic data	81
7.2.1	FHIR (Fast Healthcare Interoperability Resources)	84
7.3	Microbiome data.....	85
7.4	Food-drug interaction data	86
7.4.1	API.....	88
8	Data Inter-operability and Quality	89
8.1	Inter-operability	89
8.2	Data quality	90
9	Conclusion	91
10	References	92
11	Appendices	99

1 Publishable summary

This deliverable reviews and provides a map of food data areas relevant to FNS-Cloud and, for each data area, recommends a data exchange model, thesauri, and APIs. Existing data standards, thesauri and APIs that can form the basis of data exchange in FNS-Cloud are evaluated. Data models (entities and attributes) for each data area are based on existing data standards and models that are considered most applicable for FNS-Cloud.

Data areas covered include Agri-food, Nutrition, and Non-Communicable Diseases and Microbiome, reflecting the planned FNS-Cloud Demonstrators. Key entities that are likely to link the different data areas in FNS-Cloud relate to; Food, Food intake and lifestyle and Nutrition and health, reflecting the on-going FNS-Cloud Use Cases. To help capture information on data structures and tools for data exchange, all FNS-Cloud Beneficiaries were asked to complete two surveys; one was a survey of existing datasets and data structures and, the other, a survey of tools used for data exchange. Results from these surveys are important to ensure that WP2 provides information on data maps and APIs that are needed and relevant for WP4 Use Cases and WP5 Demonstrators. This Deliverable is also relevant for WP3, which is developing services to support standardisation and interoperability of data from other, often heterogenous, sources.

After considering API styles, our recommendation for FNS-Cloud is a REST architecture with a JSON data format. REST is a simple architecture to implement and consumes less server resources. The JSON file format also allows for smaller bandwidth consumption, which is important when transferring large amounts of data.

Data maps for the food area have been adopted from well-established standards and existing mappings amongst thesauri. The recommended data model for the food intake area is based on the data schema for the EFSA EU Menu project, but it is likely that this will need to be extended to handle intake datasets from more heterogenous research projects. The ENPADASI produced Ontology for Nutritional Studies can be used as a starting point for a data model for the nutrition and health data area, but further development and linking to relevant thesauri, including the Unified Medical Language System used for biomedical data, will be needed.



2 Introduction

Deliverable 2.1 Definition of data models and APIs reviews and provides a map of food data areas relevant to FNS-Cloud and, for each data area, a data exchange model, thesauri, and Advanced Programme interfaces (APIs) are recommended (an API is a computing interface which defines interactions between multiple software intermediaries). Existing data standards, thesauri, and APIs were reviewed, as the basis for data exchange in FNS-Cloud. Based on these existing data standards, a data model (entities and attributes) is provided for each data area. Existing APIs for data exchange are also described and were reviewed in terms of their current use and limitations for future use in FNS-Cloud. Data areas covered include:

- Agri-Food
 - Food classification and description
 - Composition
 - Labelling
 - Authenticity
 - Safety
- Food intake and Lifestyle
 - Food intake
 - Food choice and consumer behaviour
 - Socio-economic status
 - Lifestyle
- Nutrition and Health
 - Anthropometric data
 - Biomarkers
 - Phenotype
 - Genomic data
 - Microbiome

To help capture information on data structures and tools for data exchange, all FNS-Cloud Beneficiaries were asked to complete two surveys, as part of the of WP2 (task 2.1): one on existing datasets and data structures and the other on tools used for data exchange. These surveys were important to ensure that WP2 provides information on data maps and APIs that are needed by WP4 and WP5 and if needed to consult experts in the given areas on the existing standards and data transfer models for their datasets. Details are included in this Deliverable, and a summary of results from the data surveys are provided in Appendix 1 and 2. This Deliverable is also relevant to WP3 in developing services that support standardization and inter-operability of data from other, often heterogeneous, sources.

FNS-Cloud will link data related to food and nutrition (food intake, absorption of nutrients and consumer behaviour), security and people (populations, including individuals), who consume foods with indicators of nutrition and health status (including biomarkers, genotype, phenotype, and microbiome).

The output of this task is complementary to D3.1 but goes covers a completely different objective. This deliverable takes the FNS data map that was defined in the project proposal, containing FNS topics and

sub-topics and defines for each sub-topic the recommended data exchange model and APIs that system programmers should follow to support automated and electronic data exchange, and therefore interoperability. Deliverable 3.1 uses the same FNS map and makes a short data standard and guideline review to demonstrate the process of mapping them to the FNS topics. This short review was done in collaboration with T2.1 and the outcome were used as starting point for this deliverable. The focus of D3.1 is how the FNS Cloud could look like from a process engineering point of view. D3.1 describes how users starts with a research question, go to the corresponding sub-domain, get corresponding datasets from different data repositories, and receive linked datasets based on the data standards and guidelines.

In the scope of this deliverable, an FNS-Cloud Data Map was developed, using the results of the before mentioned surveys, that is presented in chapter 3. Chapter 4 explains the general concept of data transfer and APIs, while detailed data transfer models for different dataset types (based on the different FNS-Cloud Data Map areas) are defined in chapters 5 Agri-Food Data, 6 Food intake and Lifestyle and 7 Nutrition and Health Data., both covering the current state as well as recommendations for the future. The deliverable finishes with chapter 8 covering the issues around data interoperability and quality and finally conclusions in chapter 9.

3 FNS-Cloud Data Map

There have always been challenges in harmonising approaches to acquisition and publication of data related to food, nutrition, and security. While data processing and publication in electronic form makes large datasets easier to handle and more accessible, challenges related to data quality, exchange formats, and documentation have increased disproportionately. Many projects and research networks have initiated harmonisation of methods to collect, manage, and publish data, but data harmonisation has not, generally, kept pace with fast moving developments in information and communications technology. The ability to generate and exchange large amounts of data has, therefore, highlighted existing limitations in data structure associated with data management, accessibility, transfer, and re-use of data.

Data maps are an essential starting point to compare datasets and enable mapping and linking between different datasets. It is not essential for data structures to be identical, although it does make handling easier, but it is important that there are common entities and attributes that allow linking. In many cases data structures and taxonomies will depend on the system used originally for collection and the intended application.

RICHFIELDS (GA No. 654280)¹ identified, analysed, and tested implementation of a data platform infrastructure, focusing on linking data and supporting information with resources, in terms of platform content and technical feasibility. RICHFIELDS aimed to design the approach, while FNS-Cloud is taking this to the next stage and implementing the cloud infrastructure and rolling out three Demonstrators for user communities. Case studies, representing four topic areas (determinants of dietary behaviour, intake of foods and nutrients, status, and functional markers of nutritional health) and field-trials (WP4) form the basis of the FNS-Cloud Demonstrators (WP5). The case studies and field trials are designed to review the state-of-the-art and identify gaps and needs in relation to food, nutrition and security data and represent types of data relevant to FNS-Cloud, including outputs from previous initiatives. FNS-Cloud will link case study, field trial, and other datasets in the areas of food, nutrition, and security in order that the WP5 Demonstrators will go beyond the current state-of-the-art once deployed.

Figure 1. provides an outline data map that covers the data areas included in FNS-Cloud.

¹Bogaardt et al., 2019

-  - Food data
-  - Person data

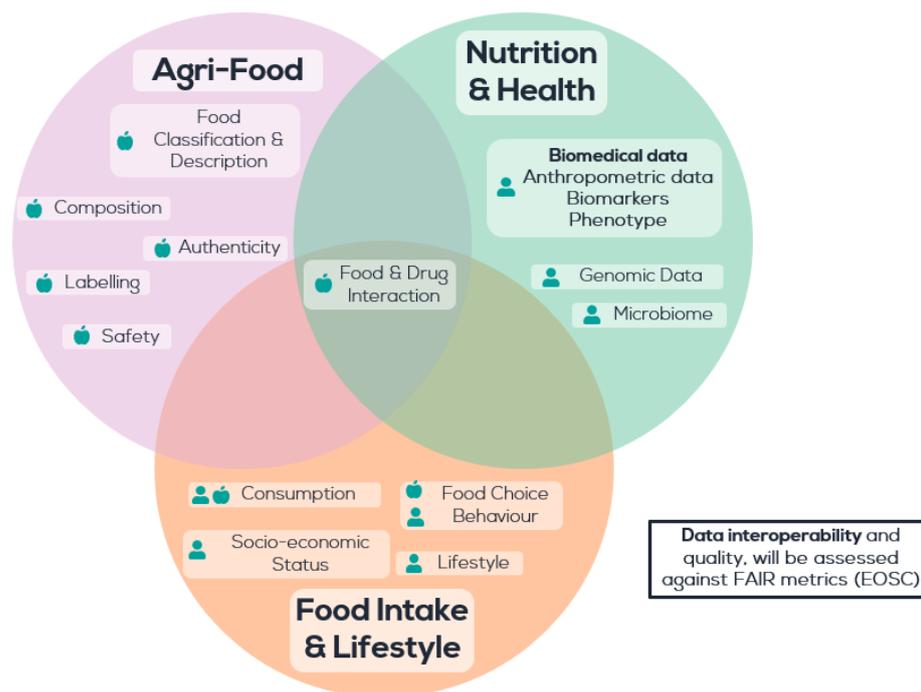


Figure 1. Outline data map for FNS-Cloud.

4 Data transfer models

Identifying existing datasets related to food nutrition security, grouping them into areas, and harmonising their format is only the first step in interoperability. Datasets from the same areas, which have common structures can be compared and analysed together. However, a common data transfer model must also be found to exchange and reuse data. This chapter focuses on the general concept of data transfer, while detailed data transfer models and API recommendations for each of the identified FNS-Cloud Data Map topics is provided in Chapters 5 Agri-Food Data, 6 Food Intake and Lifestyle and 7 Nutrition and Health Data

User communities will use a variety of systems to gather and manage data and backend implementation and database structure will differ between software. The important thing for FNS-Cloud is that systems are able to communicate with one another and “understand” data in the same way.

There are several ways in which datasets can be transferred (Figure 2.), which can be classified broadly as:

- **Manual** – data must be interpreted by an individual and transferred manually from one system to another, which is time consuming and error prone. Data can be unstructured and transformed by the person transferring the data to fit the structure of the receiving system. Examples of this data transfer are copying and pasting information found on a website or transcribing information from a written report into a data management system.
- **Semi-automatic** – data are transferred automatically into the data management system but need first to be ‘interpreted’ by a person. Data need to be structured and users must know and understand structures used in both systems. The amount of human work and time needed, as well as the potential for error in comparison to manual data transfer, are reduced. An example of a semi-automatic data transfer is an Excel worksheet imported into a data management system, where the user assigns the Excel columns to specific fields.
- **Automatic** – data can be transferred and interpreted by the data management systems without human intervention; data must be fully structured, with thesauri, and mandatory fields defined. This data transfer is quickest, allows full interoperability, and is the least error prone. The downside is that it requires harmonisation and standardisation in advance, and appropriate APIs need to be implemented that can communicate with one another.

a) Manual data transfer:

Initial unstructured data
An apple has on average 2.4 g of fibre per 100 g.



Data transferred to a data management system	
Food	Apple
Component	Fibre
Amount	2.4 g / 100g

Manual data transfer and interpretation by a person.

b) Semi-automatic data transfer:

Initial structured data in data management system A	
Food	Apple
Component	Fibre
Amount	2.4 g / 100g



Data transferred to a data management system B	
Food Name	Apple
Nutrient	Fibre
Value	2.4
Unit	gram
Matrix unit	per 100 g total food

Semi-automatic data transfer, where the person defines that Food=Food Name, Component=Nutrient.

The problem arises when the Amount=Value + Unit + Matrix unit because a dedicated script or function must be created to distribute the information into 3 separate fields automatically.

c) Automatic data transfer:

Initial structured data in data management system A	
Food	Apple
Component	Fibre
Amount	2.4 g / 100g



Data transferred to a data management system B	
Food Name	Apple
Nutrient	Fibre
Value	2.4
Unit	gram
Matrix unit	per 100 g total food

Automatic transfer of data, where the transfer is made via an API in a format, that is understandable by both systems.

Figure 2. Methods of data transfer

It is important to note that two communicating data management systems can manage datasets using different models. For example, as shown in b) and c), where the systems stored the information about the composition using different models (i.e., in the initial system A there was only the “Amount” entity, while in the second system B there are separate entities for “Value”, “Unit” and “Matrix unit”) but the data stored in both systems is basically the same. Dataset models used in the API must be standardised and brought to a common format, for all the systems communicating using the API to be able to later easily “translate” the data they are receiving into their native data model.

The channels and formats used for data transfer add another level of complexity to the process of data sharing. For example, (semi)structured datasets come in many formats:

- Excel files (.csv)
- XML files
- JSON (<https://www.rfc-editor.org/info/std90>, <http://json-schema.org/>)
- Database dumps

These datasets can be shared via different channels as:

- email
- Downloads from and/or online repository
- On physical data storage devices (e.g., USB memory stick)

These pose many issues in terms of data security (e-mails can be leaked, memory sticks lost or stolen) and might not be applicable for very large datasets. These challenges can be overcome if we allow for data transfer using secure and standardised channels. The most technically mature and state of the art approach is using APIs, that is an Application Programming Interfaces, that are further explained in the following chapter.

4.1 Application Programming Interfaces (APIs)

An interface (as defined by the Oxford dictionary) is a point at which two systems, subjects, organizations, etc. meet and interact². Two entities can operate in different ways and do not need to understand how one another works but, by creating common understanding, they can cooperate, generating added value through the cooperation. A good example is a user interface, where neither the software nor the user needs to know about background operations to work together using a common language (buttons, written commands, error messages, etc.). Similarly, software systems can communicate and exchange data or provide input/output for functionalities using an Application Programming Interface (API). Initially, APIs were understood as interfaces that are specific to certain software³. The interfaces were not standardised and allowed only functionality of specific software to be extended (e.g., website plugins). As more software was required to cross-communicate, more endpoints had to be developed, resulting in higher costs for development and maintenance. Each set of software had to create a separate API to exchange data. As

² https://www.oxfordlearnersdictionaries.com/definition/english/interface_1

³ https://en.wikipedia.org/wiki/Application_programming_interface

internet and interoperability have gained momentum, demand for API standardization has increased. By defining standard APIs, software needs only to integrate with that API. Figures 3 and 4 show how database management systems can communicate with each other, either directly or via standardized APIs.

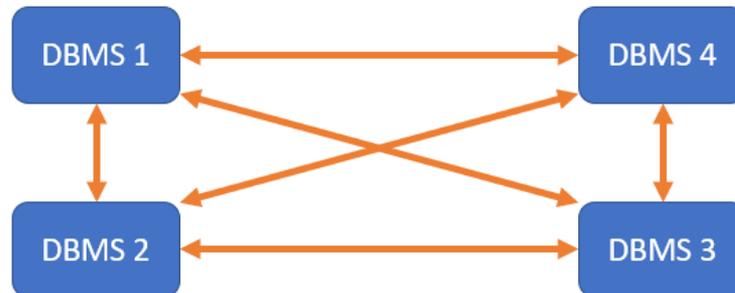


Figure 3. Four Database Management Systems with dedicated endpoints to communicate with each other.

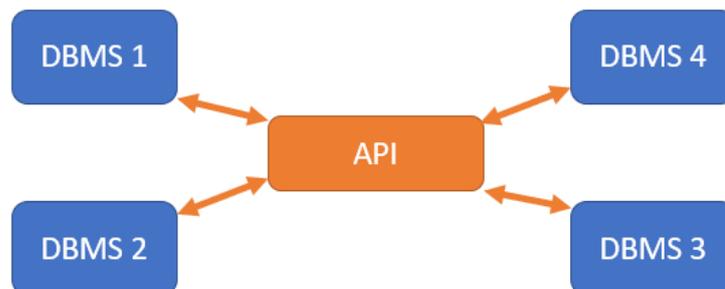


Figure 4 The same four Database Management Systems able to communicate using a standardised API.

Currently, the two most commonly utilised API styles are:

- SOAP (Simple Object Access Protocol) – an XML based messaging protocol that allows distributed elements of an application, or two separate applications, to communicate. SOAP services have a strict communication contract that both parties need to know beforehand.

REST (REpresentational State Transfer) – a software architecture style that defines a set of constraints to be used to create web services. Using HTTP as its transport protocol, REST services do not require any knowledge of the target system – the consumer needs only to know the Uniform Resource Identifier (URI) of the resource that they want to access and its parameters.

4.2 Existing data transfer harmonisation

There have already been efforts to compile food data related APIs, notably Quisper ASBL/ QSP, which was created for personalised nutrition advice services. Quisper ASBL/ QSP⁴ provides APIs⁵ for:

- Dietary reference values
- Food composition - implemented as a SOAP webservice, that is a basis for the new Food Composition REST API (described in following chapters)
- Personal nutrition questionnaires
- Ageing information, based on DNA results
- Personalised dietary advice
- eNutri app (to be connected)

The mission of Quisper ASBL is to provide scientifically valid data underpinning personalised nutrition advice service. The APIs developed were mostly REST, except for food composition, and use the EuroFIR thesauri⁶, which are well-established in the food composition user community, and are a good basis for extension into other food-related datasets.

4.3 Data transfer recommendations

As recommended in section 4.1 there are two common approaches to creating web APIs: REST and SOAP. Our recommendation for the FNS-Cloud project is to use REST⁷ API with a JSON data format.

Several factors have been taken into the consideration, when reaching this decision:

- REST is much simpler than SOAP to implement and is more widespread⁸ (therefore, more known) - allows more developers and projects to utilise and benefit from the infrastructure,
- REST is more lightweight – consumes less resources and is cheaper to maintain,
- JSON file format requires much less bandwidth than XML, which again means cheaper maintenance,
- REST is stateless, and therefore easier to scale when needed,
- SOAP offers built-in security features, which makes it a good choice for high-security applications, but for our needs most of those features aren't necessary and will only reduce performance and increase running costs for the infrastructure, for example, SOAP is still being used in banking systems, since the security features are inherently built into this framework, similar securities can be also added in REST APIs, but they are an addition,

⁴ <https://quisper.eu/>

⁵ <https://developer.quisper.eu/apis>

⁶ <http://www.eurofir.org/our-resources/eurofir-thesauri/>

⁷ <https://martinfowler.com/articles/richardsonMaturityModel.html>

⁸ <https://jaxenter.com/state-of-api-integration-report-136342.htm>

- REST/JSON is a format already utilised by EU projects – during initial investigation for this deliverable it was discovered that EOSC is already using said format for its [own API implementation](#).

Issues such as consumption of server resources and bandwidths are particularly important when transferring large amounts of data, which is to be expected in this project.

For implementation, where possible, commonly used standards should be deployed:

- URI defines resources that are accessed:
 - `/{entity}` - list of entities
 - `/{entity}/{identifier}` – single entity
- HTTP method defines operations to perform on a resource:
 - GET – retrieve resource
 - Currently we don't plan to support modifying resources in other systems via API
- Naming convention for entities and parameters: camel case with no spaces (ex. `foodName`)
- All the entity data returned should be returned in “data” param of the JSON response. For entities list endpoints, additional parameters should be provided, such as: “totalData” – stating how much results there were for provided parameters and “currentPage” – stating which data page is the data from. Example response is shown below:

```
{
  "totalData": 265,
  "currentPage": 1,
  "data": {
    { "id": ... },
    ...
  }
}
```

Additionally, we should utilise:

- Query params to filter data lists:
 - **GET** `/food?name=`
- Special query param (“\$p” parameter) – to paginate the results of query. Start page numbering from 1. We propose a default page size of 20 results. For example:
 - **GET** `/food` – load first 1-20 results
 - **GET** `/food?$p=1` – equivalent to above
 - **GET** `/food?$p=3` – load results 41-60
- Two special query params (“\$s” and “so” parameters) – to decide sorting of query results. User should be able to specify both column he wants to sort by (“\$s”) and order of sorting (“\$so” – either “asc” for ascending or “desc” for descending, by default “asc”). For example:
 - **GET** `/food` – load results without any sorting
 - **GET** `/food?$s=name` – load results sorted by “name” column in ascending order
 - **GET** `/food?$s=name&$so=desc` – load results sorted by “name” column in descending order

- Special query param (“\$ex” parameter) – to optionally load subparts of entity. For example:
 - **GET** /food/5 – should load Food entity with ID 5
 - **GET** /food/5?\$ex=components – should load Food entity with ID 5 and all the component entities connected to the food

The usage of \$expand parameter will reduce load and bandwidth, since not all entities connected will have to be loaded for each request. This allows the application developer to decide which data is required at the time.

More examples for specific datasets and data areas are provided below.

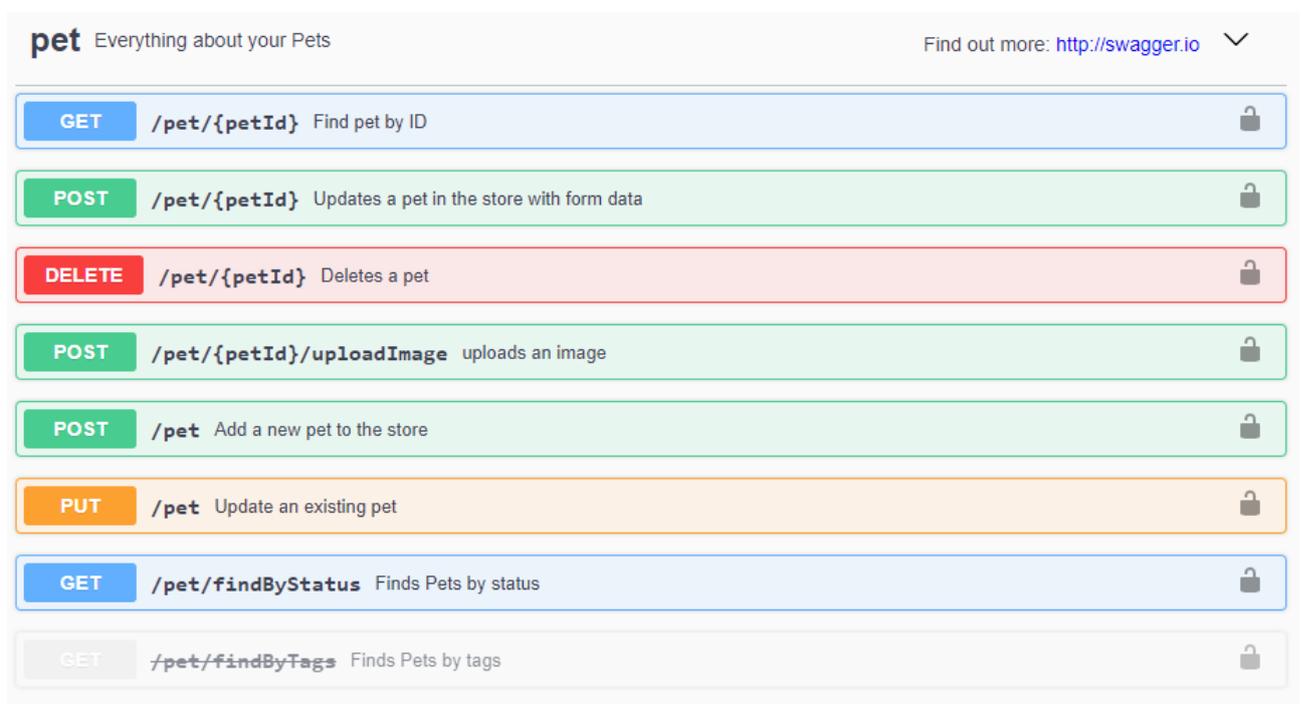


Figure 5. Example HTTP methods (operations) used in REST.

Documentation for APIs can be visualized using Swagger, an open-source tool that supports API design, development, documentation, and testing. FNS-Cloud beneficiaries and future IT collaborators who would like to integrate their tools (i.e., software, apps, algorithms) into the FNS Platform, should submit their API documentation using this software, as it will allow other platform users that have IT knowledge to easily review it, test their endpoint implementations and integrate with each other. This will ensure that the platform, with not only datasets but also documentation needed to access the data are FAIR (findable, accessible, interoperable, and reusable).

A similar approach is being tested by Quisper ASBL/ QSP.

4.3.1 Data field (attribute) types

In the following chapters, suggested data model definitions will be presented for the defined data areas. The data models will be presented in a table format, with the entity name and its attributes. A name, description and data type will be provided for the attributes as well as information if they are mandatory or unique.

In structured databases, each attribute has a certain data type assigned and this should be also implemented in the API to promote data and data format harmonisation. Below are listed types of attributes that can be used in datasets, with a description and examples:

Type abbreviation	Description	Example
STR	String; text field	Example text
DAT	Date format	2019-05-20
DAT(TIM)	Date with time	2019-05-20 14:31:45
INT	Integer	165
NUM	Decimal number, '.' Separated	1.213
FRC	Fraction; decimal between 0 and 1(inclusive)	0.33;0;1
BLN	Boolean	TRUE;FALSE
FILE	File	a picture
THS	Thesaurus entry in a string format. For all thesaurus entries separate complete list of values will be supplied.	B0123
KEY/FKEY	Key or Foreign Key – positive integer	112

In the before mentioned tables with attributes for entities, the abbreviated format names will be used for clarity.

5 Agri-Food Data

The Agri-Food domain includes data related to agriculture, horticulture, and food and drink processing technologies.

5.1 Food description and classification

Accurate and consistent description of foods included in datasets is a major determinant of data quality and whether or not values can easily be used by a range of users with different needs. Data can be provided for raw foods, processed foods and for foods as purchased or as single or combination of foods prepared for consumption (meals). Any process applied to a food is likely to change its composition and effect the food components supplied and absorbed when consumed. The different states of foods included in datasets means that a consistent and accurate approach to the description of each food, particularly where differences impact on nutrient composition, e.g., vegetables, raw or cooked; vegetables boiled with or without added salt, fish canned in brine or oil, is essential for correct mapping between datasets. Food description and classification is a key factor in mapping and merging of food composition and food intake datasets to allow combined analysis.

Food composition data is traditionally published in book or table form allowing for additional descriptive text to be added to provide more details on foods and distinguish between foods where necessary. However, searchable electronic datasets mean that, even where additional descriptive information is available, it may not always be presented alongside the food name in a structured format. Facetted systems have been developed to describe foods and to distinguish between different foods that may not be fully described by the food name alone, e.g., to describe the plant or animal source of a food in more detail; specify parts analysed and/or inedible waste; describe food processing such as cooking method, preservation method or addition of ingredients. These systems also aid the comparison of data between different datasets, particularly those from different countries (with names in native languages) or those that were intended for a different purpose.

The LanguaL (Langua alimentaria) system⁹ was developed by the United States Food and Drug Administration in the 1970s and has been modified and adopted for use in European countries as part of the EuroFIR initiative to better standardise approaches for food description used for foods in national FCDBs in European countries¹⁰. The European Food Safety Authority (EFSA) has also developed a similar system of facet descriptors for use with the FoodEx2 food list that is used as the basis for dietary intake, exposure and risk assessments. LanguaL codes are included in EFSA's FoodEx2 browser tool to allow easy matching to data that already includes LanguaL codes. The use of such facetted descriptive terms is a powerful tool to aid electronic searching of foods, e.g., for matching composition data to food intake data. The relatively recent FoodON initiative¹¹ has developed a new ontology, initially built around LanguaL terms, that represents knowledge about food that is built to interoperate with the OBO (Open Biological and Biomedical Ontology) Library and to represent entities that represent 'food'. The FoodON ontology

⁹ www.langual.org

¹⁰ Ireland and Møller, 2010;

¹¹ www.foodon.org

covers food safety, food security, the agricultural and animal husbandry practices linked to food production, culinary, nutritional, and chemical ingredients and processes.

5.1.1 Food production and processing

Processes used for food production and for home preparation of food for consumption may have a significant impact on the nutrient content and bioavailability of nutrients. Therefore, it is important for these processes to be documented in food datasets (see food description and classification, section 3.1.1).

EFSA's Foodex2 system for food classification¹² includes facet descriptors that can be used to capture the main processes used in food production. EFSA guidelines for EU Menu methodology for production of food consumption data recommend that information should be recorded for the following facets:

- source e.g., plant or animal origin
- part-consumed (i.e., skin or visible fat consumed)
- preparation/processing method
- cooking method and reheating
- preservation method
- qualitative information on the food e.g., fat, sugar, salt and caffeine content of the food
- sweetening agent
- fortification
- packaging material

EFSA Foodex2 facet descriptors are closely mapped to LanguaL facets which also allow detailed description of production and processes. LanguaL also includes a very comprehensive 'Treatment applied' facet that includes codes for processes that add, remove, or modify food components and ingredients.

5.1.2 Source of foods

The geographic origin of foods is important for purposes including trade, labelling, and food integrity (safety, quality, and authenticity). LanguaL includes a 'Geographic places and region' facet that includes climatic zones, continents, countries, regions, and fishing zones.

GS1 uses the Global Location Number (GLN) to identify their locations. The GLN can identify a company's physical locations, for example a store, a warehouse, or a berth in a port and can be used to identify an organisation as a corporate entity. The GLN is encoded in either a barcode or EPC/RFID tag to automatically identify a location.

¹² <https://www.efsa.europa.eu/en/data/data-standardisation>

5.2 Composition

5.2.1 Nutrients

Data on nutrient composition of foods is predominantly provided in the form of food composition datasets. Traditionally there wasn't a standard structure for food composition data because datasets were compiled independently for publication in country specific printed tables in books and scientific journals. Since the introduction of computerised data compilation and publication, there has been a trend towards more harmonised data structures and control of data quality through clear structured documentation of the data. The range of foods and components included is usually based on resources available to compilers and also longevity of datasets, with compilers prioritising foods that are commonly consumed and foods and nutrients that are nutritionally important¹³. Data quality is associated with various factors (including food description, component identification, sample collection, sample handling, analytical method, and laboratory performance) related to derivation of values¹⁴.

Nationally representative datasets have been produced in most developed countries and increasingly in developing countries, although coverage of foods and nutrients may be more limited. The quality of published data is high, and the data are typically produced, managed and published by groups with a high level of sustainable expertise in food composition. Sources of data used within the datasets include: analytical data produced specifically for the dataset using standard methods in accredited laboratories; data from scientific or grey literature, calculations from ingredients of composite foods, manufacturers' data, and data from other national datasets. Data is managed and published using a range of data handling tools, most commonly relational databases, although some datasets are still compiled and published in spreadsheet format. Although data standards do exist, the structure of food composition data is not yet fully standardised and depends on the source of the data and the knowledge and expertise of data compilers. Most national food composition datasets are freely available to use, even for commercial purposes. National food composition datasets are usually available for searching or download from the national dataset curators and/or publishers in standard formats (e.g., spreadsheet, database, xml) that could be incorporated into a data platform. Dataset compilers are beginning to produce APIs that enable their data to be incorporated into applications and websites (e.g., United States Department of Agriculture FoodData Central¹⁵, Swedish National Food Administration¹⁶). Documentation to describe how datasets are compiled and presented, including information on data sources and definition of components is available for most datasets and can help inform standardization and comparison of data.

There have been many collaborative projects and networks of food composition data compilers that have aimed to improve consistency and harmonization of food composition databases, so that values from different datasets are of comparable structure and quality. European projects such as EuroFOODS, Cost Action 99, the IARC European Nutrient Data Bank project^{17,18} and the work of the INFOODS (International

¹³ Greenfield & Southgate, 2003

¹⁴ Westenbrink et al., 2016

¹⁵ <https://fdc.nal.usda.gov/api-guide.html>

¹⁶ <https://www.livsmedelsverket.se/om-oss/psidata/livsmedelsdatabasen>

¹⁷ Slimani et al., 2007a

¹⁸ Slimani et al., 2007b

Network of Food Data Systems) organization¹⁹ network all made progress towards more standardised production, compilation, and management of data. These and other related projects, summarised in 'The production, management and use of food composition data'¹³ were used as the basis for the European Food Information Resource (EuroFIR) project. The EU FP6 and FP7 EuroFIR Network of Excellence (NoE) (2005–2010) and EuroFIR NEXUS (2011–2013) projects aimed to standardise and harmonise food composition data in Europe through improved data quality, database searchability and standards. To further standardise the EuroFIR quality approach, new or existing procedures and tools were developed or adopted for data interchange, food description, component identification, value documentation, recipe calculation and quality evaluation of values (<http://www.eurofir.org/our-resources/noe-and-nexus-projects/>). EuroFIR produced a range of tools to help data compilers, including procedures for documenting data values, and supported the development and publication of a European standard for food data (EN 16104:2012, Food data - structure and interchange format) that was based on the EuroFIR technical standard^{20,21,22}. INFOODS also produces guidelines for data compilation, online resources, and coordinates training activities, particularly aimed at developing countries²³.

The CEN standard²² also took into account recommendations of the GS1 Global Data Synchronisation Network (GDSN) Trade Item standard Food & Beverages extension²⁴ used in the retail industry, and the European Food Safety Authority (EFSA) Guidance on Standard Sample Description for Food and Feed²⁵ that applies to chemical contaminants and residues included in monitoring and control programs. The CEN standard is therefore a flexible standard that can support not only food nutrient and bioactive data exchange, but also data on feed and data concerning other food properties (e.g., allergen or micro-organism contents, pH, vitamin retention factors). The CEN standard allows use of a range of different thesauri, but current international training programs are focused on thesauri published by EuroFIR⁶ and by INFOODS (especially outside of Europe). While not completely identical, these thesauri have many features in common and are continuing to be harmonised so that European and International standards are compatible. A feature of these, and other thesauri (e.g., EFSA Standard Sample Description for Food and Feed, FoodEx2), is that they can easily be mapped from one to another so that datasets using different thesauri can be exchanged relatively easily. Available thesauri and classification systems that are commonly used include:

- Food classification and description (see section 3.1.1)
 - Component identification
 - Value documentation
 - Value type
 - Units
 - Matrix unit
 - Method indicator
 - Acquisition type

¹⁹ www.fao.org/infoods

²⁰ Becker et al., 2008

²¹ Becker, 2010

²² CEN, 2012

²³ <http://www.fao.org/infoods/infoods/en/>

²⁴ <http://www.gs1.org>

²⁵ EFSA, 2010

- Reference type
 - Recipe calculation

Thesauri used for value documentation are relatively standardised and contain commonly used terms that allow consistent and clear expression of nutrient values. Standard thesauri are routinely used by almost all national food composition data compilers because most compilers have been trained through either EuroFIR and/or INFOODS training programs. Data produced outside of national programmes, including data published in some scientific journals, is typically less standardised because producers are often unaware of available standards.

Details of EuroFIR thesauri are available at <http://www.eurofir.org/our-resources/eurofir-thesauri/>.

Details of INFOODS standards and guidelines are available at <http://www.fao.org/infoods/infoods/standards-guidelines/en/>.

Even though nutrient composition datasets are easily available, accessing and combining data from different datasets is not an easy task. The EuroFIR FoodEXplorer tool²⁶ is an innovative interface, which can be accessed online and allows users to simultaneously search harmonised nutrient composition datasets²⁷. The values available within FoodEXplorer are provided by national compilers but are presented in a harmonised format and search results can be downloaded into formats that enable further use. Even though the presentation of results is harmonised, there are still some differences in how values are compiled at national level and further work to fully standardise data are ongoing.

Food composition datasets usually consist of data for generic food products with only limited data for specific branded food products. Nutrient composition of branded food products is widely and freely available, from retailer and manufacturers websites and from compiled data sources (e.g., Open Food Facts²⁸, The Open Food Repo²⁹ and can potentially be combined with data for generic foods. However, most sources do not provide data that allow easy assimilation into other datasets, with branded food data generally being limited to components that are needed for food labelling. There are commercial sources of some retailer and manufacturer datasets, but they are expensive to license. Some data sources (e.g., Tesco (UK)³⁰, Brandbank³¹) have made APIs available to users to allow product information, including nutrient content, to be embedded into software applications.

GS1 has its own range of data standards that include food and beverages. The Global Data Synchronization Network (GDSN) is an internet-based, interconnected network of interoperable data pools governed by GS1 standards³². The GDSN enables food businesses to exchange standardised product data with their trading partners. The GDSN is used as a tool to support high data quality through use of authoritative data sources, real-time data synchronization, and standardization of data formatting³³. The core standards document of the GS1 system is the GS1 General Specifications that describe how GS1 barcodes and identification keys should be used. The GS1 system Architecture describes the component parts of the GS1

²⁶ <http://www.eurofir.org/our-tools/foodexplorer/>

²⁷ Finglas, et al 2014

²⁸ <https://world.openfoodfacts.org/>

²⁹ <https://www.foodrepo.org/>

³⁰ <https://www.tescolabs.com/category/api/>

³¹ <https://www.brandbank.com/>

³² <https://www.gs1.org/standards>

³³ <https://www.gs1.org/standards/gdsn>

system, how they are related and how they are related to standards published by other organisations such as ISO, UN/CEFACT or W3C. Food products are classified according to the Global Products Classification (GPC) standard. Properties of foods and beverages are described by the GS1 web vocabulary standard. Transactional information exchanged between GS1 partners is enabled by GS1 XML (Electronic Data Interchange) standards.

5.2.2 Bioactive components

Data on bioactive components in food is also available through EuroFIR via the eBASIS database^{34,35} and ePlantLibra³⁶. Data on polyphenols can also be found in the Phenol-Explorer³⁷ database, which is an online database with free and unrestricted access for all users. Some data on bioactive compounds (e.g., carotenoids) is also included in food composition datasets.

There is no specific standardised data structure for bioactive compound data and each database is able to set-up the data structure and fields according to their own needs. However, since bioactive compounds are food constituents, data structures used would be compatible with the food data standards described for nutrients. For example, the two principal resources for data on polyphenols in foods, EuroFIR's eBASIS database and Phenol-Explorer (coordinated by a team at INRA), both use the LanguaL system for classification and description. Compound classifications are not necessarily consistent between datasets and data extracted from different databases may not be directly comparable.

5.2.3 Contaminants

Contaminants may be present in foods both intrinsically and extrinsically and can be considered as food components. Contaminant data can be managed and exchanged using the same data systems as nutrients. Some contaminant data, e.g., for inorganic components including arsenic, lead, cadmium and mercury, is already included in some national food composition datasets.

Data is usually collected as part of specific dietary monitoring studies that are used to calculate population exposure to contaminants and the associated risk to the population. Total diet studies (TDS) capture a population's intake of certain contaminants and nutrients to identify potential detrimental risks from contaminants in the food chain. They are used primarily by governments as a public health safeguard and to investigate newly emerging contaminants³⁸. As part of the TDS-Exposure project (EU 7th Framework Programme 2012-2016), a module for management of TDS data was developed for the FoodCASE system to manage food composition and consumption data³⁹. The basis of the FoodCASE TDS module is the data structure used for food composition data, but a decision was made to also support management of information on the processes used within a TDS. The process information relates mainly to the purchasing and sampling of foods that are used for the samples that are finally analysed to produce values for contaminants.

³⁴ <http://www.eurofir.org/our-tools/ebasis/>

³⁵ Plumb et al., 2017

³⁶ <http://www.eurofir.org/our-tools/eplantlibra/>

³⁷ <http://phenol-explorer.eu/>

³⁸ Pite et al 2018

³⁹ Presser et al., 2016

5.2.4 Data Exchange Model

5.2.4.1 Food entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
foodId	Unique food identifier	M; U	NUM
country	Country code - Alpha-2 code from ISO 3166	M	THS
code	The food code, ID, or abbreviation used to identify the food in the data set	M	STR
name	Food name in original language, as specified by language attribute using the language code (ref. ISO 639), see FoodName entity in the EuroFIR Standard technical annex (2008) [am4].	M	STR
engName	Food name in English, with preference given to British English., see FoodNames entity	M	STR
sciName	The scientific name should add here to the following format: Genus species Author [, Year] e.g. GadusmorhuaLinnaeus, 1758, see FoodNames entity	O	STR
codex	Codex Alimentarius Food Standards code, see the Codex Alimentarius Standards list	O	THS
gs1	Global Trade Identification Number	O	THS
eNumber	If food is food additive, code according to the European E-Number system for additive standardization (included in LanguaL facet A). For more information, see the European Commission's web site on food additives	O	THS
ins	If the food is a food additive, code according to the International Numbering System for food additives according to CODEX Alimentarius. For more information, see Codex Alimentarius General Standard for Food Additives	O	THS
group	Original Food Group Code - The code from the classification system used in the original dataset	O	STR
servingSize	Refer to amounts, in grams, dl, etc., as specified by the producer	O	NUM
servingPack	Servings per pack - This can only be filled in, if you are dealing with a single article/package	O	INT
servingSuggest	Serving suggestions	O	STR
prepState	State of Preparation - Prepared or unprepared (True/False)	O	BL
productYield	Product yield after preparation / serving (cooking, dilution, draining) - Quantity after preparation	O	NUM
allergenInf	Allergen information: allergen type, allergen level - Claim(s) concerning possible allergen(s) in the food	O	STR
dietClaim	Dietary claim or use - Health, nutrition, or other claim(s), like religious claim(s)	O	STR
edibleNature	Nature of Edible Portion - Which parts of the food are included in Edible Proportion	O	STR
wasteNature	Nature of Waste - Which parts of the food are not edible and have been removed as waste, e.g., rind, bone, stone, peel, liquid from can, outer leaves, etc	O	STR
colour	Colour values are currently not further specified. More detailed recommendations are planned for further versions	O	STR
producer	Names the manufacturer(s) or producer(s) of the food(s) e.g., a farmer is considered a producer	O	STR
distributor	Use when e.g., producer not stated	O	STR
remarks	Any further remarks, or free text description of food item	O	STR

5.2.4.2 Component Values Entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
foodId	Link to Food Entity	M; U	Food
componentId	Component value ID	M; U	NUM
componentCode	Component code from EuroFIR Thesauri	M	THS
valueId	Value ID - Required only if individual sample results are also interchanged as Contributing Values	O	INT
selectedValue	The value that is considered the best representative according to the decision of the data compiler, previously referred to as "Best Location". This value may be based on an evaluation of analytical results from one or more samples, a composite sample or derived from literature data. The characteristics of the value are given by the Value Type. Additional properties, e.g., statistical, can be given (see Statistical Values). Generally, this property is required. In some cases, however, it might not be possible to assign it (e.g., value unknown, value undecidable because of distribution showing more than one cluster of values). Selected Value may then be left empty with Statistical values given (e.g., minimum, maximum)	O	NUM
unit	Unit	M	THS
matrix	Matrix Unit	M	THS
valueType	The Value Type is designed to further describe the value in Selected Value or to give a qualitative description of the value when no Selected Value can be given	M	THS
selectedAcqType	SelectedValue acquisition type - Gives categories for the origin of a value, e.g., an evaluated food composition database or table, a scientific publication, analytical results commissioned by the compiler or results calculated by the compiler. Some Acquisition Types usually relate to the Reference associated with the value	M	THS
acqType	Value acquisition type - Gives categories for the origin of a value, e.g., an evaluated food composition database or table, a scientific publication, analytical results commissioned by the compiler or results calculated by the compiler. Some Acquisition Types usually relate to the Reference associated with the value	M	THS
genDate	The date when this particular value was generated, e.g., date of analysis or compilation	O	DAT
evalDate	The most recent date the value in question was evaluated or validated	O	DAT
remark	Any further remarks, or free text description for the value	O	STR
contribValuesNum	Number of values contributing to the value given as Selected Value. Defined as the number of analytical portions of the food or the number of contributing values (e.g., values taken from food composition tables)	O	int
analyticPortionSize	The size of the sample, prepared from the laboratory sample, from which test portions are removed for testing or for analysis	O	NUM
analyticPortionReplicates	The number of times (sub-) samples of a specific analytical sample is being analysed	O	INT
mean	Mean value - The arithmetic mean value of the statistic	O	NUM
median	Median value - The median value of the statistic	O	NUM
min	Minimum value -The minimal value within the statis	O	NUM
max	Maximum value- The maximal value within the statistic	O	NUM
standardDev	Standard Deviation - Should be used for normal distributions only	O	NUM

standardErr	Standard Error	O	NUM
methodId	Reference to the method used to generate value	O	<i>Method</i>
referenceId	One or more Reference, using the Reference Code(s) of the associated References record(s). See Reference Entity	O	<i>Reference[]</i>

5.2.4.3 Method Entity

Attribute	Description	(M)andatory/ (O)bligatory (U)nique	Type
id	Primary key of Method Specification entity	M	STR
type	Type of Method used to generate the value. From the list given in EuroFIR Method Type thesaurus	M	Thesauri (Method Type)
indicator	From the EuroFIR Method Indicator thesaurus	M	Thesauri (Method Indicator)
parameter	Further method information for calculation methods, e.g., NCF and FACF. Mandatory only for calculated protein and fatty acid values	M	STR
name	The name of the analytical/calculation method in the English language (preference to British English)	M	STR
originalName	The name of the analytical/calculation method in the original language	M	STR
officialCode	Gives abbreviation for official method or accreditation organization / system, e.g., AOAC 985.29 method, NMKL (Nordic) system, COFRAC 60 accreditation (France). This may be multiple occurring data, if more than one organisations standard is listed.	O	STR
description	Free-text field describing the method	O	STR
remarks	Any further remarks, or free text description for the method	O	STR
referenceId	Link to Reference entity providing information about the method used by the laboratory	O	<i>Reference</i>

5.2.4.4 Reference Entity

Attribute	Description	(M)andatory/ (O)bligatory (U)nique	Type
code	Reference standard code	M; U	STR
type	Gives categories for types of documents or 'publication media' according to the list given in Reference Type thesaurus	M	THS
acqType	Nature/origin of reference (e.g., food composition table, food label, independent laboratory) given in Acquisition Type thesaurus	M	THS
citation	Reference citation in a format currently used for scientific publications	M	STR
www	The internet address (URL) of the file / data source (WWW or FTP)	O	STR
remarks	Any further remarks, or free text description for the reference	O	STR

5.2.4.5 Sender Information Entity

Attribute	Description	(M)andatory/ (O)bligatory (U)nique	Type
country	Country code from the ISO Thesauri	M	THS

sender	Person(s) who provided the dataset	M	STR
organisation	Organisation that provided the dataset	M	STR
superOrganisation		M	STR
postalAddress	Postal address of organisation that provided dataset	M	STR
telephone	Telephone number or organisation that provided dataset	M	STR
email	Email address of person/organisation that provided dataset	M	STR
www	url of organisation that provided dataset	M	STR
remarks	Any further remarks on the sender	O	STR
content	Description of dataset content	M	STR
shortContent	Short description of dataset content	M	STR
responsibleBody	Organisation responsible for the dataset	M	STR
legalRestrictions	Legal restrictions related to use of dataset	M	STR
summary		M	STR
bibliographicRef	Citation reference for the dataset	M	STR
remarks	Any further remarks on the bibliographic reference	O	STR
creationReason	Reason for creation of dataset	O	STR
language	Language(s) of the dataset	O	STR
datasetCreated	Date of dataset creation	O	DAT
acqType	Thesauri (Acquisition Type)	M	THS

5.2.5 Thesauri

Name	Reference
Country Code	https://www.iso.org/obp/ui/#search https://www.iso.org/iso-3166-country-codes.html
CODEX	http://www.fao.org/fao-who-codexalimentarius/codex-texts/list-standards/en/
GS1	https://www.barcodefaq.com/1d/gs1-id-key/
E-Number	https://www.food.gov.uk/business-guidance/approved-additives-and-e-numbers
INS	http://www.fao.org/tempref/codex/Meetings/CCFAC/ccfac31/INS_e.pdf
Component	http://www.eurofir.org/eurofir-thesauri/
Unit	http://www.eurofir.org/eurofir-thesauri/
Matrix Unit	http://www.eurofir.org/eurofir-thesauri/
Value Type	http://www.eurofir.org/eurofir-thesauri/
Reference Type	http://www.eurofir.org/eurofir-thesauri/
Method Type	http://www.eurofir.org/eurofir-thesauri/
Method Indicator	http://www.eurofir.org/eurofir-thesauri/
Acquisition Type	http://www.eurofir.org/eurofir-thesauri/

5.2.6 APIs

5.2.6.1 Food APIs

HTTP method	URI	Description	Result
GET	/food	List of foods	Food[]
GET	/food?name={name}	List of foods that match name	Food[]
GET	/food?comp_code={code}&comp_value={value}&comp_op={operation}	List of foods that have specific amount of component	Food[]
	code – EuroFIR code, value – numeric value, operation – one of: >, <, =		
GET	/food/{foodId}	Single Food by Id	Food

5.2.6.2 Component APIs

HTTP method	URI	Description	Result
GET	/component	List of component entities	Component[]
GET	/component?name={name}	List of component entities that match name	Component[]
GET	/component/{code}	Single Component by EuroFIR code	Component

5.2.6.3 Methods APIs

HTTP method	URI	Description	Result
GET	/method	List of methods	Method[]
GET	/method?name={name}	List of methods that match name	Method[]
GET	/method?type={method_type}	List of methods belonging to specific type	Method[]
GET	/method?indicator={indicator}	List of methods belonging to specific indicator	Method[]
GET	/method/{id}	Single Method by Id	Method

5.2.6.4 Reference APIs

HTTP method	URI	Description	Result
GET	/reference	List of references	Reference[]
GET	/reference?citation={citation}	List of references that match citation	Reference[]
GET	/reference?code={code}	Single Reference by Code	Reference

5.2.6.5 Sender Information APIs

HTTP method	URI	Description	Result
GET	/sender	Information about dataset sender	Sender

5.3 Labelling

Information on food packaging labels is provided according to legislation⁴⁰ and is intended for ‘protection of consumers’ health and interests by providing a basis for final consumers to make informed choices and to make safe use of food, with particular regard to health, economic, environmental, social and ethical considerations. The regulations apply to all foods intended for the final consumer, including foods delivered by mass caterers, and foods intended for supply to mass caterers.

Labelling information provides data that is important for accurate description of foods and includes information on:

- Nutrient composition (mandatory and voluntary)
- Portion size
- Ingredients
- Country of origin
- Place of provenance
- Allergens
- Information on processing (e.g., added water, date of freezing)

5.3.1 Branded Food

Although Branded Food Data might be considered a subset of Food Composition data, and many compilers store and manage their Branded Food Data alongside their regular Food Composition data, using the same dataset scheme and software, there are some aspects of Branded Foods that aren’t that covered by the proposed Food Composition exchange model and some attributes of Food Composition aren’t applicable to Branded Foods.

Branded Food data is mainly collected from:

- packaging (by taking photos and transferring the data manually or by using text recognition software)
- retail store websites (by copying the text manually or automatically by APIs or web scraping)
- directly from producers (in Excel files or another file format)

Regardless of the source, in general, for branded datasets there is less information available about how the component values were obtained (no information available about analytical methods used, no detailed method indicators can be assigned, no sample pooling occurs, etc.) and there is more information connected to the food, for example additives used, packaging material, allergens and so on. Branded food usually only includes values for nutrients that are required for food labelling, meaning that values for

⁴⁰European Commission, 2011

minerals, vitamins, fatty acids, and individual sugars are not often available. The lack of values for many nutrients is a significant limitation if data is to be used for calculation of nutrient intakes and these values must be estimated from other sources (e.g., from values for generic foods, by calculation based on ingredients).

Food Composition datasets usually only have one or two languages used (national language and English language), and two languages are only used for the Food Name. Branded foods, more often than not, have the full label information in several language (e.g., Swiss foods have packaging information in at least four languages: DE, FR, IT, EN).

There are some identifiers that do not apply to generic or aggregated food composition datasets, like the GS1 standards. Including information to branded foods records like:

- GLN (Global Location Number)
- GTIN (Global Trade Item Number)
- SSCC (Serial Shipping Container Code)
- GRAI (Global Returnable Asset Identifier)
- GIAI (Global Individual Asset Identifier)
- GSRN (Global Service Relation Number)

These add another dimension to the branded foods, uniquely distinguishing all products, logistic units, locations, and assets across the supply chain from manufacturer to consumer.

5.3.1.1 Data Exchange Models

https://www.gs1.org/sites/default/files/gs1_xml_technical_user_guide_to_release_3_i2.pdf

5.3.1.1.1 Food entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
country	ISO 3166 code	M	THS
id	Primary key	M;U	KEY
name	Food name	M;U	STR
englishName	Food name in English	M;U	STR
timestamp	date and time of data collection	M	DAT(TIM)
technicalName		O	STR
ingredientList	List of ingredients listed on the package	O	STR
additions	Additions listed on the package	O	STR
allergenInformation		O	STR
servingsPerPack		O	NUM
servingSuggestions		O	STR
finalPreparation		O	STR
samplingPlace	place of data collection, for example a store name	O	STR
producer	Producer name	O	STR
dietaryClaims	Health&nutrition claims	O	STR
netWeight	Net weight value	O	NUM
netWeightUnit	Net weight unit	O	THS
marketShare	market share value	O	FRC
physicalState	Cooled, Frozen, Dried, Uncooled, Conserved/can, Others	O	THS
packagingMaterial	Plastic, Metal, Paper	O	THS
producer	Format: Producer/subproducer	O	STR
producerInformation	Name and address of the producer	O	STR
productLine	Format: Product line/subproduct line	O	STR
brand		O	STR
brandInformation		O	STR
distributor	Distributor information	O	STR
sourceId	If the data was imported from a different DBMS, the ID can be kept	O	STR
GTIN	Global Trade Item Number from the GS1 standard	O	STR
density		O	STR
marketShareValue	% of market share	O	FRC
marketShareType	Market share of product or category	O	THS
purchasedQuantityValue		O	NUM
purchasedQuantityUnit		O	THS
foodEx2	FoodEx2 exposure and facet classification	O	THS
image	Images of the food/package	O	FILE

5.3.1.1.2 Food Classification systems

- FoodEx2
- LanguaL
- own classification

GPC (Global Product Classification) bricks classifies products by grouping them into categories based on their essential properties as well as their relationships to other products. The GPC covers a wide range of products sold around the world and has also a segment dedicated to Food/Beverage/Tobacco. The most basic building block of GPC is an eight-digit numeric code, known as a Brick code. There are Bricks for everything from a car to a bottle of milk.

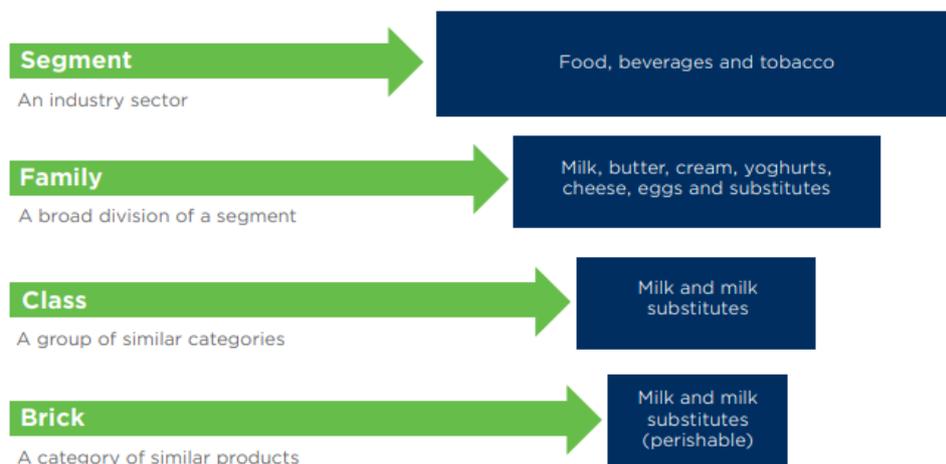


Figure 6. GPC hierarchy.

5.3.1.1.3 Component entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
foodId	key for food table	M	FKEY
componentId		M	KEY
componentCode	From the EuroFIR Thesauri	M	THS
value	value for this component in this food	M	NUM
unit		M	THS
matrixUnit	From the EuroFIR Thesauri	M	THS
valueType	From the EuroFIR Thesauri, AR (as reported) set by default	M	THS
acqType	From the EuroFIR thesauri, L (Food label, product information) set by default	M	THS

5.3.1.2 Thesauri

For Branded data GS1 and EuroFIR thesauri should apply. Additional thesauri might be needed for fields like packaging material or physical state.

5.3.1.3 API

The API should be identical to the Food Composition API, just different data might be returned. An example API that allows branded data transfer can be found here:

http://zswt2.foodcase-services.com:51455/FoodCASE_WebServiceBranded/#/default/postJson

5.4 Authenticity

Determining the authenticity of foods means to detect

1. mislabelled food,
2. substitution of original ingredients by cheaper ones,
3. not declared additives (adulterants) and processing practices, and
4. incorrect labelling or reporting of geographic and species origin, and method of production.

To control the authenticity of foods and detect food frauds, development and application of advanced analytical tools and infrastructure is essential. Research has developed the following methods using the most widely instrumental techniques presented in Figure 7:

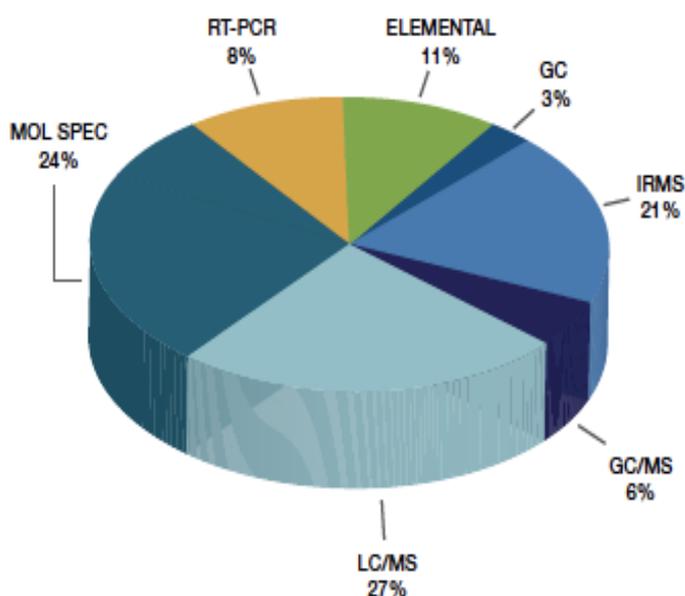


Figure 7. Most widely instrumental techniques for monitoring adulteration and verifying authenticity.

- Stable isotope analysis (light elements – H, C, N, O and S; heavier elements – Sr, Pb)

The potential to determine the geographical origin of animal or plant derived material using stable isotope signatures is well established in food authentication studies. The transfer of isotope signals from the bio-elements (H, C, N, O, S) present in local feed and water to animal and plant tissue is understood and forms the basis of the approach. Similarly, the geochemical isotope signature of a particular region can be traced through the use of strontium isotope analysis (and sulphur to a lesser extent).

A fundamental part of this approach to determine the geographical origin of a suspect food sample in an objective way is by statistical comparison of its stable isotopic composition to a database of samples of which the geographical origin is known.

- Elemental analysis

As with stable isotopes, element concentrations in plants and animals are mainly related to the geological and microclimatic characteristics of the site of cultivation and farming practices.

- DNA speciation

DNA-based methods are particularly suited to determining if meat from a particular species of animal is present in a product.

- Metabolomics

Metabolomics is the systematic study of the unique chemical fingerprints that specific cellular processes leave behind. The metabolome represents the collection of all metabolites in a biological cell, tissue, organ, or organism, which are the end products of cellular processes.

- Metagenomics

Metagenomics is a technique whereby DNA is extracted from environmental samples and sequenced. Bioinformatics techniques then are used to identify the microbial species from which the DNA was derived, including species that never have been cultivated in the laboratory.

- Spectroscopic techniques – fast screening methods, such as FTIR, RAMAN

A common theme of food authentication and traceability studies is the requirement for a database of genuine samples to which the sample can be compared to establish its authenticity. The same database could be further used for geographical origin determination. For effective data processing, the large number of independently measured parameters is needed from which the most crucial are element and isotope parameters. Parameters can be then statistically evaluated in order to identify key tracers that differentiate the regions or countries of interest.

EU-funded projects, where these kind of methods and supporting tools have been developed and evaluated, include: ERA Chair ISO-FOOD⁴¹, ESFRI Metrofood-PP⁴², RealMED⁴³. In ISO-FOOD, i) an ontology representing knowledge base within the domain of isotopes for food science, named ISO-FOOD, was created by Eftimov et al⁴⁴, while in RealMED ii) a food authenticity database and a data management and exploration tool, named FoodTrack, were developed⁴⁵. The ISO-FOOD ontology and the FoodTrack database and tool are relevant for the FNS-Cloud stakeholders as well, therefore the underlying data model is described in more details in the following subsection of this deliverable.

5.4.1 Data Exchange Model

⁴¹ <http://isofood.eu>

⁴² <https://www.metrofood.eu>

⁴³ <https://realmedproject.weebly.com>

⁴⁴ Eftimov et al., 2019

⁴⁵ <http://foodtrack.ijs.si>

For determining the authenticity and traceability of foods, isotopes of light elements need to be determined in selected food commodities. The results of such measurements are collected and, in most cases, organised using a relational database. However, when the results should be compared with data collected from other laboratories, it is required first to structure the data using the same format and meta-data, which is used to describe the measurements performed. In most cases, this is done by manually preparing the data sets that need to be compared. To integrate the datasets with a minimal effort, we propose a formal representation of the isotopic knowledge in the domain of food research and identify the requirements for such representation. We have respected the opinion recently published in Pauli et al.,⁴⁶ where the authors explain why a centralised repository for isotopic data is required and provide a shared vision for IsoBank that would offer a viable and powerful framework to organise, consolidate, and share stable isotope data across disciplines.

The ISO-FOOD ontology dedicated to food is described in detail in Eftimov et al.,⁴⁴. Our aim is to link it with the FNS-Cloud ontology (to be developed in WP3, Task 3.3). In figure 8, meta-data (including provenance information) for describing the measurement of isotopes in food/drink items is presented.

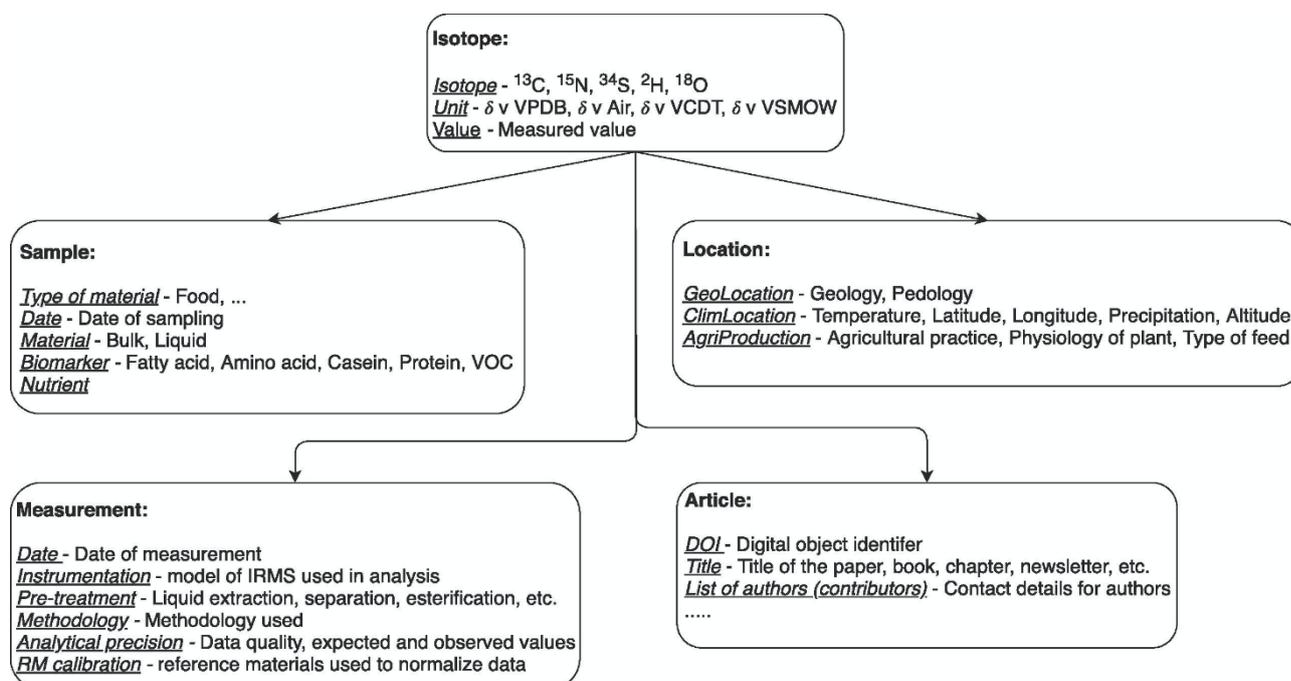


Figure 8. A schematic representation of the proposed database structure, outlining how contributors and users would interface with samples, analyses, measurements, and datasets for describing stable isotope data

5.4.1.1 Isotope data model

Code	Name	Description	Type	Mandatory
------	------	-------------	------	-----------

⁴⁶ Pauli et al., 2017

isotope	Isotope	Isotope measured in the food	Thesaurus	Yes
unit	Unit	Unit	Thesaurus	Yes
value	Value	Measurement value	Float	Yes

5.4.1.2 Sample data model

Code	Name	Description	Type	Mandatory
typeOfMaterial	Type of material	Food	String	Yes
date	Date	Date of sampling	Date	Yes
material	Material	Material (e.g., bulk, liquid)	Thesaurus	Yes
biomarker	Biomarker	Fatty acid, amino acid, casein, protein, VOC	Thesaurus	Yes
nutrient	Nutrient	Nutrient	Thesaurus	Yes

5.4.1.3 Measurement data model

Code	Name	Description	Type	Mandatory
date	Date	Date of measurement	String	Yes
instrumentation	Instrumentation	Model of IRMS used in analysis	Thesaurus	Yes
preTreatment	Pre-treatment	Freeze-dry, liquid extraction, separation, esterification, etc.	Thesaurus	Yes
methodology	Methodology	Methodology used	Thesaurus	Yes
analyticalPrecision	Analytical precision	Data quality, expected and observed values	Thesaurus	Yes
RMcalibration	RM calibration	Reference materials used to normalise data	Thesaurus	Yes

5.4.1.4 Geographical location data model

Code	Name	Description	Type	Mandatory
geology	Geology	Geology	Thesaurus	Yes
pedology	Pedology	Pedology	Thesaurus	Yes

5.4.1.5 Climate location data model

Code	Name	Description	Type	Mandatory
temperature	Temperature	Temperature	float	Yes
latitude	Latitude	Latitude	float	Yes
longitude	Longitude	Longitude	float	Yes
precipitation	Precipitation	Precipitation	float	Yes
altitude	Altitude	Altitude	float	Yes

5.4.1.6 Agriculture production

Code	Name	Description	Type	Mandatory
------	------	-------------	------	-----------

agriculturalPractice	Agricultural practice	Agricultural practice	Thesaurus	Yes
physiologyOfPlant	Physiology of plant	Physiology of plant	Thesaurus	Yes
TypeOfFeed	Type of feed	Type of feed	Thesaurus	Yes

5.4.1.7 Article data model

Code	Name	Description	Type	Mandatory
doi	DOI	Digital object identifier	String	Yes
title	Title	Title of the paper, book, chapter, newsletter, etc.	String	Yes
listOfAuthors	List of authors (contributors)	Contact details for authors	String	Yes

5.4.2 Thesauri

5.4.2.1 Isotope

Isotope	Description – Natural abundance in %
¹ H	99.9885
² H	0.0115
¹² C	98.93
¹³ C	1.07
¹⁴ N	99.632
¹⁵ N	0.368
¹⁶ O	99.757
¹⁷ O	0.038
¹⁸ O	0.205
³² S	94.93
³³ S	0.76
³⁴ S	4.29
³⁶ S	0.02

5.4.2.2 Unit

The ratios between heavier and lighter isotope (²H/¹H, ¹³C/¹²C, ¹⁵N/¹⁴N, ¹⁸O/¹⁶O, ³⁴S/³²S) are expressed in δ-notation in ‰.

5.4.2.3 Material

Material	Description
Bulk	Sample as a whole – usually solid, pulp
Liquid	Liquid samples and/or water extracted from the sample

5.4.2.4 Biomarker

Biomarker	Description
FattyAcid	Fatty acid
AminoAcid	Amino acid
Casein	Casein
Protein	Protein
VOC	Volatile Organic Compounds

5.4.2.5 Nutrient

For the list of nutrients, the EuroFIR component Thesauri is used.

5.4.2.6 Instrumentation

Instrumentation	Description
	Model of IRMS used in analysis
EA-IRMS	Elemental Analyser coupled to Isotope Ratio Mass Spectrometer
TC/EA Pyrolysis IRMS	High Temperature Conversion Elemental Analyzer coupled to Isotope Ratio Mass Spectrometer
TG-IRMS	Trace Gas unit coupled to Isotope Ratio Mass Spectrometer
MultiFlow Bio	Equilibration Unit coupled to Isotope Ratio Mass Spectrometer
LC-IRMS	Liquid Chromatography coupled to Isotope Ratio Mass Spectrometer
GC-C-IRMS	Gas Chromatography coupled to Isotope Ratio Mass Spectrometry through combustion (C) unit
DI-IRMS	Dual-Inlet Isotope Ratio Mass Spectrometer

5.4.2.7 Pre-treatment

Pre-treatment	Description
Freeze-dry	Freeze-dry
Demineralization	Demineralization using 1M HCl
Distillation	Purification of alcohol
Extraction	Liquid-liquid extraction, solid phase extraction
Separation	Separation techniques: adsorption, centrifugation, chelation, chromatography
Esterification	Esterification for fatty acid analysis

5.4.2.8 Analytical precision

Analytical precision	Description
0.1‰	$^{18}\text{O}/^{16}\text{O}$ – water, $^{13}\text{C}/^{12}\text{C}$ - ethanol
0.2‰	$^{13}\text{C}/^{12}\text{C}$ - bulk, protein, casein, pulp
0.3‰	$^{15}\text{N}/^{14}\text{N}$ - bulk, $^{36}\text{S}/^{34}\text{S}$ - bulk, $^{18}\text{O}/^{16}\text{O}$ – bulk
0.5‰	$^{13}\text{C}/^{12}\text{C}$ - biomarkers
1.0‰	$^2\text{H}/^1\text{H}$ - bulk, water

5.4.2.9 RM calibration

For the list of reference materials, the thesauri from Reference Material are used (more details in Section 5.5.2.1).

5.4.2.10 Geology

Geological background	Description
Igneous	Basic: basalt, dolerite, gabbro; Intermediate: andesite, micro-diorite, diorite; Acid: rhyolite, micro-granite, granite
Limestone	Carbonate sedimentary rock
Shale/Mudstone/Clay/Loess	Fine-grained terrigenous clastic rocks

5.4.2.11 Pedology

Soil Classification	Description
Gravel	> 50% of coarse fraction retained on 4.75 mm sieve, very small, irregular pieces of rock and stone
Sand	granular material composed of finely divided rock and mineral particles
Silt	is granular material of a size between sand and clay, whose mineral origin is quartz and feldspar
Clay	a finely-grained natural rock or soil material that combines one or more clay minerals
Organic	Peat, decomposition of plants and animals

5.4.2.12 Agricultural practice

Agricultural practice	Description
Integrated	cropping methods and other agricultural production techniques which fulfil both ecological and economic demands
Conventional	no restriction in pesticides and mineral fertilisers use
Organic	application of naturally derived products as defined by organic certification programs (animal manure, compost); ecological, biological

5.4.2.13 Physiology of plant

Physiology of plant	Description
C3	Plants use C3 photosynthesis - Calvin pathway
C4	Plants use C4 photosynthesis - Hatch-Slack pathway
CAM	Plants use crassulacean acid metabolism – C3-C4 intermediate photosynthesis

5.4.2.14 Type of feed

Type of feed	Description
C3	Feed based on C3 plants: grass
C4	Feed based on C4 plants: maize
aquaculture	Farmed feeding of fish
wild	Wild feeding of fish

5.4.3 API

http method	URI	Description
Get	foodtrack.ijs.si/api/{commodity name}	Returns a list of samples and their data matching the food commodity name
Get	foodtrack.ijs.si/api/{commodity name}?isoFlag={}&species={name of species}&smpType={sample type}&country={country name}&location={name of location}&latitude={lat. coordinates}&longitude={lon. coordinates}&altitude={altitude}&temperature={temperature}&precipitation={precipitation}&year={year of sampling}	Returns a filtered list of samples and their data matching the food commodity name. Filters can be omitted. The isoFlag determines whether to include isotopes/macro/essential/environmental/geological/toxic properties of the given sample.

5.5 Food safety

5.5.1 Analytical Data and total diet study (TDS) data

Analytical data and TDS data are analyte or compound values which are analytically measured in laboratories.

For food composition data, a data level model exists from Greenfield and Southgate¹³, which was taken over by EuroFIR²⁰. This model defines the levels: data source, archival data, reference database and user database/printed and computerised tables. The model represents the data life cycle for food composition data. The data source level represents public and private technical literature containing analytical data in its original format while in the archival data, the original data is transposed into an electronic system without amalgamation or modification but scrutinised for consistency. The reference level is the pool where values have been converted into standard units and nutrients are expressed uniformly. Finally, the user database contains weighted and averaged data that are fit for users' purpose. For instance, a food item does not contain 20 different protein values but only one and the 20 source values are aggregated into one representative value.

While the food composition exchange format is targeted on the user database level, this section is focusing on the exchange of data from the first two levels which is analytical data. In addition, the food composition data contains a thesaurus with about 1000 nutrients while analytical data contains around 19,400 analytes.

It must also be mentioned that two datasets are merged here into one for simplification and because they logically belong together. One dataset is about the samples and their sub-samples and the information about their selection, collection, storage, and handling. The other dataset is about the data that analytical laboratories produce which is information about the analytical method, the analytical approach, and the results. Sometimes the laboratories have no information about the samples and only know the food matrix and a sample code. But sometimes these two steps are done in close collaboration and all information is available.

The EuroFIR data structure, in particular the FDTP XML format, could be used as a data exchange format. But the FDTP XML format showed some issues in practice because the data files reached several 100 MB and programmes had problems handling them. In general, XML is more verbose than JSON and harder to understand for data managers and programmers. The new EuroFIR EXCEL format, which forms the basis for the food composition data exchange format, is targeted for the user database level and, for instance, does not contain information about the samples.

EuroFIR thesauri are very helpful and will be taken over where appropriate.

In the recent months, the FoodCASE discussion group within EuroFIR made remarkable advances in regard to sample and analytical data management. FoodCASE is an information system designed to manage food composition, consumption, TDS, and branded data. The food composition module showed in different countries that not all information can be stored or cannot be stored in a unified way and some deviations were recently accepted. Together with the analytical and documentation discussion groups, a new approach was defined and agreed upon⁴⁷. This new approach takes care of many details which were not

⁴⁷ Matusczak et al., 2020

considered in many other data formats. The new approach, for instance, has entities for samples, subsamples, primary samples, analytical values, and replicate values and documents the differences between these entities. It was discussed with many European organisations working with occurrence data and is based on the EuroFIR and the TDS-Exposure data models, improving them in several places.

In the TDS-Exposure project⁴⁸, a data format for internal storage was defined in a software specification for FoodCASE⁴⁹. The entities and their relations were evaluated with different TDS experts and tested in several studies during the project. The EuroFIR data model was used as a basis and extended where necessary for TDS. The new model was later extended in the German MEAL study⁵⁰. This model was the first that differentiated between sample and subsamples and introduced an additional level of detail. It also introduced vial entities that are sent to labs for analysis.

The EFSA SSD2 format^{51,25,52} was already described in the section about reference material. Its comprehensive structure and thesauri definitions makes it a good candidate also in this data area. SSD V1 and SSD V2 were designed for reporting occurrence data. Also, the support and maintenance of EFSA is promising that the format will be sustainable.

5.5.1.1 Recommendation

Although the EFSA format is a strong candidate, the new initial data level format defined in the FoodCASE discussion group of EuroFIR is more promising. The reason is the more detailed entity-relation structure reflecting the complex procedure between sampling and laboratory analysis and the latest knowledge and experience from several country organisations went into this approach. The new approach explains visually many of the entities and their relation and is therefore more understandable. The new approach is more comprehensive and the EFSA model can easily be derived. It is therefore recommended to use the new EuroFIR approach in FNS-Cloud and to use the SSD2 format when exchanging data with EFSA.

5.5.1.2 Data Exchange Model

5.5.1.2.1 Sampling

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
sourceId	Identifier in the source system	O	STR
id	Unique identifier	M; U	STR

⁴⁸ <http://www.tds-exposure.eu/>

⁴⁹ Presser et al., 2012

⁵⁰ <http://www.bfr-meal-studie.de/en/the-bfr-meal-study.html>

⁵¹ EFSA, 2014

⁵² EFSA, 2020

name	Name for the sampling	O	STR
reason	Overall context for the project and thus sampling procedure (e.g., contamination study, consumption survey...)	O	STR
strategy	Short description of the sampling strategy	O	STR
country	Country in which the sampling procedure took place	O	THS
reference	e.g., a publication in which the sampling plan is described	O	STR

5.5.1.2.2 Sample

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
sourceId	Identifier in the source system	O	STR
id	Unique identifier	M; U	STR
code	The code, ID, or abbreviation used to identify the sample in the dataset	M	STR
name	Name of the initial food the Sample is connected to	M	STR
engName	Food name in English, with preference given to British English	M	STR
sampleCode	Unique Sample Code that links the Sample with its AVs (and Primary Samples). A Sample can have only one AV per component	M; U	STR
amount	Sample amount	O	NUM
unit	Unit for amount	O	THS
compositeSample	Determines if this is a composite or a single sample	O	BLN
homogenizationEquipment	Use of mills and sample dividers	O	STR

homogenizationCooling	Use of cooling agents like liquid nitrogen, dry ice	O	STR
sampleHandling	General handling of sample before arrival at laboratory, e.g., sample transport, storage conditions and duration.	O	STR
dateOfArrivalAtLab	Date of sample arrival at the laboratory	O	DAT
labStorage	Conditions of sample storage at the lab (e.g. -20°C freezer)	O	STR
remarks	Free text field for remarks	O	STR
samplingId	Reference to sampling entity	O	NUM

Analytical Values are obtained from Samples that can consist of a single or multiple Primary Samples. Primary Samples are for example different brands of chocolate bars, that are then mixed into a single (lab) Sample. This pooling is normally done to reduce the number of samples and the costs for laboratory analysis when budget is an issue. If a Sample was created using several Primary Samples, pooling information needs to be added. This kind of Sample is a Composite Sample. Samples that are made of one Primary Sample (Primary Sample = Sample) will be called Single Samples.

5.5.1.2.3 Subsample

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
sourceId	Identifier in the source system	O	STR
id	Unique identifier	M; U	STR
sampleSourceId	Reference to the sample	M	STR
unitDescription	Description of primary sample size e.g. Can of tomatoes, bag of rice, piece of apple and loaf of bread	M	STR
noOfUnits	Number of bought primary sample unit e.g. 2 cans of tomatoes	M	NUM
unitAmount	Package weight/volume e.g. Can of tomatoes has 280 g	O	NUM
unit	Unit from label e.g. Can of tomatoes has 280 g	O	THS
productName	Name of the sampled product	O	STR
genericName	e.g. sales description of the food item	O	STR
producer		O	STR
brand		O	STR

cultivar/breed/variety/species	Additional field for describing whether different cultivars of apples or meat from different breeds were sampled	O	STR
husbandry	Type of farming	O	THS
countryOfOrigin	ISO 1366	O	THS
dateOfSampling		O	STR
placeOfSampling		O	STR
region	Region of sampling, e.g. Bavaria	O	STR
ingredients	List of ingredients from the package	O	STR
expirationDate	Given on the package	O	STR
storageInStore	e.g. ambient, frozen	O	STR
packaging	Type of packaging (glass, paper, plastic, can...)	O	STR
incomingStorage	Storing of primary samples before any homogenisation step	O	STR
remarks			
processingType	Food processing	O	STR
processingTime	Duration of processing	O	STR
processingTemperature	Temperature of processing	O	STR
addedIngredients	Added ingredients during processing, e.g. fat and salt	O	STR
cutting	Information about e.g. whether the sample is peeled	O	STR
images		O	FILE

If we have a composite sample, we need additionally the contribution information:

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
amount	Contributing amount to the pooled sample	M	NUM
unit		M	THS
contributionRemarks		O	STR

5.5.1.2.4 Nutrient value

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
subsampleId	Reference to the subsample id	M	STR
nutrient	termExtendedName of PARAM catalogue	M	THS

value		O	NUM
valueType		M	THS
unit		O	THS
matrixUnit		O	THS

5.5.1.2.5 Analytical value

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
id	Unique identifier	M; U	STR
sampleId	Reference to the sample id	M	STR
analyte	termExtendedName of PARAM catalogue	M	THS
value	Numeric value of the analytical value (AV), created based on the technical replicate measurement (RM) values. Usually calculated as a mean of the RMs.	O	NUM
unit	Unit of the AV (for example [g]), list of available selections shared with the Unit Thesauri for Values.	O	THS
matrixUnit	Matrix Unit of the AV (for example “per 100g edible portion”), list populated by the EuroFIR Matrix Unit Thesauri.	O	THS
valueType	Value type of the AV (for example “mean”), list populated by the EuroFIR Value type Thesauri.	M	THS
methodId	Reference to method used for analysis	M	THS

5.5.1.2.6 Technical replicate

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
analyticalValueId	Reference to the analytical value id	M	STR

value	Numeric value of the analytical value (AV), created based on the technical replicate measurement (RM) values. Usually calculated as a mean of the RMs.	O	NUM
unit	Unit of the AV (for example [g]), list of available selections shared with the Unit Thesauri for Values.	O	THS
matrixUnit	Matrix Unit of the AV (for example “per 100g edible portion”), list populated by the EuroFIR Matrix Unit Thesauri.	O	THS
valueType	Value type of the AV (for example “mean”), list populated by the EuroFIR Value type Thesauri.	M	THS

5.5.1.2.7 Method

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
analyticalValueId	Reference to the analytical value id	M	STR
name	The name of the analytical/calculation method in English	M	STR
originalMethodName	The name of the analytical/calculation method in the original language	M	STR
officialMethod		O	THS
valueType	Value type of the AV (for example “mean”), list populated by the EuroFIR Value type Thesauri.	M	THS

5.5.1.3 Thesauri

5.5.1.3.1 Nutrient and analyte

For the list of analytes, the PARAM catalogue from EFSA is used. The catalogue has around 20,000 entries and therefore it will not be listed here. The catalogue can be downloaded over the EFSA Catalogue Browser⁵³.

An issue is that food composition data uses the EuroFIR component thesaurus. We therefore need a mapping of the PARAM catalogue and the EuroFIR component thesaurus.

5.5.2 Metrology

METROFOOD-RI – is a research infrastructure that aims to promote scientific excellence in the field of food quality and safety, which will provide high level metrology services for enhancing food quality & safety and supporting the traceability and sustainability of the agri-food systems⁴². It comprises an important cross-section of highly interdisciplinary and interconnected fields throughout the food value chain, including agri-food, sustainable development, food safety, quality, traceability and authenticity, environmental safety, and human health. The mission is to enhance quality and reliability of measurement results and make available and share data, information, and metrological tools, in order to enhance scientific excellence in the field of food quality and safety.

METROFOOD-RI consists of a service-oriented architecture providing a platform for sharing and integrating data, knowledge and information about food analysis, food composition (nutrients and contaminants) and markers. The intention is to collate analytical results and merge with existing data and provide tools for use.

METROFOOD-RI was cited as an emerging project in the ESFRI Roadmap 2016 and entered the Roadmap 2018⁵⁴.

5.5.2.1 Reference Material Data

Reference material and certified reference material in relation to food data are controls or standards used to check the quality and metrological traceability of products, to validate analytical measurement methods, or for the calibration of instruments. Certified reference materials are produced with a certificate of values, uncertainty, and metrological traceability. Accreditation can be achieved by national and/or international standards such as ISO/IEC 17025.

In general, a reference material is similar to food composition data, TDS data or branded food data where a food item has specific characteristics, in particular content of chemical substances. It is therefore possible to apply one of these data models, use their thesauri and API definitions but we do not recommend it, because some attributes are different, and terminology is also different.

In its early phase project, METROFOOD-RI defined a data model for reference material.

⁵³ EFSA. 2019

⁵⁴ <https://www.esfri.eu/esfri-roadmap>

The data model is not yet published but can be used in its current version 4 for the FNS-Cloud. The data model is based on data models from EuroFIR²⁰ and TDS-Exposure^{46,47} and has entities data source, reference material, values, method, proficiency testing, sample, subsample, analytical limits, and thresholds. METROFOOD-RI also implemented some APIs but did not yet harmonise the signatures of these methods. METROFOOD-RI is currently relying on the substance thesaurus of SSD2, the so-called PARAM catalogue⁵¹ with around 20'000 entries, defined some own thesauri and uses some of the EuroFIR thesauri.

The SSD2 format from EFSA was created to report occurrence data to EFSA from several food safety domains²⁵. The following diagram (Figure 9.) shows the relations between the entities.

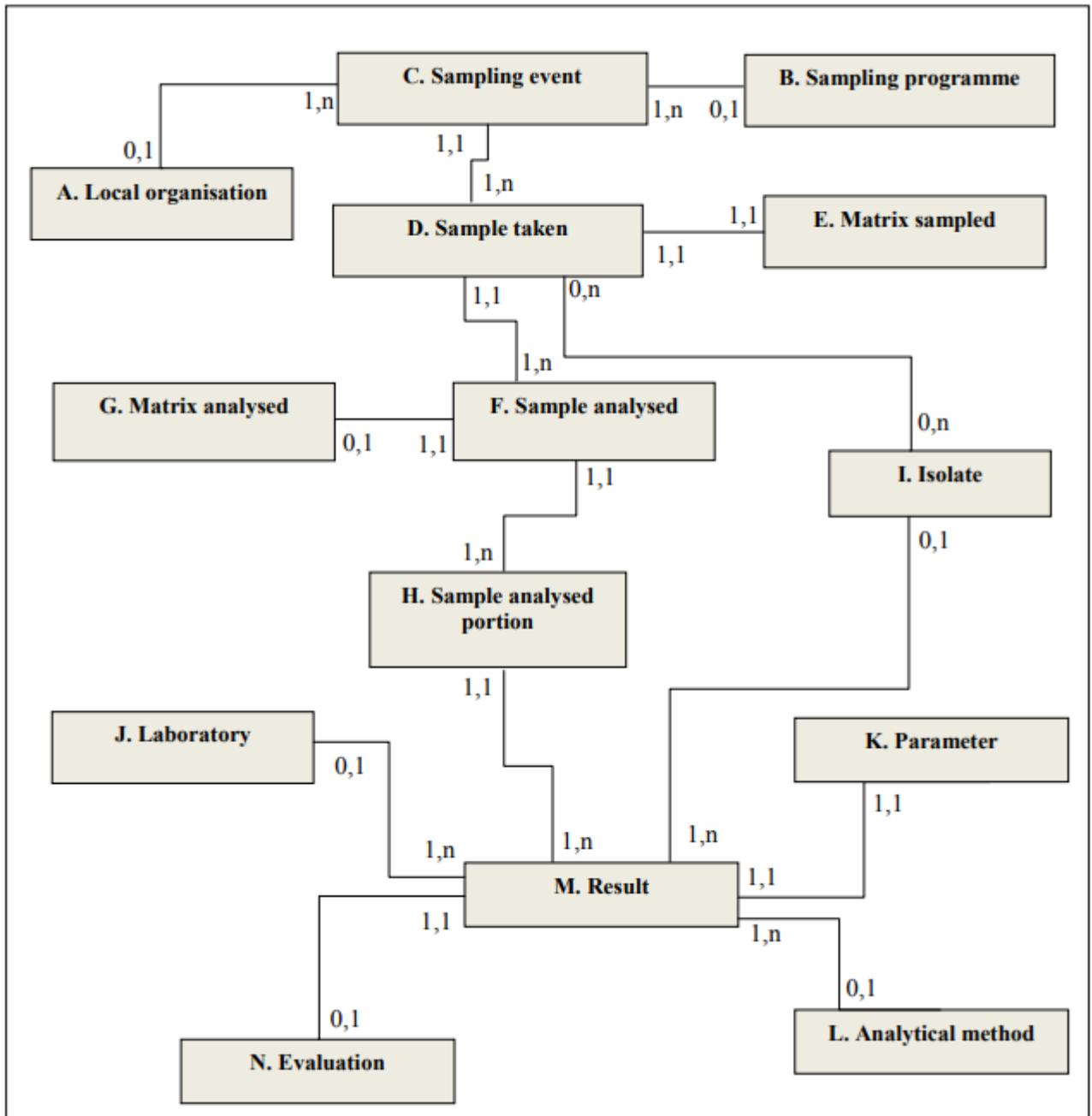


Figure 9. Structure of the main SSD2 entities

The SSD2 format is comprehensive and well elaborated. It contains many thesauri including the PARAM thesaurus, which represents the reportable chemical substances and contains around 19,400 substances. There are no API definitions for SSD2.

In addition to the big METROFOOD-RI initiative, there are many online and offline, private, and public shops where reference materials can be bought. A public example is the certified reference materials catalogue of the European Commission (EU) Joint Research Center (JRC)⁵⁵.

Legal notice | Cookies | Contact | Search | English (en) ▾

Shopping cart

JOINT RESEARCH CENTRE
Certified reference materials catalogue

Home Frequently asked questions Legal notice My account/login Contact us

European Commission > JRC Science Hub > Knowledge>Reference and Measurement > Reference Materials
Home » Search (apple) » ERM-BC516 APPLE (dietary fibre)

ERM-BC516 APPLE (dietary fibre)
Article in stock

How to order
Search tips
Catalogue/price list (pdf)
Accreditation
How to read our certificates
User support / Application Notes
Development of GMO CRMs
News

Put in shopping cart

EUR 130,00

Proximates and conventional properties

Search CRMs
Find

Browse CRMs
» By application field
» By analyte group
» By material/matrix

Downloads for this reference material
Certificate ERM-BC516 certificate.pdf
Certification report ERM-BC516_report.pdf

Product information	
CRM code	ERM-BC516
Description on the invoice	DRIED APPLE
Sales unit	bottle
Net mass	25
Gross mass	120
Mass unit	Gram (g)
Storage temperature	-20 °C

Figure 10. An example of dietary fibre in an apple from the EU JRC certified reference materials catalogue.

Another interesting reference material online shop is labmix24.com where reference materials from different providers like NIST, NCS or IAEA are collected and can be ordered⁵⁶. At the time of writing, no information was provided if the web app live queries all these reference material providers or if the web app has an internal database where samples can be managed.

Several of these online shops were investigated to see if they publish information about an API or data exchange model. None of the investigated online shops provide any information and it must be assumed that proprietary data models and APIs are used.

⁵⁵ <https://crm.jrc.ec.europa.eu/>

⁵⁶ <https://www.labmix24.com/crmsearch/>

There are some sources for the thesaurus of chemical substances available. Already mentioned was the PARAM catalogue from the EFSA Data Collection Framework (DCF). Another one is ChEBI which stands for Chemical Entities of Biological Interest, provided by the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI). ChEBI is a freely available, manually annotated database of small molecular entities⁵⁷. EMBL-EBI and ChEBI are part of the ELIXIR Research Infrastructure.

5.5.2.1.1 Recommendation

METROFOOD-RI is on the way to become a self-sustainable and long-term operational Research Infrastructure and its data model and thesauri are therefore the preferred choice. The ChEBI thesaurus is comprehensive but with the focus on chemistry and bioscience provides more information than needed for this data area. In the context of reference material, the SSD2 format is not completely fitting. Some entities and attributes are not needed while others like price are missing.

The data model and the thesauri are listed in the following sections which correspond to the unpublished METROFOOD-RI version at the time of writing but for the future, publications of METROFOOD-RI should be followed. The API is directly derived from food composition.

5.5.2.1.2 Data Exchange Model

5.5.2.1.2.1 Data Source

Code	Name	Description	Type	Mandatory
organisationName	Organisation name	Name of the sender organisation.	STR	Yes
datasetName	Dataset name	Name of the dataset. Can be used to indicate version of the dataset, e.g. CH food comp V5.2	STRHS	Yes
country	Country	ISO Alpha-2 code of country where food comes from.	Thesaurus	Yes

⁵⁷ <https://www.ebi.ac.uk/training/online/course/chebi-quick-tour/what-chebi>

5.5.2.1.2.2 Reference Material

Code	Name	Description	Type	Mandatory
sourceId	Source unique identifier	Unique identifier in the source system.	STR	Yes
sourceCode	Source code	Code or abbreviation from the source system.	STR	No
name	Name of the RM	Name of the RM as registered by the producer	STR	Yes
englishName			STR	Yes
foodMatrix	Food matrix	Name of the food matrix.	STR	Yes
englishFoodMatrix	English food matrix	English name of the food matrix.	STR	Yes
description	Description	Description of the RM	STR	No
producer	Producer	Producer of the RM	STR	Yes
country	Country	Country of the RM producer.	THS	No
physicalForm	Physical form	Solid, liquid, frozen, lyophilised etc.	THS	No
rmType	Type of RM	Type of the RM (Reference Material, Certified Reference Material)	THS	No
amount	Amount	Amount of the RM aliquot	INT	No
amountUnit	Amount unit	Unit of amount	THS	No
price	Price	Price of the RM	Price	No
priceUnit	Cost unit	Unit of the RM prize	THS	No
quantityInStock	Quantity in stock	This is the amount if available in storage.	INT	No
generationDate	Generation date	Date when RM values were produced.	DAT	No
validDate	Valid until date	Date until RM values are valid.	DAT	Yes
availableDate	Available until date	Date until when RM is available.	DAT	No
recomStorCond	Recommended storage condition	Recommended storage condition after receiving RM.	STR	No
remark	Remarks			No

5.5.2.1.2.3 Reference Material Value

Code	Name	Description	Type	Mandatory
sourceId	Source unique identifier	Unique identifier in the source system.	STR	Yes
rmlId	Reference material id	Link to RM	STR	Yes
analyte	Analyte	Name of the analyte for which the RM is characterised.	EFSA SSD2 thesaurus	Yes
valueCharacter	Character of the analyte value	Possible entries are: certified value, reference value or information value	THS	No
value	Value	The value of the property.	NUM	Yes
unit	Unit	Unit of measurement of the value.	THS	Yes
matrixUnit	Matrix unit	Matrix unit of the value.	THS	Yes
Method	Method	Methodology used to generate value.	THS	No
uncertainty	Uncertainty	Uncertainty of value.	NUM	No
maxUncertainty	Max uncertainty	Maximum uncertainty for value.	NUM	No
number	Number of contributing values	Number of data points producing value.	INT	No
mean	Mean		NUM	No
median	Median		NUM	No
minimum	Minimum		NUM	No
maximum	Maximum		NUM	No
stdDev	Standard deviation		NUM	No
stdError	Standard error		NUM	No
remark	Remarks		STR	No

5.5.2.1.2.4 Method

Code	Name	Description	Type	Mandatory
sourceId	Source unique identifier	Unique identifier in the source system.	STR	Yes
name	Method name	Name of the analytical method in English	STR	Yes
officialMethod	Official name	Gives abbreviation for official method or accreditation organisation/system, e.g. AOAC 985.29 method, NMKL (Nordic) system, COFRAC 60 accreditation (France). This may be multiple occurring data, if more than one organisations' standard is listed.	STR	No
methodType	Method type	One of the following entries: reference, official, standard method or nothing.	THS	No

description	Description	Free-text field describing the method.	STR	No
foodMatrix	Food matrix	Name of the food matrix for which method can be used.	STR	No
analytes	Analytes	Analyte for which the method can be used.	Collection of thesauri entries	No
year	Year	Year when method was introduced.	INT(4)	No
references	References	Publication about method. Link to Reference entity.	Collection of strings	No

5.5.2.1.3 Thesauri

5.5.2.1.3.1 Physical Form

Physical form	Description
Solid	Solid form
Liquid	Liquid form
Frozen	Frozen
Lyophilised	Lyophilised

5.5.2.1.3.2 Type of RM

Type of RM	Description
Reference Material	Not certified RM
Certified Reference Material	Certified RM

5.5.2.1.3.3 Analyte

For the list of analytes, the PARAM catalogue from EFSA is used. The catalogue has around 20, 000 entries and therefore it will not be listed here. The catalogue can be downloaded (<https://github.com/openefsa/catalogue-browser/wiki>) over the EFSA Catalogue Browser⁵¹

5.5.2.1.3.4 Character of the analyte value

Character of the analyte value	Description
Certified value	A certified value has the highest confidence in its accuracy in that all known or suspected sources of bias have been fully investigated.
Reference value	A reference value is a best estimate where not all known and suspected sources of bias have been investigated.
Information value	An information value is a value that is of interest, but not sufficient information is available to assess uncertainty.

5.5.2.1.3.5 Method type

Method type	Description
Reference method	
Official method	
Standard method	

5.5.2.1.4 API

http method	URI	Description
Get	/rm/{rmlId}	Finds reference material by Id
Get	/rm	Returns a list of reference material
Get	/rm?name={name}	Returns a list of reference material matching the name
Get	/rm?englishname={english name}	Returns a list of reference material matching the English name
Get	/rm/{rmlId}?relations={relations}	Returns a reference material found by Id with the given relations

6 Food Intake and Lifestyle Data

'Food intake and lifestyle' data include data related to food consumed by individuals, including food intake, food purchase and food preparation information. Often lifestyle data is collected as part of the food intake database which can consist of demographic data (e.g. age, sex, location, family status), anthropometry data (e.g. weight, height, waist circumference), medical data (e.g. disease risk factors, medical history, BP, biochemical status (e.g. lipids, glucose), family history of disease), socio-economic data (e.g. education, salary, qualification, job), activity data (e.g. physical activity (type and time spent), sedentary activity (type and time spent), and sleep (type and time spent)). Some field studies in WP4 collect some of this lifestyle data and these parts are included in the data exchange model below, directly attached to the subject entity.

Food intake data can be collected using a wide range of methods and software tools. The applicability of a selected method depends on the intended use of the data, including the study population, and sometimes on the resources available to users. As with production and publication of food composition data, there have been efforts to standardise approaches to dietary assessment methods used in national dietary monitoring surveys. There are many software tools available for dietary assessment, with different levels of complexity/types of data and functionality, depending on their purpose and intended user groups.

6.1 Consumption and lifestyle data

6.1.1 EFSA EU Menu guidelines

Harmonisation of methods used for collecting high quality food consumption data for use in dietary exposure assessments has been a priority for EFSA. Furthermore, guidance on "General principles for the collection of national food consumption data in the view of a pan-European dietary survey" was published in 2009, and a pan-European food consumption survey, also known as the "EU Menu", was launched. The guidance has since been updated by the EFSA Evidence Management Unit (DATA) and the EU Menu Working Group and has been endorsed by the EFSA Network on Food Consumption Data. The current guidance⁵⁸ provides recommendations for the collection of more harmonised food consumption data among the EU Member States for use in dietary exposure assessments of food-borne hazards and nutrient intake estimations under the remit of EFSA's scientific panels.

The guidance includes recommendations on the methods to be used and instructions for data to be collected in accordance with the FoodEx2 classification system. The guidance focusses on methods that should be used and the data that should be collected but does not specify particular data structures. Specific software is not recommended but the recommendations for use of data collection and/or dietary software focus on the desired features that are required for the collection of high-quality data. Detailed recommendations are included for food description and classification (FoodEx2), portion size determination and for collecting information on consumption of food supplements and background

⁵⁸ EFSA, 2014

information of the survey subjects. There is also a general recommendation for data transfer (EFSA data transmission schema).

Software features that are required to meet EFSA guidelines include:

- Input of food, beverages and food supplements consumed during the survey days in accordance with common 24-hour recall rules (e.g. applying the “Multiple-Pass method”).
- During data entry, it should be possible to automatically search, describe and quantify each item using the national quantification methods (e.g. a validated picture, national standard portions based on real weights and similar). However, regardless of how the information is collected, this is converted to a gram amount for each food included in nutrient assessment.
- The level of description of foods should be based on, or at least be compatible with, the FoodEx2 facet descriptor system and should include the following minimum features:
 - The tool should cover descriptor information indicated in the recommendation for food descriptions
 - All self-made composite dishes should be disaggregated and described on the component/ingredient level “as purchased”/“as ingredient” if possible
 - In the case of single and composite foods, and the ingredients of composite dishes, information should be collected on:
 - source e.g. plant or animal origin (if not implicit from the name, FoodEx2 source facet);
 - part-consumed (i.e. skin or visible fat consumed, FoodEx2 part-consumed-analysed facet);
 - preparation/processing method (FoodEx2 process facet);
 - cooking method and reheating (FoodEx2 process facet);
 - preservation method (FoodEx2 process facet);
 - qualitative information on the food e.g. fat-, sugar-, salt- and caffeine-content of the food (FoodEx2 qualitative-info facet e.g. full-fat, semi-skimmed, without added sugar, caffeine-free etc.);
 - sweetening agent (FoodEx2 sweetening-agent facet, information may be added at a later stage by national food consumption/composition data experts);
 - When a food/recipe is quantified, the system is able to automatically convert food quantities “as reported” to “as finally consumed” (e.g. cooked and/or without inedible part) or this step must be provided at a later stage of the data collection process using pre-defined algorithms and standard food-specific coefficients (e.g. raw-to-cooked yield factors, density, or edible part coefficients) which can be easily updated.
 - Checks (electronic or checks by interviewer), pathways to be followed during data input and probing questions, so as not to miss the collection of mandatory information and foods that are easily forgotten (e.g. beverages, snacks, food supplements). Systematic quality controls should be performed throughout the data input procedure. The software should check systematically for all information reported by the subject and entered into the program so that possible errors and suspicious answers and outlier values can be detected and clarified with the subject during or after the survey. It must be possible to check for outliers and inconsistencies in the national database. Additional quality checks based on energy values of foods and intake per day are considered an asset.

- Maintenance procedures for the different databases must be ensured. Like any open-ended method, the databases should be updated regularly so that new foods, recipes, and other information reported by the study subjects can be added. To maintain a high level of control and standardisation of the different databases and to facilitate updating, it is important that only one version of the software is available in each country, and that any modifications to the country-specific files are centralised at country level and, if possible, collected at European level.
- The program must keep track of the interview time per subject and provide output files for the valid population on all consumption events including details on all foods, beverages and supplements consumed, such as recipes as available, food descriptors, portions consumed, place of consumption, time of consumption etc. and, preferably, energy and macronutrient intake data in electronic format. These output files must be functional to perform statistical analyses. The software must allow storage, output, and export of the different databases in a standardised way in accordance with the EU Menu data transmission schema.
- The following mandatory databases should be incorporated in the software program: FoodEx2 food descriptors (or compatible), portion sizes, standard recipes, and yield factors. Energy composition of foods for quality control measures is considered an asset. If not included in the supporting documents provided by EFSA, the databases must be developed at country level.
- The software provider should set the minimum training period needed and plan the training content needed to successfully use the software, to become familiar with the protocols and databases incorporated in the tool, and to be able to successfully perform interviews in an acceptable time and facilitate the “training of local trainers”. Controlled test interviews should be included and performed with acceptable quality during the training and before carrying out survey interviews.
- The following optional databases could also be included in the software tool: energy requirements, food supplement composition, brand, and packaging information. All these databases should be developed at the country level.

These guidelines imply that software for food consumption that is intended to collect and report data at this level, must have a complex data structure and development of such software is highly specialised. Software that is intended to collect personal consumption data (e.g. apps for fitness, lifestyle, food choice, weight loss), may be less complex and may not include more specialised features such as food classification and description codes, although it would be expected that the basic data structure would be similar.

6.1.2 Software examples

There are numerous examples of software available for collection and analysis of food intake, many of them web based or available as apps. The applicability of methods and software features depends on the purpose of data collection and the intended use. Factors include what is being studied (e.g. specific components, types of foods), population group (e.g. number of participants, age, country, characteristics) and the timeframe for the study. The Diet@Net (DIETary Assessment Tools NETwork) partnership, which included experts in the field of dietary assessment, nutritional epidemiology, public health, and clinical

studies from 8 UK universities and institutes, developed the Nutritools⁵⁹ website which supports dietary assessment through guidance and access to validated interactive dietary assessment tools⁶⁰.

Examples of commonly used software tools that are compatible with the requirements for data collection and management of food intake data, including linking to food composition data, for the EU Menu project include:

GloboDIET is a computerised program to collect 24-hour dietary recall interview, either face-to-face or by telephone, and food dairies using its data entry application, which has been developed at the International Agency for Research on Cancer (IARC), together with multiple end-users and multidisciplinary partners⁶¹. It is a highly standardised methodology which has been designed, validated, and implemented as a reference methodology for both European nutritional epidemiology studies and for future Europe-wide nutritional surveillance^{62,63}. The GloboDIET methodology has been successfully implemented in European nutritional epidemiology projects, such as the European Prospective Investigation into Cancer and Nutrition (EPIC) calibration study and the European Food Consumption Validation (EFCOVAL). In addition, it is used as a common standardised dietary methodology in seven National Surveillance systems in Europe (Austria, Belgium, France, Germany, Malta, The Netherlands, and Switzerland) under the double EU-Menu and GloboDIET umbrella. The GloboDIET methodology is fully compliant with EU menu recommendations for dietary software for dietary assessment^{56,64} and it has already been customised, validated, and tested for risk assessment and food safety, through different EU projects i.e. EFCOVAL, PANEU, PANCAKE.

The GloboDIET databases consist of about 70 databases (“Common” and “Country-specific”) on foods, recipes, and dietary supplements as well as on related information such as place and time of food consumption, special diets etc. Common databases are the backbone of the standardisation within and between countries, for instance food and recipe classifications, facets/descriptors to describe foods in a comparable way, quantification methods and probing questions. In contrast, country-specific databases are meant to capture the differences in diet existing within and between countries, for instance, food and recipe lists; brand name lists; facets and probing, which are specific to foods and recipes; quantification methods for each food item and recipe; coefficients for edible portion; cooking; density; fat left on the dish; default ratios for fat, sauce and sweeteners added to foods and recipes; and databases for quality controls. These dataset formats are ‘.end’ (easily opened with text or word file) or ‘.xls’, and systematically labelled in local language as well as in English, to ease data reading, exchanges and pooling.

Diet Assess and Plan (DAP) software was designed to meet public health nutrition challenges in Central Eastern European Countries (CEEC) and Balkan countries⁶⁵. DAP is a platform for standardised and harmonised food consumption collection, comprehensive dietary intake assessment and nutrition planning. Its unique structure enables application of national food composition databases from the European food composition exchange developed by EuroFIR and in addition allows communication with

⁵⁹ <https://www.nutritools.org/>

⁶⁰ Hoosen et al., 2019

⁶¹ Slimani et al., 2000

⁶² de Boer et al., 2011

⁶³ Slimani et al., 2011

⁶⁴ Gavrieli et al., 2014

⁶⁵ Gurinović et al., 2018

other tools. DAP is used for daily menu and/or long-term diet planning in diverse public sector settings, foods design/reformulation, food labelling, nutrient intake assessment and calculation of the dietary diversity indicator. DAP is compatible with the EuroFIR technical standard for food composition and with the requirements for EFSA's EU Menu project.

FoodCASE was originally developed as a software for management of food composition data, but it is common for food composition data compilers to also be responsible for national dietary consumption data and a module for food consumption data was added to increase compatibility. The idea was to store food composition and food consumption information together, since the food list of the food composition dataset can be used for the consumption survey and since the two datasets should be linked for calculation of nutrient intake. In addition to dietary consumption data the module also stores additional data on the people who consumed food, such as anthropometry, medical (e.g., disease and risk factors), socio-economic, physical activity. The concept of the food consumption module is that a consumption interview is a set of questions, to which the interviewee provides answers. The first set of questions concerns the foods that are consumed and the corresponding set of answers to be stored. Such food related answers contain not only information about the food and the consumed amount, but also information about consumption time and how the food amount was quantified. In addition to 'standard' questions there are many other possible additional questions (e.g., dietary habits) meaning that a default set of questions is hard to define, and flexibility and customisation options are needed. The implemented concept allows users to define their own questions, group questions and define sub-questions and the answers to each question and sub-question are then stored and attached to an interview. The tool supports different data collection types, such as 24-h recalls and a food frequency questionnaire (FFQ). However, regarding these methodologies, although levels of food information may be different (FFQs may collect food information at the food group level as well as individual foods), the data can be managed in the same way. The main difference is that, in the former case the dietary intake is specified by the absolute amount eaten, while in the second case it is specified by the frequency of intake. The concept of facets and descriptors is also a feature of FoodCASE to allow users to customise the vocabulary and assign descriptors (LanguaL or FoodEx2) to food answers.

FoodCASE is currently used worldwide to manage food data and is used by governmental organisations, research, and private institutions. Organisations that produce and manage food composition and consumption data are increasingly replacing their existing data management systems with FoodCASE because of the relatively low-cost open-source system and the network of expert users that support its continued development.

6.1.3 Data Exchange Model

EFSA's EU menu data schema for data transfer is used for transfer of data between EU Member States and EFSA. The schema has been successfully used and refined since publication and will provide a basis that covers the main types of data that are likely to be exchanged and it is therefore recommended as a model for transfer of food consumption data. If necessary, the schema could be extended for additional functionality by reference to external standards (e.g. GS1 GDSN). The schema is specific to EFSA's guidelines for national dietary monitoring surveys, and many food intake datasets from other sources will not contain all the data required by EFSA and may also contain additional data types. Therefore, some modifications may be needed so that the data model is more widely applicable, including review of mandatory and optional status of attributes.

6.1.3.1 Subject entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
Survey	Acronym of the dietary survey	M	STR
Country	Country of the dietary survey	M	THS
Orsubcode	Unique subject identifier	M;U	NUM
Panswer	Person who provided the answer	M	THS
Gender	Gender	M	THS
Birthday	Birthday	M	NUM
Birthmonth	Birth month	M	NUM
Birthyear	Birth year	M	NUM
Age	Age in years	M	NUM
Weight	Body weight in kg from the first measurement	M	NUM
Height	Height in cm from the first measurement	M	NUM
Fantday	Date of the first anthropometric measurements (day)	M	NUM
Fantmonth	Date of the first anthropometric measurements (month)	M	NUM
Fantyear	Date of the first anthropometric measurements (year)	M	NUM
Sweight	Body weight in kg from the second measurement, if any	O	NUM
Sheight	Height in cm from the second measurement, if any	O	NUM
Santday	Date of the second anthropometric measurements (day), if any	O	NUM
Santmonth	Date of the second anthropometric measurements (month), if any	O	NUM
Santyear	Date of the second anthropometric measurements (year), if any	O	NUM
Mweight	Method used to measure body weight	M	THS
Mheight	Method used to measure height	M	THS
Geo	Region, area or city of residence	M	THS
Engryintake	Average energy intake over the survey period in Kcal per day	M	NUM
Unovrep	Subject identified as under or over reporter	M	THS
Wf	Weighting factor used to normalize for age groups, gender, regions ...	M	NUM

Specialcon	Subject identified as being in special conditions	M	THS
Specdiet	Subject identified as having particular eating pattern	M	THS
Nhousehold	Size of household - Number of individuals in the household	M	NUM
Labours	Labour status of the subject (Not applicable in case of children)	O	THS
Labourm	Labour status of the mother of the subject (Only applicable in case of children)	O	THS
Professs	Professional category of the subject (Not applicable in case of children)	O	THS
Educations	Description of the current education level or highest diploma obtained by the subject	O	THS
Educationm	Description of the current education level or highest diploma obtained by the mother	O	THS
Educationf	Description of the current education level or highest diploma obtained by the father	O	THS
Activity	Description of the activity level	M	STR
Sleep	Sleep data	O	STR
Ethnic	Self-defined ethnic group	M	STR
Commentsubject	Text field to be used in order to provide additional information about the subject or to report on possible problems related to him/her.	O	STR

6.1.3.2 Consumption entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
Orsubcode	Unique subject identifier	M	NUM
Survey	Acronym of the dietary survey	M	STR
Day	Ordinal number of the survey day	M	NUM
Week	Code of the weekday of consumption	M	THS

Season	Code of the season of consumption	M	THS
Conday	Date of consumption (day)	M	NUM
Conmonth	Date of consumption (month)	M	NUM
Conyear	Date of consumption (year)	M	NUM
Exceptionday	The subject reported to have followed an exceptional diet in the specific day because of a special event (e.g. sickness, wedding party, religious event, etc.)	M	THS
Timehour	Time of consumption (hours)	M	NUM
Timeminutes	Time of consumption (minutes)	M	NUM
Meal	Code of the meal as defined within the dietary survey. If not available, the time of consumption will be used by EFSA to eventually assign eating occasion to meals.	M	THS
Place	Place of consumption in English	M	THS
Eatseq	Ordinal number of the eating occasion within the meal. Each different food, recipe and composite food determines an eating occasion.	M	NUM

Orrecipecode	<p>Unique original identifier for the recipe or composite food when applicable.</p> <p>This code must be repeated for each ingredient belonging to the recipe or composite food.</p>	M	NUM
Foodexrcode	<p>EFSA food identifier (see attached document) - Only from the "Composite dishes" category</p>	M	THS
OrrecipeDESC	<p>Description of the recipe or composite food when applicable (in the original language).</p> <p>This code must be repeated for each ingredient belonging to the recipe or composite food.</p>	M	STR
EnrecipeDESC	<p>Description of the recipe or composite food when applicable (in English).</p> <p>This code must be repeated for each ingredient belonging to the recipe or composite food.</p>	O	STR
Amountrecipe	<p>Amount consumed of the total recipe or composite food (in grams as consumed).</p> <p>This code must be repeated for each ingredient belonging to the recipe or composite food.</p>	M	NUM
Orfoodcode	<p>Unique identifier for the food or for the ingredient in case of recipe or composite food</p>	M	NUM

Amountfraw	Amount (edible) consumed of the food (ingredient in case of recipe or composite food) before processing	M	NUM
Amountfcooked	Amount (edible) consumed of the food (ingredient in case of recipe or composite food) after processing	M	NUM
Unitmeas	Unit of measurement for the amount (edible) consumed of the food or of the ingredient in case of recipe or composite food. Grams for all foods and beverages, Units for supplements and medicines	M	THS
Brand	Brand name	O	STR
Foodexcode	EFSA food identifier including all facet descriptors	M	THS
Ofacets	Facets specific for the eating occasion	M	THS
Ofacetscode	Original (National) facets identifier specific for the eating occasion, if any	O	STR
Orfacets	Facets description specific for the eating occasion in the original language, if any	O	STR
Enfacets	Facets description specific for the eating occasion in the English language, if any	O	STR

6.1.3.3 Food entity

Attribute	Description	(M)andatory/ (O)ptional (U)nique	Type
Orfoodcode	Unique original (National) food identifier	M	NUM
Survey	Acronym of the dietary survey	M	STR
orfoodname	Food description in the original language	M	STR
Engfoodname	Food description in the English language	M	STR
Foodexcode	EFSA food identifier including all facet descriptors	M	THS
Commentsfood	Text field to be used in order to provide additional information about the food (e.g. facets) or to report on possible problems related to its classification	O	STR
Energy	Amount of energy per 100 grams edible portions of the food before processing (in Kcal)	M	NUM
Water	Amount of water per 100 grams edible portions of the food before processing (in grams)	M	NUM

Fat	Amount of total fat per 100 grams edible portions of the food before processing (in grams)	M	NUM
Carb	Amount of total carbohydrates per 100 grams edible portions of the food before processing (in grams)	M	NUM
Proteins	Amount of proteins per 100 grams edible portions of the food before processing (in grams)	M	NUM
Alcohol	Amount of alcohol per 100 gram edible portions of the food before processing (in grams)	M	NUM

6.1.4 Thesauri

Thesauri used for food composition (EuroFIR, INFOODS, GS1) are compatible for use with food consumption data. EFSA specific controlled vocabulary will need to be reviewed and compared with other available controlled vocabularies that could also be used. Some controlled vocabularies used by EFSA will need to be mapped to other relevant vocabularies that are widely used in the biomedical area (see section 7.1), e.g. Unified Language Medical System⁶⁶ and SNOWMED CT⁶⁷.

Additional terms relating to the people consuming and reporting intake could be compiled based on the EFSA EU Menu data transfer schema which includes the following thesauri:

⁶⁶ <https://www.nlm.nih.gov/research/umls/index.html>

⁶⁷ <http://www.snomed.org/>

Name	Reference
Country Code	https://www.iso.org/iso-3166-country-codes.html
Person who provided the answer	A1 Subject himself/herself A2 Father A3 Mother A4 Other caretaker
Gender	G1 Male G2 Female G3 Unclassified
Method used to measure body weight/height	W1 Measured W2 Self reported W3 Unclassified
Region, area or city of residence	3rd level of the NUTS classification (Nomenclature of territorial units for statistics), http://epp.eurostat.ec.europa.eu/portal/page/portal/nuts_nomenclature/introduction
Under or over reporter	U1 Under reporter U2 Normal U3 Over reporter U4 Unclassified

Subject identified as being in special conditions	C1 Normal condition C2 Lactating C3 Pregnant C4 Chronic / long term disease C5 Unclassified
Subject identified as having particular eating pattern	D1 Normal diet D2 Vegetarian diet D3 Slimming diet D4 Diet related to health conditions (unspecified) D41 Diet related to health conditions (celiac) D42 Diet related to health conditions (diabetes) D43 Diet related to health conditions (allergy) D5 Unclassified Important: it is possible to use multiple codes in case more than one type is applicable. In this case codes should be separated by the \$ symbol. In case of unspecified health conditions please specify the reason(s) in the COMMENTSSUBJECT field

Labour status	<p>L0 Not applicable</p> <p>L1 Working for pay or profit</p> <p>L2 Unemployed</p> <p>L3 Pupil, student, further training, unpaid work experience</p> <p>L4 In retirement or early retirement or has given up business</p> <p>L5 Permanently disabled</p> <p>L6 In compulsory military or community service</p> <p>L7 Fulfilling domestic tasks</p> <p>L8 Currently not at work due to maternity, parental, sick leave or holidays</p> <p>L9 Other</p>
Professional category	<p>P1 Manager</p> <p>P2 Professional</p> <p>P3 Technician and associate professional</p> <p>P4 Clerical support worker</p> <p>P5 Service and sales worker</p> <p>P6 Skilled agricultural, forestry and fishery worker</p> <p>P7 Craft and related trades worker</p> <p>P8 Plant and machine operators, and assembler</p> <p>P9 Elementary occupation</p> <p>P10 Armed forces occupation</p> <p>P11 Other</p>

Education	<p>E0 Not applicable</p> <p>E1 No formal education or below ISCED</p> <p>E2 Primary education (ISCED 1)</p> <p>E3 Lower secondary education (ISCED 2)</p> <p>E4 Upper secondary education (ISCED 3)</p> <p>E5 Post-secondary but non-tertiary education (ISCED 4)</p> <p>E6 First stage of tertiary education (ISCED 5)</p> <p>E7 Second stage of tertiary education (ISCED 6)</p>
weekday of consumption	<p>W1 Monday</p> <p>W2 Tuesday</p> <p>W3 Wednesday</p> <p>W4 Thursday</p> <p>W5 Friday</p> <p>W6 Saturday</p> <p>W7 Sunday</p> <p>W8 Unclassified</p>
Season	<p>S1 Spring</p> <p>S2 Summer</p> <p>S3 Fall</p> <p>S4 Winter</p> <p>S5 Unclassified</p>

Exceptional diet day	E1 No E2 Yes, unspecified E3 Yes, consumed more than normal E4 Yes, consumed less than normal E5 Unclassified
Meal type	M0 Before breakfast M1 Breakfast M2 Snack between breakfast and lunch M3 Lunch M4 Snack between lunch and dinner M5 Dinner M6 Snack after dinner M7 Unclassified
Place of consumption	P1 At home P2 Out of home P3 Unclassified
Food classification and description	FoodEX2

7 Nutrition and Health Data

‘Nutrition and health’ data includes all data related to the nutritional and health status of individuals and populations, including descriptive characteristics of individuals and populations, biomarkers, genomic data, phenotype data and microbiome data.

Human Nutritional Science studies the effects of food components on metabolism, health, performance, and disease resistance of humans, also encompassing the study of human behaviour related to food choices. Nutritional epidemiology, on the other hand, assesses the relations between diet, nutrients and health, and disease outcomes⁶⁸. Nutrition data are heterogeneous in terms of quality and nature and harmonization of data on nutrition and health is generally limited, even though data are being collected in a vast number of national and international research and monitoring projects. Where there have been initiatives to harmonise data, the guidelines and standards appear to be disseminated within specific scientific communities and definitions and concepts used in different datasets are often from different sources. Even when datasets are shared, there will be a need for retrospective data harmonization to enable re-use of data.

The ENPADASI project⁶⁹ included investigation of structured and standardised data storage, and actions required for clinical data sharing. The project specifically focused on the general rules to share and reuse data based on EU national policies, addressing ethics, data protection, data sharing policies and intellectual property. Within the ENPADASI project, the design of dedicated data sharing infrastructure was planned with related analysis of the data sharing issues. The need for standardization was identified as crucial to allow mapping of data from different sources and the project considered standardization of study metadata and phenotypic data (e.g. clinical data, dietary intake, lifestyle and physical activity, metabolomics, and transcriptomics). A new more user-friendly version of the Phenotype database was developed, and the upload of additional studies resulted in the identification of bugs and new features that were implemented. A technical definition of the database infrastructure that connects the intervention studies and observational studies was discussed. Nutritional terms were identified, based on the templates and uploaded studies, that were mapped to existing ontologies and new ontologies were developed for nutritional terms.

7.1 Biomedical data

Biomedical data is information that relates to human health and therefore includes anthropometric data, phenotype, data on disease characterisation and biomarkers of health. Information related to socio-economic status (e.g. employment status, education) and data related to lifestyle (e.g. physical activity, alcohol consumption, smoking) may also be part of biomedical datasets.

The development of ontologies has played an important role in biomedical data and knowledge representation, integration, sharing and analysis. For example, The Ontology for Biomedical Investigations

⁶⁸ Leaf & Weber, 1987

⁶⁹ <http://www.enpadasi.eu/>

(OBI)⁷⁰ has been used for representation of a wide range of investigations⁷¹. Biomedical ontologies are consensus-based controlled biomedical vocabularies of terms and relations with associated definitions, which are logically formulated to promote automated reasoning⁷². It is essential to reach interoperability between different ontologies and different databases/datasets. There have been a number of other initiatives aimed at harmonising and standardizing data sharing across the biomedical area including the BioSHaRe project⁷³ that was part of the Maelstrom research network⁷⁴ that brings together an international team of epidemiologists, statisticians, and computer scientists to answer some of the challenges of cross-cohort research collaborations. The overall goal is to facilitate collaborative epidemiological research through rigorous data documentation, harmonization, integration and co-analysis. These projects consisted of partnerships between collaborators in related fields but do not appear to have produced comprehensive and overarching standards that would be applicable to FNS-Cloud.

Ontology for Nutritional Studies (ONS)

The ENPADASI project developed the Ontology for Nutritional Studies (ONS) to facilitate the harmonization and integration of biological samples and integrate terms related to food description, medical science, genetics, genomic data and nutritional science methods for diet and health research⁷⁵. A key principle was to avoid the definition of new terms where they were present in other ontologies already in use. Terms to be included in the ONS were collected among partners of the ENPADASI consortium, as well as from templates for data and metadata upload into the DASHIN databases⁷⁶.

To enhance interoperability with other ontologies, the ONS built on a subset of the Ontology for Biomedical Investigations (OBI). Terms related to food description were included by importing a subset of terms from the FOODON ontology, which itself was built on the LanguaL system for food description. Where new terms, that were not available in existing ontologies, were needed they are prefixed with 'ONS' followed by an underscore and a sequential seven-digit number. New terms were annotated with text to define the term. The ONS is hosted in a GitHub repository⁷⁷ and was uploaded to the Bioportal repository⁷⁸.

Central concepts of ONS include:

- Diet - defined as the regular course of eating and drinking adopted by a person or animal (ONS_0000080)
 - Usual diet - defined as the regular course of eating and drinking adopted by a population in a certain geographical area, or in a certain cultural setting, or following certain common eating behaviour

⁷⁰ Bandrowski et al., 2016

⁷¹ Ong and He, 2016

⁷² Cimino & Zhu X, 2006

⁷³ <https://cordis.europa.eu/project/id/261433/reporting/fr>

⁷⁴ <https://www.maelstrom-research.org/>

⁷⁵ Vitali et al., 2018

⁷⁶ <https://dashin.eu/interventionstudies/>

⁷⁷ <https://github.com/FrancescoVit/Ontology-for-Nutritional-Studies>

⁷⁸ <http://bioportal.bioontology.org/ontologies/ONS>

- Prescribed diet - defined as a diet prescribed by a physician/nutritionist to meet specific nutritional needs of a person
- Intervention diet - defined as the diet administered during an intervention study.
- Food component - defined as any substance that is distributed in foodstuffs. It includes materials derived from plants or animals, such as vitamins or minerals, as well as environmental contaminants
 - Nutrient – a food component used by the body for normal physiological functions that guarantee survival and growth.
 - Food bioactive -a food bioactive is a food component other than those needed to meet basic human nutritional needs
 - Contaminant - is unwanted food component that makes the food no longer suitable for use
 - Additive - is a component added to food to improve or preserve it
- Food - defined as a complex matrix that is consumed by a person through the process of eating or drinking
 - Raw food – is an uncooked, unprocessed food that is consumed in its natural state
 - Processed food- is the result of the process of home or industrial food preparation
- Measurement datum - is an information content entity that is a recording of the output of a measurement such as produced by a device
 - Biomarker

ONS is the first systematic effort providing a solid and extensible formal ontology framework for nutritional studies, where integration of new information can be easily achieved by the addition of extra modules (i.e. Nutrigenomics, Metabolomics, Nutrikinetics, Quality appraisal, etc.). Nutritional researchers who might not necessarily be familiar with ontologies and concept standardization, can find in ONS a single knowledge entry point for a unified and standardised terminology for their studies. ONS is intended to be a collaborative development that can respond to the needs of the FNS-Cloud and research community.

ELIXIR

ELIXIR⁷⁹ is an intergovernmental organisation that brings together life science resources from across Europe, including databases, software tools, training materials, cloud storage and supercomputers. The aim of ELIXIR is to coordinate these resources so that they form a single infrastructure that makes it easier for scientists to find and share data, exchange expertise, and agree on best practices. ELIXIR includes an interoperability platform that aims to establish Europe-wide standards that can be used to describe life science data and aspects of the platform may be relevant to FNS-Cloud.

The portfolio of ELIXIR Recommended Interoperability Resources is regularly evaluated for quality assurance and quality control and potentially useful resources include:

- FAIRsharing⁸⁰ – collection of standards, including ontologies, manually curated from a variety of sources, including BioPortal, MIBBI and the Equator Network.

⁷⁹ <https://elixir-europe.org/>

⁸⁰ <https://fairsharing.org/>

- Ontology Look Up Service⁸¹ - repository for biomedical ontologies that provides a single point of access to the latest ontology versions that can be browsed or accessed via an API.

Unified Medical Language System (UMLS)

The UMLS, or Unified Medical Language System⁸², published by the National Library of Medicine (US), is a set of files and software tools that brings together many health and biomedical vocabularies and standards to enable interoperability between computer systems⁶⁴. The UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records. UMLS is intended for several uses including processing texts to extract concepts, relationships or knowledge and to develop information retrieval systems. Licences for use are available to individuals but not to groups or organisations.

The UMLS Metathesaurus is a multi-purpose, and multi-lingual vocabulary database consisting of information about biomedical and health-related concepts, their names, and relationships between them. UMLS includes terms and codes from various vocabularies. The Metathesaurus is organized by concept or meaning and links alternative names and views of the same concept and identifies useful relationships between different concepts. The concepts are assigned one or more semantic type from the UMLS semantic network. The major groupings of semantic types are for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas and are intended for broad use with terminologies used in different domains.

There are a large number of vocabularies included in UMLS and including some that are likely to be relevant to FNS-Cloud:

- Clinical Care Classification (CCC)
- Medical Entities Dictionary (CPM)
- Gene Ontology (GO)
- International Statistical Classification of Diseases and Related Health Conditions, Tenth Revision (ICD10)
- MeSH
- NCI thesaurus
- SNOWMED Clinical Terms

UMLS provides two implementations (REST format and OAP format) of the UMLS API. The REST API provides a set of convenient URI patterns and Json output that offer links for important UMLS entities such as CUIs, atoms, and subsets such as the SNOMED CT-> ICD-10-CM map. The SOAP implementation installs all the classes for the SOAP API using Maven, code samples are available mainly in Java. Details and examples of the API are available⁸³.

SNOWMED CT

SNOWMED CT is a comprehensive, multilingual clinical healthcare terminology that is used worldwide and enables consistent, processable representation of clinical content in electronic health records. Concepts

⁸¹ <https://www.ebi.ac.uk/ols/index>

⁸² Bodenreider, 2004

⁸³ <https://documentation.uts.nlm.nih.gov/>

are organised into hierarchies with descriptions that relate concepts to each other. Within hierarchies, concepts are organised from general through to more detailed, allowing data to be recorded at detailed levels and then aggregated to broader terms.

Top level concepts that are the major branch of the SNOWMED CT hierarchy are:

- Clinical finding - result of a clinical observation, assessment or judgment
- Procedure - activities performed in the provision of health care
- Situation with explicit context - concepts in which the clinical context is specified as part of the definition of the concept, includes presence or absence of a condition, whether finding is current or in the past
- Observable entity - question or assessment which can produce an answer or result
- Body structure - normal and abnormal anatomical structures
- Organism - organisms of significance in human and animal medicine
- Substance - represents general substances, the chemical constituents of pharmaceutical/biological products, body substances, dietary substances and diagnostic substances
- Pharmaceutical / biologic product - drug products
- Specimen - entities that are obtained (usually from the patient) for examination or analysis
- Special concept - concepts that don't play a part in the formal logic of the concept model of the terminology, but which may be useful for specific use cases
- Physical object - natural and man-made physical
- Physical force – physical forces that can play a role as mechanisms of injury
- Event - occurrences excluding procedures and interventions (e.g. flood, earthquake).
- Environments and geographical locations - types of environments as well as named locations such as countries, states and regions
- Social context - social conditions and circumstances significant to health care (e.g., occupation, spiritual or religious belief)
- Staging and scales - represents assessment scales and tumour staging systems
- Qualifier value - the values for some SNOMED CT attributes, where those values are not subtypes of other top-level concepts. (e.g., left, abnormal result, severe).
- Record artefact - content created for the purpose of providing other people with information about record events or states of affairs. (e.g., patient held record, record entry, family history section).
- SNOMED CT Model Component - contains technical metadata supporting the SNOMED CT release.

There are also a range of attributes that are used to define clinical finding concepts and examples that could be relevant to FNS-Cloud include:

- Finding site - specifies the body site affected by a condition
- Severity - used to sub-class a clinical finding concept according to its relative severity
- Finding method - specifies the means by which a clinical finding was determined
- Finding informer - the person (by role) or other entity (e.g., a monitoring device) from which the clinical finding information was obtained.

- Has specimen - specifies the type of specimen on which a measurement or observation is performed.
- Component - refers to what is being observed or measured by a procedure.
- Property - specifies the kind of property being measured.
- Scale type - refers to the scale of the result of an observation of a diagnostic test.
- Measurement method - specifies the method by which a procedure is performed.
- Specimen substance - specifies the type of substance of which a specimen is comprised.
- Specimen source identity - specifies the type of individual, group or physical location from which a specimen is collected

7.1.1 Data Exchange Model

The data model should be based on the ONS ontology design that consists of food, components, measurements, study metadata. Biomedical data should be mapped to concepts and vocabularies used in the UMLS.

7.1.1.1 Food Entity

In ONS, foods are listed as FoodON food classes that are based on LanguaL facet B source codes. A link can be made between ONS foods and other food data types (composition, consumption) using the Food Entity for those data types.

7.1.1.2 Component Entity

ONS contains component classes that can be linked to other data types using the Component Entity. Components included in ONS are very limited, e.g., protein, polyunsaturated fatty acid and definitions of type of component, e.g., nutrient, bioactive compound, contaminant.

7.1.1.3 Study Metadata Entity

ENPADASI produced a checklist for minimal requirements for study metadata⁸⁴.

Descriptors included are:

- Full name of the study (STR)
- Country of the study (STR)
- Description of the study aim within the investigation (STR)
- Principal investigator (name) for the study described (STR)
- Contact information of the contact person of the study/experiment (STR)
- Funding body/bodies for the investigation (STR)
- Upload if available

⁸⁴ Pinart et al., 2018

- or provide the URL Study web link for the investigation or study (URL)
- Registration number of the study (i.e., clinicaltrials.gov)
- IRB/IEC approval number —
- Informed consent —
- Study protocol and any protocol deviation/amendments —
- Questionnaires —
- SOPs for samples collection —
- Publications: type and DOI or file location —
- Other: Please specify type of document (i.e., data dictionaries and research proposals)
- Data-sharing policy: study terminated (THS)
- Data-sharing policy: data (THS)
- Aggregate data-sharing policy (i.e., descriptive statistics) (THS)
- Metadata (THS)
- Data analysis permission (THS)
- Study design (THS)
- Provide a short description of the study (STR)
- Study population (THS)
- Particular dietary, physiologic, or nutritional characteristics of target population (STR)
- Population representativeness (THS)
- Type of sampling (THS)
- Describe control group (STR)
- Describe type of controls (STR)
- Start/end of recruitment DD/MM/YYYY–DD/MM/YYYY (STR)
- Follow-ups - Describe time points and actions taken (STR)
- Total number of participants recruited Total; M; F (STR)
- Age range of the study participants (STR)
- Method for dietary or nutritional assessment Dietary records (THS)
- Reference of the main food-composition table used (or URL) (STR)
- Type of food assessed Food (THS)
- Nutrient and food intake data (THS)
- Physical activity Objective measurement (THS)
- Tobacco use (THS)
- Alcohol consumption (THS)
- Anthropometry Weight (THS)
- Sociodemographic information (THS)
- Study outcomes and time points of assessment: (STR)
- Total number of sample donors (number of individuals with biological samples) (STR)
- Type of biological samples and total number of sample donors per sample type (THS)
- Fasting (THS)
- Relative time points of sampling (STR)
- Type of omics Biomarkers (THS)
- Measurement (i.e., metabolite profiling) (THS)
- Technology (THS)

7.1.1.4 Measurement Type Entity

Measurement types and properties, e.g., tissue type, assay type.

7.1.1.5 Measurement Value Entity

Values for reported measurements and associated data (e.g., study code, sample code, tissue type, sample date, unit, matrix unit, method).

7.1.2 Thesauri

ONS ontology codes, mapped to UMLS codes.

7.2 Genomic data

When it comes to genomic data, FASTQ has emerged as a common file format for storing and sharing biological sequencing data. It was developed at Wellcome Trust Sanger Institute by bundling the FASTA sequence with its quality data. It was an ad hoc solution that became a standard, mostly due to its simplicity.

FASTQ is text-based and normally uses four lines per sequence⁸⁵:

- Line 1 begins with a '@' character and is followed by a sequence identifier and an optional description
- Line 2 is the raw sequence letters
- Line 3 begins with a '+' character and is an optional repeat of the Line 1
- Line 4 quality score for the sequence in Line 2 (must have the same length)

```
@SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGCTTTTTTGTGGAAACCGAAAGG
GTTTGAATTTCAAACCCTTTTCGGTTCCAACCTTCCAA
AGCAATGCCAATA
+SRR014849.1 EIXKN4201CFU84 length=93
3+&$#"7F@71, '";C?,B;?6B;:EA1EA
1EA5'9B:?:#9EAOD@2EA5':>5?:%A;A8A;?9B;D@
/=<?7=9<2A8==
```

Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

FASTQ needs further harmonisation, as there exist at least three incompatible variants:

- Sanger

⁸⁵ <http://maq.sourceforge.net/fastq.shtml>

- Solexa
- Illumina

The three described FASTQ variants, with columns giving the description, format name used in OBF projects, range of ASCII characters permitted in the quality string (in decimal notation), ASCII encoding offset, type of quality score encoded and the possible range of scores

Description, OBF name	ASCII characters		Quality score	
	Range	Offset	Type	Range
Sanger standard				
fastq-sanger	33–126	33	PHRED	0 to 93
Solexa/early Illumina				
fastq-solexa	59–126	64	Solexa	–5 to 62
Illumina 1.3+				
fastq-illumina	64–126	64	PHRED	0 to 62

There have been attempts made to define conversions between these variants⁸⁶.

The Sequence Alignment/Map Format⁸⁷ is similar to FASTQ as it is also a text format, with the header line starting with @ and the sequences being simple ASCII character strings. The difference is that it is TAB-delimited and has 11 mandatory fields for essential alignment information such as mapping position, and variable number of optional fields for flexible or aligner specific information. It is being developed by Samtools⁸⁸ - an organisation for genome sequencing developers that has also developed the VCF Variant Call Format⁸⁹. The variant call format (VCF) is a generic format for storing DNA polymorphism data such as SNPs, insertions, deletions and structural variants, together with rich annotations. VCF data are usually stored in a compressed manner and can be indexed for fast data retrieval of variants from a range of positions on the reference genome. The format was developed for the 1000 Genomes Project and has also been adopted by other projects such as UK10K, dbSNP and the NHLBI Exome Project⁹⁰.

Both the FASTQ and SAM/BAM formats are accepted by the Genomic Data Commons NIH for molecular sequencing data⁹¹, while FASTQ files contain information about the reads, SAM/BAM files also provide information about the mapping to the reference genome.

While FASTQ and SAM/BAM focus more on the raw sequencing data, another interesting aspect of genomic data are variants, that is genetic variations or differences that that make each person's genome unique. When 99.5% of the human genome is the same for all humans, we should focus on these differences for health related determination for individuals. We find these variants by comparing the DNA

⁸⁶ Cock et al., 2010

⁸⁷ <https://samtools.github.io/hts-specs/SAMv1.pdf>

⁸⁸ <https://samtools.github.io/>

⁸⁹ <https://samtools.github.io/hts-specs/VCFv4.2.pdf>

⁹⁰ https://osp.od.nih.gov/wp-content/uploads/Genomic_Data_Standards_Resources_and_Initiatives.pdf

⁹¹ <https://gdc.cancer.gov/about-data/data-standards>

sequence we're analysing to the DNA sequence of a reference genome maintained by the Genome Reference Consortium (GRC).

The Global Alliance for Genomics & Health (GA4GH) is the biggest authority when it comes to developing technical standards for data sharing in relation to genomics⁹², integrating over 500 leading organizations working in healthcare, research, patient advocacy, life science, and information technology including the NCI Genomic Data Commons⁸⁹. They have several Foundational and Technical Workstreams working on or developing different aspects of the standards and policies. They use the real-world Driver Projects to define and develop standards that are most needed for the international genomics community to share data. Adopting Driver Projects of the v1.0 Release Candidate include the VICC⁹³, ClinGen⁹⁴, and the BRCA Exchange⁹⁵.

From all the defined workstreams, the ones dedicated to Cloud⁹⁶ and Genomic Knowledge Standards⁹⁷ seem to be most relevant to the FNS-Cloud project.

The GA4GH Cloud Work Stream is working on implementing a Data Repository Service (DRS) API that embodies a RESTful service philosophy and uses JSON in requests and responses and standard HTTPS for information transport⁹⁸.

While there were several previous projects dedicated to sharing genomic variation data, like dbSNP⁹⁹, ClinVar¹⁰⁰, COSMIC¹⁰¹ and DECIPHER¹⁰², there have still been some inconsistencies and need for harmonisation of data models, in particular the representation of an allele. In 2016 the Variant Modelling Collaboration (VMC) has started working on the VMC Data Model and Specification to standardise the terminology and exchange of allele, haplotype and genotype data. The VMC project was incorporated into GA4GH⁹⁰ and the 1.0 release was approved in September 2019. The GA4GH Variation Representation Specification (as the specification is now called) is under active development and accessible on GitHub (<https://github.com/ga4gh/vr-spec/>) in a machine readable format and on <https://vr-spec.readthedocs.io/en/1.0/>

The GA4GH has also developed an ontology called DUO (Data Use Ontology), that will be further explored as part of the WP3 tasks and deliverables¹⁰³.

⁹² <https://github.com/ga4gh>

⁹³ <https://cancervariants.org/>

⁹⁴ <https://reg.clinicalgenome.org>

⁹⁵ <https://brcaexchange.org>

⁹⁶ https://www.ga4gh.org/work_stream/cloud/

⁹⁷ https://www.ga4gh.org/work_stream/genomic-knowledge-standards/

⁹⁸ https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.0.0/docs/#_version_information

⁹⁹ Sherry et al., 2001

¹⁰⁰ Landrum et al., 2013

¹⁰¹ <https://cancer.sanger.ac.uk/cosmic>

¹⁰² <https://decipher.sanger.ac.uk/>

¹⁰³ <https://github.com/EBISPOT/DUO>

Another initiative working on genomic data standardization is the Genomic Standards Consortium¹⁰⁴, that supports a wide range of projects working on a Genomic Contextual Data Markup Language¹⁰⁵, genomics metadata¹⁰⁶, Environment Ontology¹⁰⁷ or a Genomic Contextual Data JSON format¹⁰⁸ to name a few.

Complete Genomics¹⁰⁹ lists on their website several open-source tools, which might be useful in the FNS-Cloud Project.

7.2.1 FHIR (Fast Healthcare Interoperability Resources)

As the GA4GH focuses on genomics data in order to improve human health, they also take into account the FHIR (Fast Healthcare Interoperability Resources) that is a specification to enable the transfer of healthcare information over standard APIs. FHIR is being developed by HL7, that has been producing healthcare data exchange and information modelling standards for over 20 years and is a successor to HLv2, HLv3 and the RIM, and CDA.

There is a growing need for medical data interoperability and FHIR is trying to be a response to that need, by making electronic health records available, discoverable, and understandable. The goal is also to support automated clinical decision support and other machine-based processing, so the data must also be structured and standardised.

The basic building block in FHIR is a Resource. All exchangeable content is defined as a resource. Resources all share the following set of characteristics¹¹⁰:

- A common way to define and represent them, building them from data types that define common reusable patterns of elements
- A common set of metadata
- A human readable part

Although FHIR focuses on the topic of healthcare in relation to patient data management in hospitals and healthcare facilities as well as the general operations and workflows of such facilities and not directly on the different aspects of food and nutrition security, for data related to patient information and studies conducted on patients FNS-Cloud should align its data format recommendations with the FHIR standard, in order to allow interoperability and exchange of data.

The FHIR Resources of interest for FNS-Cloud are the ones connected to Public Health & Research, that is (<https://www.hl7.org/fhir/resourcelist.html>):

- ResearchStudy
- ResearchSubject

¹⁰⁴ <https://gensc.org/>

¹⁰⁵ <https://gensc.org/projects/gcdml/>

¹⁰⁶ <https://gensc.org/projects/m5/>

¹⁰⁷ <https://gensc.org/projects/the-environment-ontology-envo-project/>

¹⁰⁸ <https://gensc.org/projects/gcdj/>

¹⁰⁹ <https://www.completegenomics.com/public-data/analysis-tools/third-party-tools/>

¹¹⁰ <https://www.hl7.org/fhir/overview.html>

and their connected Resources.

It is important to note, that the FHIR standard development is a work in progress and not many of the Resources have yet reached the normative maturity level (<https://www.hl7.org/fhir/versions.html#maturity>).

FHIR standard uses JSON and XML formats to describe the content of its defined resources and encourages its users to support both formats for data exchange.

FHIR defines the following methods for exchanging data between systems:

- RESTful API
- Messaging
- Documents
- Services
- Database / Persistent Storage

Where RESTful API is the only one being balloted as normative and there are no plans to further develop the Messaging, Documents and Services methods. Database / Persistent Storage is a new area with considerable action at this time, and many production implementations, though RDF itself is not getting much use. At this time, HL7 is monitoring implementer experience and feedback to see whether additional standardization is required.

7.3 Microbiome data

In the area of microbiome study, different types of data are collected and analysed. Taking the DIME (Dietary Bioactives and Microbiome Diversity) study being conducted by QIB as an example, we can name several types of collected data:

- Gut **microbiota**
- Urine **metabolomics**
- **Anthropometric** data (blood pressure, BMI, etc.) – can be common across a lot of studies.
- Blood tests
- Oral GTT (Glucose Tolerance Test) and continuous glucose monitoring
- Dietary assessment (**food intake**). Link to earlier section
- Record of stool form and frequency
- Gastrointestinal Transit Time
- Reported mood and stress scores

The human gut microbiota, that is the bacteria, archaea and viruses that reside in the gastrointestinal tract, are vitally important for the health of their host. The gut is home to trillions of microorganisms (predominantly bacteria) with the richest concentration located in the large intestine or colon. It is estimated that the number of microbiota cells in humans roughly equals the number of human cells.

Thanks to the advances in high throughput DNA sequencing technologies, and the decrease of cost related to such studies, we can identify individual species present in a person's gut and analyse the taxonomic

diversity of the microbiota. Also the functional capacity of the microbiota can now be measured through shotgun whole-genome sequencing technologies which have recently emerged and allow sequencing of the total gut microbial community DNA and the subsequent matching of the resulting sequences to known functional genes.

The analysed data will be basically genomic data of the microbiota, so an already mentioned FASTQ data format can be used.

There already exists a Cloud-based solution for microbial data - the CLIMB project (Cloud Infrastructure for Microbial Bioinformatics¹¹¹). FNS-Cloud should investigate their solutions in order to align its recommendations for these types of data moving forward. Another tool is MG-RAST¹¹², that is an open-source web app for automatic phylogenetic and functional analysis of metagenomes, the Integrated Microbial Genomes and Microbiomes (IMG/M) system¹¹³ that supports the annotation, analysis and distribution of microbial genome and microbiome datasets sequenced at DOE's Joint Genome Institute (JGI), CLARK for sequence classification¹¹⁴ and MGnify¹¹⁵ a free resource for the assembly, analysis, archiving and browsing all types of microbiome derived sequence data.

Other projects worth mentioning are the International Human Microbiome Standards (IHMS) project¹¹⁶, that focused on all key aspects of from human sample identification, collection and processing to DNA sequence generation and analysis, the NIH Human Microbiome Project¹¹⁷ and the Earth Microbiome Project, that provides metadata guides and templates for MIxS packages (that are part of the aforementioned Genomic Standards Consortium)¹¹⁸ with a dedicated package for human gut

(http://press3.mcs.anl.gov/gensc/files/2016/07/MIxShumangut_210514.xls).

7.4 Food-drug interaction data

Interactions of food components with drugs is a growing health concern. They can include negative interactions such as decreasing drug bioavailability or blocking the mechanism of action and positive interactions, such as maximizing the drug effect, reducing side effects or both.

However, food/diet-drug interactions are quite challenging as food consumption is not well documented on patient profiles and when interaction information is available, it is dispersed in the scientific literature, clinical trial reports, package leaflets for medicine and others. In fact, although some efforts have been made to organize this type of information (drugs.com; medscape.com; Memorial Sloan Kettering Cancer Center interaction tool; Rxisk.org; Rxlist.com; Webmd.com among others) these databases contain limited and incomplete annotations with low overlap between them.

¹¹¹ <https://www.climb.ac.uk/>

¹¹² <http://www.mg-rast.org/>

¹¹³ <https://img.jgi.doe.gov/>

¹¹⁴ Ounit et al., 2012

¹¹⁵ <https://github.com/EBI-Metagenomics>

¹¹⁶ <http://www.microbiome-standards.org/#>

¹¹⁷ <https://hmpdacc.org/>

¹¹⁸ <http://www.earthmicrobiome.org/protocols-and-standards/metadata-guide/>

In order to uncover these interactions two approaches will be used in the FNS-Cloud Project: a transcriptomics and a text mining approach.

The Gene Expression Omnibus¹¹⁹ contains many transcriptomic datasets that are publicly available. The goal is to look for datasets in which a cell line or patients have been treated or fed a certain food bioactive and compare those expression profiles with a control. Datasets can be described by:

- Technology
- Platform
- Title
- Summary
- Compound/s used
- GSM number
- Entrez id
- Samples

For each study sample metadata will have to be obtained. These can be described by:

- Compound
- Concentration
- Time point
- Cell line
- Cell type
- Tissue
- GSM number
- Entrez id

Some of this information can be accessed directly through their API, but others will require web scraping or downloading files. All this information can be presented in JSON format.

Log fold changes (logFC) for expression of genes between conditions, and differentially expressed genes (DEG) can be obtained for each compound. One problem is that there are many systems for gene naming and identification. We will use Entrez gene ids to identify the genes. A table for each compound can be obtained with:

- entrez gene ids
- logFC values w.r.t control
- Average expression
- t values
- p values
- adjusted p values
- bayes factor

¹¹⁹ GEO, <https://www.ncbi.nlm.nih.gov/geo/>

Transcriptomic data will come from either microarray (one channel or two) or RNAseq data. For microarray data, raw data in the form of CEL files or data normalized by the researchers can be downloaded and analysed. We will be using the normalized data since it consumes less space, and the lengthy normalization step is already done for us. It contains metadata preceded by a '!', and then a table of normalized values, where columns correspond to samples and rows to probe sets. For RNAseq data, count data is available in most studies. Here, columns correspond to samples, and rows to a given gene. This data will have to be normalized.

There have been initiatives, such as CMAP¹²⁰, which have already analysed the expression profile of many drug compounds and have tools that allow the comparison of their results with ours. CMAP lets us compare our DEGs with their expression profiles, to infer drug-food interactions from transcriptomic data.

Most information online is in the form of unstructured textual data. With human language technologies and text mining techniques we can extract possible drug-food interactions from this textual data. This will entail being able to determine which documents can contain useful information, and inside these documents, extract entities relating to food, entities relating to drug, and their interactions.

This information will come from many sources: drug bank, drugs.com, Medscape.com, Rxisk.org, Nutrichem 2.0, Webmd, PubMed, Arxiv, clinicalTrials, and dailyMed. Other information sources are not discarded and may be used, including the possibility to use information from blogs, forums and news to complement traditional academic sources. This information, in some cases can be accessed through an API, but in other cases will require web crawling techniques.

A challenge will lie in standardizing this information: compound names for studies in GEO do not follow a specific convention, CMAP has its own compound naming system, and no doubt the various sources used in the text mining approach will use many different naming conventions. These names will have to be mapped to the other naming standards used in the project in order to have any consistency across the tools and to make them easier to use.

In the BLUEPRINT project a REST API with JSON data transfer model was created, inspired by the ICGC DCC Data model. The designed model was called EPICO Data model and is freely available at <https://github.com/inab/EPICO-data-model>

Based on the model the Blueprint data analysis portal¹²¹ was developed, to explore the behaviour of genes, pathways or genome regions across the Blueprint datasets with source code for portal¹²² and API¹²³ freely available.

7.4.1 API

Dedicated APIs for the Food-Drug Interactions tool will be defined and developed later in the project.

¹²⁰ <https://cmap.ihmc.us/>

¹²¹ http://blueprint-data.bsc.es/release_2016-08/#/

¹²² <https://github.com/inab/epico-data-analysis-portal>

¹²³ <https://github.com/inab/EPICO-REST-API>

8 Data Inter-operability and Quality

8.1 Inter-operability

One of the key principles of FNS-Cloud is that its services will have the capacity to link with new resources and enable crosstalk amongst them; therefore, access to FNS-Cloud data will be open access, underpinned by FAIR (findable, accessible, interoperable and re-useable) principles¹²⁴.

The principle of data being findable relies on data being assigned a globally unique persistent identifier and data being provided with metadata that allows data to be described, identified and indexed as a searchable resource. A dedicated high-level metadata structure for FNS Cloud resources (both datasets and tools and services) is being developed in D2.2 to be used in the FNS Cloud catalogues, that will have search functionalities, that will enhance the findability of resources. More detailed and area specific metadata models are being developed in D2.3 for the microbiome study and in WP3 for nucleotide data as demonstrators. Both apps will be integrated with the central catalogues and the nucleotide metadata app will provide APIs for searching by these topic-specific metadata.

Accessible data means that data and metadata provided by FNS-Cloud must be retrievable by their identifier using a standardized communications protocol that is free and universally implementable. To make data interoperable and re-usable, the data maps described in this deliverable must be developed to ensure they are accessible and use relevant controlled vocabularies that also follow FAIR principles.

WP2 tasks 2.3 and 2.4. will integrate document and meta(data) repositories and implement authentication and authorisation infrastructure for FNS-Cloud. FairSpace¹²⁵ will be used to provide access to FNS data, add metadata, and store donated datasets. FairSpace will also provide data governance and metadata functions that enhance existing data sources, and track compliance with FAIR principles. FairSpace functionality to publish/ federate metadata for other open data catalogues (e.g. via OIA-PMH¹²⁶, will be used to improve findability of data. These tasks will check that other data repositories used in FNS-Cloud, provide data in suitable formats and allow open access. Services will access data repositories for WP4 Use Cases and WP5 Demonstrators and those use cases will need to use or map data to the data maps finalized within the project.

FNS-Cloud WP8 will develop a sustainable Open Science and Open Innovation governance framework and sustainable business model for FNS-Cloud in accordance with EU requirements and needs of the project and user communities. The proposed structure and resulting documents will guide design of software underpinning FNS-Cloud architecture and integration, specifically governance, operations, and sustainability, and be adapted for daily working methodologies. This deliverable has reviewed data maps and APIs that are relevant and can be used (or adapted for use) within the project and recommendations

¹²⁴ Wilkinson et al. 2016

¹²⁵ <https://thehyve.nl/cases/fair-vre-institut-curie/>

¹²⁶ <https://www.openarchives.org/pmh/>

will need to be aligned with tasks 8.5 and 8.6 of WP8 to ensure that FNS-cloud data maps and APIs meet requirements for the final implementation.

8.2 Data quality

Provision of high-quality data will be vital for the sustainability of FNS-Cloud. It would be expected that data providers will already have implemented a degree of quality control before providing their data to FNS-Cloud, but it is important that there is a system of quality checks that can help to ensure the quality of the process of adding data to FNS-Cloud and ensuring that it is re-usable.

It's important to note, that there can be different levels when we speak about data quality. One would be on the record level, that can be influenced by measurement methods and procedures, the precision of measurement, etc. and the other on the dataset level, where the quality can be influenced by limited metadata and thesauri and data models that do not directly describe the reality causing a loss of data. The FNS Cloud, as an e-infrastructure, should focus mostly on the second aspect, and promote high quality and standards on the dataset level first. Automatic evaluation of quality on the record level can be difficult, especially when the data is kept in an unstructured format. An approach is being explored in D2.2, where human users of the data can rate it, to provide transparency and feedback on the data being published.

Production and management of food composition data has already been subject to development of procedures to ensure quality of published data. EuroFIR developed a quality management framework¹⁴, with a flow chart of the compilation process with standard operating procedures (SOPs) to assure critical steps being the starting point. Recommendations for food description, component identification, value documentation, recipe calculation, quality evaluation of values, guidelines to assess analytical methods, document and data repositories and training opportunities were harmonized as elements of the quality framework and these form the basis of the data maps and controlled vocabularies related to food composition data recommended in this deliverable. This quality management framework has been adopted, at least in part, by most national food composition data compilers and individual components of the framework have continued to be developed and adopted as 'best practice', with training support provided by EuroFIR and INFOODS.

WP4 task 4.4 plans to develop a strategy to assess data quality of food intake and consumer behaviour datasets. Strategies on when and how to merge data will be provided for user communities with respect to data fusion and merging as well as linking to dietary intake data. SOPs will be published in a standardised format to ensure high-quality data collection, comparability, security, and confidentiality of personal data. Data quality strategies developed will be utilised, including training support, to determine the quality of existing and emerging FNS data.

While SOPs can be used to support process management, quality management and checking of data is challenging and is likely to rely on automated checking of metadata and descriptive codes and terms from controlled vocabularies that are used in datasets. Once data maps are finalized, data management systems and data transfer systems can be used to automatically check that datasets comply with the vocabularies and data formats required by the data maps.

9 Conclusion

This deliverable introduces the concept of data maps and APIs for use in FNS-Cloud. Existing datasets and tools that are available to the project were identified by the results of a survey of project partners (Annex 1 and 2). The availability and applicability of existing standards, thesauri and APIs are reviewed for each of the data domains that are relevant to the project. Limitations and gaps in existing data structures are discussed where applicable. For each data area, a recommended data map, for data transfer, is provided along with recommendations for thesauri and APIs.

The data maps for the food area are generally well-established standards and there are existing mappings between thesauri. The recommended data model for the food intake area is based on the data schema for the EFSA EU Menu project but is likely to need to be extended to handle intake datasets from more heterogeneous research projects.

The ENPADASI produced Ontology for Nutritional Studies can be used as a starting point for a data model for the nutrition and health data area but further development and linking to relevant thesauri will be needed. The Unified Medical Language System, that includes SNOWMED CT, brings together many health and biomedical vocabularies and standards that could be used to aggregate data from the biomedical area. After considering API styles the recommendation for FNS-Cloud is to use REST architecture with a JSON data format.

The final recommended data models will be used by WP3 to develop services that support standardization and inter-operability of data from heterogeneous sources.

10 References

1. Bogaardt, M.-J; Geelen, A; Zimmermann, K; Finglas, P; Raats, M; Mikkelsen, B; Poppe, K; van't Veer, P. (2018) Designing a research infrastructure on dietary intake and its determinants. *Nutrition Bulletin*, 43, 301–309.
2. Oxford Learner's Dictionary. https://www.oxfordlearnersdictionaries.com/definition/english/interface_1. Accessed 12/05/2020.
3. Wikipedia. https://en.wikipedia.org/wiki/Application_programming_interface. Accessed 12/05/2020
4. Quisper. <https://quisper.eu/>. Accessed 12/05/2020
5. Quisper. Quisper developer portal. <https://developer.quisper.eu/apis> . Accessed 12/05/2020
6. EuroFIR. EuroFIR thesauri. <http://www.eurofir.org/our-resources/eurofir-thesauri/>. Accessed 12/05/2020
7. Fowler, M. <https://martinfowler.com/articles/richardsonMaturityModel.html>. Accessed 12/05/2020
8. Jaxenter.com. The State of API Integration: SOAP vs. REST, public APIs and more. <https://jaxenter.com/state-of-api-integration-report-136342.html>. Accessed 26/05/2020
9. LanguaL. www.langua.org. Accessed 12/05/2020
10. Ireland, J; Møller, A; (2010). LanguaL Food Description: a Learning Process. *European Journal of Clinical Nutrition* 64, S44-48
11. FoodON. A farm to fork ontology. www.foodon.org. Accessed 12/05/2020
12. EFSA. Data standardisation. <https://www.efsa.europa.eu/en/data/data-standardisation>. Accessed 26/05/2020
13. Greenfield, H; Southgate, D. (2003). *Food Composition Data: Production, Management and Use*. FAO Rome.
14. Westenbrink, S; Roe, M; Oseredczuk, M; Castanheira, I; Finglas, PM; (2016) EuroFIR quality approach for managing food composition data; where are we in 2014? *Food Chemistry* 193 69-74
15. U.S. Department of Agriculture. Food data central API guide. <https://fdc.nal.usda.gov/api-guide.html>. Accessed 12/05/2020
16. Swedish Food Agency. Food data. <https://www.livsmedelsverket.se/om-oss/psidata/livsmedelsdatabasen>. Accessed 12/05/2020
17. Slimani, N; Deharveng, G; Unwin, I; Southgate, DAT; Vignat, J; Skeie, G; et al. The EPIC nutrient database project (ENDB): a first attempt to standardise nutrient databases across the 10 European countries participating in the EPIC study. *Eur J Clin Nutr.* 2007 Sep;61(9):1037-56. PubMed PMID: WOS:000249276900001. English.
18. Slimani, N; Deharveng, G; Unwin, I; Vignat, J; Skeie, G; Salvini, S; et al. Standardisation of an European end-user nutrient database for nutritional epidemiology: what can we learn from the EPIC Nutrient Database (ENDB) project? *Trends Food Sci Tech.* 2007;18(8):407-19. PubMed PMID: WOS:000248632900002. English.

19. Food and Agriculture Organization of the United Nations. International Network of Food Data Systems (INFOODS). www.fao.org/infoods. Accessed 12/05/2020
20. Becker, W; Møller, A; Ireland, J; Roe, M; Unwin, I; Pakkala, H. (2008). Proposal for structure and detail of a EuroFIR standard on food composition data. II. Technical Annex: D1.8.19. Danish Food Information, Roskilde. http://www.eurofir.org/?page_id=12. Accessed 12/05/2020
21. Becker, W. (2010). Towards a CEN standard on food data. *European Journal of Clinical Nutrition*, 64 (Suppl 3), S49–S52.
22. CEN. (2012). European standard. Food data – structure and interchange format. EN16104:2012. <http://www.sis.se/en>. Accessed 12/05/2020
23. Food and Agriculture Organization of the United Nations. International Network of Food Data Systems (INFOODS). Standards and guidelines. <http://www.fao.org/infoods/infoods/standards-guidelines/en/>. Accessed 12/05/2020
24. GS1. <http://www.gs1.org>. Accessed 12/05/2020
25. EFSA (2010). Standard sample description for food and feed. *EFSA Journal*, 8(1), 1457.
26. EuroFIR. Food EXplorer. <http://www.eurofir.org/our-tools/foodexplorer/>. Accessed 12/05/2020
27. Finglas, P; Berry, R; Astley, S. (2014). Assessing and improving the quality of food composition databases for nutrition and health applications in Europe: the contribution of EuroFIR. *Adv Nutr.* 2014 Sep;5(5):608S-614S
28. Open Food Facts. <https://world.openfoodfacts.org/>. Accessed 12/05/2020
29. The Open Food Repo. <https://www.foodrepo.org/>. Accessed 12/05/2020
30. Tesco. Tesco labs. <https://www.tescolabs.com/category/api/>. Accessed 12/05/2020
31. Nielsen Brandbank. <https://www.brandbank.com/>. Accessed 12/05/2020
32. GS1. Standards. <https://www.gs1.org/standards>. Accessed 12/05/2020
33. GS1. GDSN Standards. <https://www.gs1.org/standards/gdsn>. Accessed 12/05/2020
34. EuroFIR. eBasis. <http://www.eurofir.org/our-tools/ebasis/>. Accessed 12/05/2020
35. Plumb, J; Piga, t S; Bompola, F; Cushen, M; Pinchen, H; Norby, E; et al. eBASIS (Bioactive Substances in Food Information Systems) and Bioactive Intakes: Major Updates of the Bioactive Compound Composition and Beneficial Bioeffects Database and the Development of a Probabilistic Model to Assess Intakes in Europe. *Nutrients*. 2017 Apr;9(4). PubMed PMID: WOS:000401355600005. English.
36. EuroFIR. ePlantlibra. <http://www.eurofir.org/our-tools/eplantlibra/>. Accessed 12/05/2020
37. Phenol-Explorer. <http://phenol-explorer.eu/>. Accessed 12/05/2020
38. Pité M., Pinchen H., Castanheira I., Oliveira L., Roe M., Ruprich J., Rehurkova I., Sirot V., Papadopoulos A., Gunnlaugsdóttir H., Reykdal O., Lindtner O., Ritvanen T., Finglas P. (2018). Quality Management Framework for Total Diet Study centres in Europe. *Food Chemistry* 240 405-414.
39. Presser, K; Weber, D2; Norrie, M. FoodCASE: A system to manage food composition, consumption and TDS data. *Food Chem.* 2018 Jan 1;238:166-172. doi: 10.1016/j.foodchem.2016.09.124. Epub 2016 Sep 19
40. European Commission, 2011. Regulation of the European Parliament and of the Council of 25 October 2011 on the provision of food information to consumers, amending Regulations (EC) No 1924/2006 and (EC) No 1925/2006 of the European Parliament and of the Council, and repealing Commission Directive 87/250/EEC, Council Directive 90/496/EEC, Commission Directive

- 1999/10/EC, Directive 2000/13/EC of the European Parliament and of the Council, Commission Directives 2002/67/EC and 2008/5/EC and Commission Regulation (EC) No 608/2004, 1169/2011/EU. In: Official Journal, L 304/18, 22/11/2011.
41. ISO-Food. <http://isofood.eu>. Accessed 12/5/2020
 42. METROFOOD-RI. <https://www.metrofood.eu>. Accessed 12/5/2020
 43. REALMed. <https://realmedproject.weebly.com>. Accessed 12/5/2020
 44. Eftimov, T; Ispirova, G; Potocnik, D; Ogrinc, N; Koroušić Seljak, B. (2019). "Iso-food ontology: A formal representation of the knowledge within the domain of isotopes for food science", Food chemistry, vol. 277, pp. 82-390.
 45. REALMed. Pursuing authenticity and valorization of Mediterranean traditional products. <http://foodtrack.ijs.si/>. Accessed 26/05/2020
 46. Pauli, JN; Newsome, SD; Cook, JD; et al. (2017). Opinion: Why we need a centralized repository for isotopic data. PNAS March 21, 2017 114 (12) 2997-3001; <https://doi.org/10.1073/pnas.1701742114>
 47. Matuszczyk A., Pferdmenges, L, Presser Karl et al., (2020), FoodCASE Specification: Revision of the Initial Data Level, FoodCASE Technical Report
 48. TDS-Exposure. <http://www.tds-exposure.eu/>. Accessed 12/05/2020
 49. Presser K. et al. (2012), Software Requirement Specification of FoodCASE-Risk, TDS-Exposure project, grant no. 289108, Deliverable 6.2
 50. BfR (The German Federal Institute for Risk Assessment). <http://www.bfr-meal-studie.de/en/the-bfr-meal-study.html>. Accessed 23/3/2020
 51. EFSA (European Food Safety Authority) (2014). Guidance on Data Exchange version 2.0. EFSA Journal 2014, 12(12): 3945. 173 pp. doi:10.2903/j.efsa.2014.3945.
 52. EFSA (European Food Safety Authority) (2020), Harmonised terminology for scientific research, published on Zenodo: https://zenodo.org/record/3243215#.XnTM_qhKiUk. Accessed 20/03/2020
 53. EFSA (European Food Safety Authority) (2019). EFSA Catalogue Browser User Guide. EFSA Supporting publications 16 (11). <https://doi.org/10.2903/sp.efsa.2019.EN-1726>.
 54. ESFRI. ESFRI Roadmap. <https://www.esfri.eu/esfri-roadmap>. Accessed 20/03/2020
 55. European Commission, Joint Research Center, Available online at <https://crm.jrc.ec.europa.eu/>. Accessed 20/03/2020
 56. LabMix24. <https://www.labmix24.com/crmsearch/>. Accessed 20/03/2020
 57. ChEBI. <https://www.ebi.ac.uk/training/online/course/chebi-quick-tour/what-chebi>. Accessed 23/03/2020
 58. EFSA (European Food Safety Authority) (2014). Guidance on the EU Menu methodology. EFSA Journal 2014;12(12):3944
 59. Nutritools. <https://www.nutritools.org/>. Accessed 12/05/20
 60. Hooson, J; Hutchinson, J; Warthon-Medina, M; Hancock, N; Greathead, K; Knowles, B; Vargas-Garcia, E; Gibson, L; Bush, L; Margetts, B; Robinson, S; Ness, A; Alwan, N; Wark, P; Roe, M; Finglas, P; Steer, T; Page, P; Johnson, L; Roberts, K; Amoutzopoulos, B; Burley, V; Greenwood, D; Cade, J. on behalf of the DIET@NET consortium (2019). A systematic review of reviews identifying UK validated dietary assessment tools for inclusion on an interactive guided website for researchers: www.nutritools.org. Critical Reviews in Food Science and Nutrition, DOI: 10.1080/10408398.2019.1566207

61. Slimani, N; Ferrari, P; Ocke, M; Welch, A; Boeing, H; van Liere, M; et al. Standardization of the 24-hour diet recall calibration method used in the European Prospective Investigation into Cancer and Nutrition (EPIC): general concepts and preliminary results. *Eur J Clin Nutr.* 2000 Dec;54(12):900-17. PubMed PMID: WOS:000165965600008. English.
62. de Boer, EJ; Slimani, N; van 't Veer, P; Boeing, H; Feinberg, M; Leclercq, C; et al. The European Food Consumption Validation Project: conclusions and recommendations. *Eur J Clin Nutr.* 2011 Jul;65:S102-S7. PubMed PMID: WOS:000292448100013. English.
63. Slimani, N; Casagrande, C; Nicolas, G; Freisling, H; Huybrechts, I; Ocke, MC; et al. The standardised computerised 24-h dietary recall method EPIC-Soft adapted for pan-European dietary monitoring. *Eur J Clin Nutr.* 2011 Jul;65:S5-S15. PubMed PMID: WOS:000292448100002. English.
64. Gavrieli, A; Naska, A; Berry, R; Roe, M; Harvey, L; Finglas, P; Glibetic, M; Gurinovic, M; Trichopoulou, A; 2014. Dietary monitoring tools for risk assessment. EFSA supporting publication 2014:EN-607, 287 pp. Available online: www.efsa.europa.eu/publications. Gurinović, M; Milešević, J; Kadvan, A; Nikolić, M; Zeković, M; Djekić-Ivanković, M; Dupouy, E; Finglas, P; Glibetić, M; Development, features and application of DIET ASSESS & PLAN (DAP) software in supporting public health nutrition research in Central Eastern European Countries (CEEC), *Food Chemistry* 238 (2018) 186–194 <http://dx.doi.org/10.1016/j.foodchem.2016.09.114>
65. Gurinovic, M.; Milesevic, J.; Kadvan, A.; Nikolic, M.; Zekovic, M.; Djekic-Ivankovic, M.; Dupouy, E.; Finglas, P.; Glibetic, M. Development, features and application of Diet Assess & Plan (DAP) software in supporting public health nutrition research in central eastern european countries (CEEC). *Food Chem.* 2018, 238, 186–194
66. Unified Medical Language System. <https://www.nlm.nih.gov/research/umls/index.html>. Accessed 12/05/20
67. SNOMED International. <http://www.snomed.org/>. Accessed 12/05/20
68. Leaf, A; Weber, PC. A new era for science in nutrition. *Am J Clin Nutr.* 1987;
69. European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI). <http://www.enpadasi.eu/>. Accessed 12/05/20
70. Bandrowski, A; Brinkman, R; Brochhausen, M; Brush, MH; Bug, B; Chibucos, MC; Clancy, K; Courtot, M; et al. (2016) The Ontology for Biomedical Investigations. *PLoS One.* 11(4): e0154556. doi: 10.1371/journal.pone.0154556
71. Ong, E; He, Y. Community-based Ontology Development, Annotation and Discussion with MediaWiki extension Ontokiwi and Ontokiwi-based Ontobedia. *AMIA Jt Summits Transl Sci Proc.* 2016;2016:65–74.
72. Cimino, JJ; Zhu, X. (2006). The practical impact of ontologies on biomedical informatics. *Yearb Med Inform.* 2006:124-35.
73. CORDIS. Biobank Standardisation and Harmonisation for Research Excellence in the European Union. <https://cordis.europa.eu/project/id/261433/reporting/fr>. Accessed 12/05/2020
74. Maelstrom. <https://www.maelstrom-research.org/>. Accessed 12/05/2020
75. Vitali, F; Lombardo, R; Rivero, D; Mattivi, F; Franceschi, P; Bordoni, A; Trimigno, A; Capozzi, F; Felici, G; Taglino, F; Miglietta, F; De Cock, N; Lachat, C; De Baets, B De Tré, G; Pinart, M; Nimptsch, K; Pischon, T; Bouwman, J; Cavalieri, D; and the ENPADASI consortium. ONS: an

- ontology for a standardized description of interventions and observational studies in nutrition. *Genes & Nutrition* (2018) 13:12 <https://doi.org/10.1186/s12263-018-0601-y>
76. Phenotype database. <https://dashin.eu/interventionstudies/>. Accessed 12/05/2020
 77. Ontology for Nutritional Studies. <https://github.com/FrancescoVit/Ontology-for-Nutritional-Studies>. Accessed 12/05/2020
 78. Bioportal. Ontology for nutritional studies. <http://bioportal.bioontology.org/ontologies/ONS>. Accessed 12/05/2020
 79. ELIXIR. <https://elixir-europe.org/>. Accessed 26/05/2020
 80. FAIRsharing. <https://fairsharing.org/>. Accessed 12/05/2020
 81. Ontology Lookup Service. <https://www.ebi.ac.uk/ols/index>. Accessed 12/05/2020
 82. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PubMed PMID: 14681409; PubMed Central PMCID: PMC308795.
 83. UMLS API technical documentation. <https://documentation.uts.nlm.nih.gov/>. Accessed 12/05/2020
 84. Pinart M; Nimptsch, K; Bouwman, J; Dragsted, LO; Yang, C; De Cock, N; Lachat, C; Perozzi, G; et al. (2017). Joint Data Analysis in Nutritional Epidemiology: Identification of Observational Studies and Minimal Requirements. *The Journal of Nutrition Methodology and Mathematical Modeling*; First published online February 27, 2018; doi: <https://doi.org/10.1093/jn/nxx037>.
 85. FASTQ Format Specification. <http://maq.sourceforge.net/fastq.shtml>. Accessed 12/05/2020
 86. Cock, PJA; Fields, CJ; Goto, N; Heuer, ML; Rice, PM. (2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res.* 2010 Apr; 38(6): 1767–1771. Published online 2009 Dec 16. doi: 10.1093/nar/gkp1137
 87. Samtools. Sequence Alignment/Map Format Specification. <https://samtools.github.io/hts-specs/SAMv1.pdf>. Accessed 26/05/2020
 88. Samtools organisation and repository. <https://samtools.github.io/>. Accessed 12/05/2020
 89. Samtools. The Variant Call Format (VCF) Version 4.2 Specification. <https://samtools.github.io/hts-specs/VCFv4.2.pdf>. Accessed 26/05/2020
 90. Genomic Data Standards Resources and Initiatives Cited in the Supplemental Information to the Genomic Data Sharing Policy. (https://osp.od.nih.gov/wp-content/uploads/Genomic_Data_Standards_Resources_and_Initiatives.pdf). Accessed 12/05/2020
 91. National Cancer Institute, Genomic Data Commons. Data Standards. <https://gdc.cancer.gov/about-data/data-standards>. Accessed 12/05/2020
 92. Global Alliance for Genomics and Health. <https://github.com/ga4gh>. Accessed 12/05/2020
 93. Variant Interpretation for Cancer Consortium. <https://cancervariants.org/>. Accessed 12/05/2020
 94. Clinical Genome Resource. ClinGen Allele Registry. <https://reg.clinicalgenome.org>. Accessed 12/05/2020
 95. BRCA Exchange. <https://brcaexchange.org>. Accessed 12/05/2020
 96. Global Alliance for Genomics and Health. Cloud Vision Statement. https://www.ga4gh.org/work_stream/cloud/. Accessed 12/05/2020
 97. Global Alliance for Genomics and Health. Genomic Knowledge Vision Statement. https://www.ga4gh.org/work_stream/genomic-knowledge-standards/. Accessed 12/05/2020

98. Global Alliance for Genomics and Health. Data Repository Service. https://ga4gh.github.io/data-repository-service-schemas/preview/release/drs-1.0.0/docs/#_version_information. Accessed 12/05/20
99. Sherry, ST; Ward, M-H; Kholodov, M; Baker, J; Phan, L; Smigielski, EM; Sirotkin, K. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* 2001 Jan 1; 29(1): 308–311. doi: 10.1093/nar/29.1.308
100. Landrum, MJ; Lee, JM; Riley, GR; Jang, W; Rubinstein, WS; Church, DM; Maglott, DR. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014 Jan 1; 42(Database issue): D980–D985. Published online 2013 Nov 14. doi: 10.1093/nar/gkt111
101. Catalogue of Somatic Mutations in Cancer. <https://cancer.sanger.ac.uk/cosmic>. Accessed 12/05/2020
102. DECIPHER. <https://decipher.sanger.ac.uk/>. Accessed 12/05/2020
103. EBISPO/DO. Ontology for consent codes and data use requirements <https://github.com/EBISPO/DO>. Accessed 12/05/2020
104. Genomic Standards Consortium. <https://genc.org/>. Accessed 12/05/2020
105. Genomic Standards Consortium. GCDML <https://genc.org/projects/gcdml/>. Accessed 12/05/2020
106. Genomic Standards Consortium. M5 GSC project description 2012. <https://genc.org/projects/m5/>. Accessed 12/05/2020
107. Genomic Standards Consortium. The Environment Ontology (EnvO) project. <https://genc.org/projects/the-environment-ontology-envo-project/>. Accessed 12/05/2020
108. Genomic Standards Consortium. GCDJ: Genomic Contextual Data JSON. <https://genc.org/projects/gcdj/>. Accessed 12/05/2020
109. Complete Genomics. Compatible Third-Party Tools. <https://www.completegenomics.com/public-data/analysis-tools/third-party-tools/>. Accessed 12/05/2020
110. FHIR Standard for health care data exchange. Release 4. <https://www.hl7.org/fhir/index.html>. Accessed 12/05/2020
111. Cloud Infrastructure for Microbial Bioinformatics (CLIMB). <https://www.climb.ac.uk/>. Accessed 12/05/2020
112. MG-RAST Metagenomics analysis server. <http://www.mg-rast.org/>. Accessed 12/05/2020
113. Joint Genome Institute. Integrated Microbial Genomes and Microbiomes. <https://img.jgi.doe.gov/>. Accessed 12/05/2020
114. Ounit, R; Wanamaker, S; Close, TJ; Lonardi, S. CLARK: fast and accurate classification of metagenomic and genomic sequences using discriminative k-mers. *BMC Genomics.* 2015; 16(1): 236. Published online 2015 Mar 25. doi: 10.1186/s12864-015-1419-2
115. MGnify. <https://github.com/EBI-Metagenomics>. Accessed 12/05/2020
116. International Human Microbiome Standards (IHMS). <http://www.microbiome-standards.org/#>. Accessed 12/05/2020
117. NIH Human microbiome Project. <https://hmpdacc.org/>. Accessed 12/05/2020
118. Earth Microbiome Project. <http://www.earthmicrobiome.org/protocols-and-standards/metadata-guide/>. Accessed 12/05/2020

119. Gene Expression Omnibus. <https://www.ncbi.nlm.nih.gov/geo/>. Accessed 12/05/2020
120. CMAP. <https://cmap.ihmc.us/>. Accessed 12/05/2020
121. INAB. EPICO data model. <https://github.com/inab/EPICO-data-model>. Accessed 12/05/2020
122. Blueprint epigenome data analysis portal. http://blueprint-data.bsc.es/release_2016-08/#!/. Accessed 12/05/2020
123. INAB. EPICO REST API. <https://github.com/inab/EPICO-REST-API>. Accessed 12/05/2020
124. Wilkinson, MD; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3:160018 doi: 10.1038/sdata.2016.18 (2016)
125. FAIR Virtual Research Environment. <https://thehyve.nl/cases/fair-vre-institut-curie/>. Accessed 12/05/2020
126. Open Archives Initiative. Protocol for Metadata Harvesting. <https://www.openarchives.org/pmh/>. Accessed 12/05/2020

11 Appendices

Appendix 1: Summary results of the information collection exercise relating to datasets to be utilised in the FNS-Cloud project. Full details can be accessed [here](#)

	Data set name	Project partner(s)	Task No.	Ownership, availability, access limitations	Data map classification Primary data type, purpose and users	Related tools or services available (that utilise the data)	URL for dataset and/or key references
Agri-Food							
1	EuroFIR Food composition datasets (via FoodExplorer)	EuroFIR + IMDEA Food + other FNS partners PARTNER	T4.5.2, T5.3.3	National datasets are owned and updated by countries, EuroFIR ownership of modified/standardised datasets in FoodExplorer. Available for project.	Nutrient Composition Food composition (Energy, macro- and micronutrients)	Range of EuroFIR tools and services (FoodExplorer, Food Basket)	http://www.eurofir.org/our-tools/foodexplorer/
2	NEVO Dutch food composition database, available online; NEVO-online vs 2019/6.0	RIVM PARTNER	T4.1, T4.2, T4.4, T5.2.1, T5.2.2	Data is owned by Ministry of Public Health and maintained at RIVM. All contacts go via RIVM. 2. Yes, the dataset is available for use in FNS-cloud. 3. No limitations, except that we require a reference to our dataset, when it is used. Available for project.	Nutrient Composition 1. Food composition data 2. Food research, dietary and nutritional advice and education, consumer information, food labeling etc 3. See 2	LEDA database for branded foods; FoodExplorer database, which currently contains NEVO-online 2013/4.0 and will be updated with NEVO-online 2019/6.0 in the next couple of months	https://www.rivm.nl/en/dutch-food-composition-database direct access to searchable data: https://nevo-online.rivm.nl/ The dataset can be downloaded directly from the website.
3	CoFID	QIB (added by IMDEA) PARTNER	T4.5.2, T5.3.3	Public Health England (PHE) is responsible for maintaining up-to-date data on the nutrient content of the UK food supply. Publicly available.	Nutrient Composition Food nutrients	None	https://www.gov.uk/government/publications/composition-of-foods-integrated-dataset-cofid
4	FoodB	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1. Genome Canada, Genome Alberta, and Genome British Columbia. 2. Publicly available.	Nutrient Composition Comprehensive resource on food constituents, chemistry and biology.	None	https://foodb.ca/
5	CNR_ISPAAM_BoMi Prot	CNR EXTERNAL	T4.1, T5.2	Available from BoMiProt at http://bomiprot.org Publicly available.	Nutrient Composition Proteins present in bovine milk	None	http://bomiprot.org

6	Swiss food composition database V6.1, 2019 and previous versions	PMT PARTNER	T4.2	1) BLV 2) yes 3) none Publicly available.	Nutrient Composition the data is about food composition and the target users are food data compilers and the general public	FoodCASE	https://naehwertdaten.ch/en/
7	CNR_ISPAAM_AGEs-containing proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project with permissions.	Composition Database of advanced glycation end-product-containing proteins in different milk commercial products	None	Renzone et al., J Proteomics. 2015 Mar 18;117:12-23
8	CNR_ISPAAM_lactosylated proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project with permissions.	Composition Database of lactosylated proteins in different milk commercial products	None	Arena et al., Proteomics. 2010 Oct;10(19):3414-34; Arena et al., J Proteomics. 2011 Oct 19;74(11):2453-75
9	CNR_ISPAAM_carbonylated proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project with permissions.	Composition Carbonylated proteins in different milk commercial products	None	Milkovska-Stamenova et al., Food Chem. 2017 Aug 15;229:417-424
10	CNR_ISPAAM_milk metabolites_MCDB	CNR EXTERNAL	T4.1, T5.2	Publicly available from Milk MCDB at http://www.mcdb.ca/ Publicly available.	Composition Metabolites present in bovine milk	None	http://www.mcdb.ca/
11	Bioactive data (eBASIS/ePlantLIBRA)	QIB/EuroFIR/Hylobates + IMDEA Food PARTNER	T4.5.2, T5.3.3	eBASIS and eplantlibra are hosted and maintained by EuroFIR. eBASIS is managed by two institutions, the composition data by QIB and bioeffects data by University College Cork, Ireland. ePlantlibra was developed as part of the Plantlibra EU FP7 project. Yes, available for project (normally membership). Available for project.	Bioactive Composition 1) Composition and biological activity databases for bioactive (non-nutrient) compounds in plant foods (eBASIS) and food supplements (eplantlibra) 2) used for a variety of research/industry uses, including estimation of bioactive intakes. 3) Researchers, food manufacturers, students all use the database	The dataset is searchable via an online interface	http://ebasis.eurofir.org/ http://eplantlibra.eurofir.org/
12	Phenol-Explorer	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1. Funded by French government, the Institut National du Cancer (France), Unilever, Danone and Nestlé. Publicly available.	Bioactive Composition Comprehensive database on polyphenol content in foods	None	http://phenol-explorer.eu/downloads

13	Food Waste Data on side streams (FoodWasteExplorer)	QIB/EuroFIR PARTNER	WP3.1	FoodWasteExplorer has been developed as part of the EU project, REFRESH. The dataset is maintained by EuroFIR and is deployed on the EuroFIR cloud server. Publicly available.	Food Waste Stream Composition Food waste composition; valorisation approaches; food waste description; data references	The dataset is searchable via an online interface with a number of different ways to explore the data	https://www.foodwasteexplorer.eu/home
14	LFCT-AUTH/ATR-FTIR spectroscopic dataset	AUTH PARTNER	T4.1, T5.2	Yes, available for FNS Cloud Available for project.	Authenticity 1. virgin olive oil spectroscopic fingerprint 2. targeted and non-targeted metabolomics; 3. researchers on quality and authenticity issues		https://onlinelibrary.wiley.com/doi/full/10.1002/ejlt.201200317
15	CNR_ISPAAM_bovine milk bioactive peptides_MBPDB	CNR EXTERNAL	T4.1, T5.2	Publicly available from Milk Bioactive Peptide Database at http://mbpdb.nws.oregonstate.edu Publicly available.	Bioactive Composition Bioactive peptides present in bovine milk	None	http://mbpdb.nws.oregonstate.edu
16	CNR_ISPAAM_ovine milk bioactive peptides_MBPDB	CNR EXTERNAL	T4.1, T5.2	Publicly available from Milk Bioactive Peptide Database at http://mbpdb.nws.oregonstate.edu Publicly available.	Bioactive Composition Bioactive peptides present in ovine milk	None	http://mbpdb.nws.oregonstate.edu
17	CNR_ISPAAM_caprine milk bioactive peptides_MBPDB	CNR EXTERNAL	T4.1, T5.2	Publicly available from Milk Bioactive Peptide Database at http://mbpdb.nws.oregonstate.edu Publicly available.	Bioactive Composition Bioactive peptides present in caprine milk	None	http://mbpdb.nws.oregonstate.edu
18	CNR_ISPAAM_human milk bioactive peptides_MBPDB	CNR EXTERNAL	T4.1, T5.2	Publicly available from Milk Bioactive Peptide Database at http://mbpdb.nws.oregonstate.edu Publicly available.	Bioactive Composition Bioactive peptides present in human milk	None	http://mbpdb.nws.oregonstate.edu

19	LEDA Dutch branded food database	RIVM PARTNER/EXT ERNAL	T4.2, T5.2.2	<p>1. Data is owned by the Ministry of Public Health and maintained at the Dutch Nutrition Centre in collaboration with RIVM 2. We will select food from the database that can be used in the project for research purposes 3. Data is not publicly available, and it is not allowed to disseminate it within the project, except to those who directly need it for the work within the WPs. It is not allowed to make the data available on the FNS-intranet or elsewhere. This is due to the fact that both RIVM and Netherlands Nutrition Centre signed license agreements including such constraints.</p> <p>Only available to tasks that directly need it.</p>	<p>Branded Foods</p> <p>1. Label data provided directly or indirectly by food producers 2. Main purpose is to use the data for research purposes (monitoring food reformulation, food intake surveys, NEVO database, and some other) and for educational purposes (informing the consumers on healthy foods and presence of allergens) 3. See 2</p>	NEVO-online vs 2019/6.0	https://www.voeding.scentrum.nl/levensmiddelenbank Only available in Dutch; data is not publicly available
20	Slovenian branded food database	Nutris, JSI, Lifyly PARTNER	T4.4	<p>Ownership is shared between Nutris and JSI.</p> <p>Available for project.</p>	<p>Branded Foods</p> <p>Labeled food data present in the Slovenian market.</p>	The dataset is being utilised in several services in Slovenia including Šolski lonec, food database editing tool Bazil.si, mobile application VešKajJeš etc.	N/A
21	Contaminants_simplified_SSD2_BE-FPS_Nickel_FNS	UGent PARTNER	T4.3	<p>Data is owned by Gent University and available for WP4</p> <p>Available for project.</p>	<p>Contaminants</p> <p>1) Nickel occurrence/contamination in different foods available in the European (Belgian) market. 2) The main purpose is to use the data for exposure assessment, TDS and risk assessment, this data can be used for research and educational purposes (informing the consumers especially the allergic individuals to nickel about nickel occurrence/ intake through consuming different food products). 3) see 2.</p>		N/A

22	LFCT-AUTH/virgin olive oil phenol composition dataset	AUTH PARTNER	T4.1, T5.2	Yes, available for FNS Cloud Available for project.	Composition/Quality the data are about total phenol, total hydroxytyrosol and tyrosol contents; 2. regulatory assessments 3. food control authorities; food industry; researchers on quality issues		https://pubs.acs.org/doi/abs/10.1021/jf5005918 , https://www.mdpi.com/1420-3049/24/6/1044 , https://www.mdpi.com/1420-3049/24/13/2429 ,
23	Standard GDSN - Global Data Synchronization Network	GS1 Slovenia PARTNER	T3.1, T4.4, T5.2	GDSN standard owned by GS1 AISBL. The GS1 standards are available (to the fullest extent possible) on a royalty-free basis and free to use by anyone in and out of the project. To implement the standard a party may need to use a GS1 company prefix and GS1 identification keys. The GS1 company prefix and the GS1 identification keys are licensed from GS1 Member Organisations and subject to licensing fees and / or membership fees. Publicly available but may require licensed keys to implement fully.	Quality Master data including label data are provided directly by food producers, suppliers and distributors in standard GDSN format. Main purpose is to synchronise these data between suppliers and retailers for doing business through the whole supply chain. Data sets also covered all requirements of EU regulation 1169/2011 on the provision of food information to consumers.	For now NONE	www.gs1.org/gdsn
24	Isotopic data	FEM PARTNER	T4.1, T5.2	To be discussed.	Authenticity 1. isotopic data of olive oils, fishes, dairy products 2. Traceability purposes, check of authenticity 3. researchers on quality and authenticity issues, regulatory bodies	isoscapes (to be implemented in FNS Cloud)	N/A

25	CNR_IBBA_TBP-based dataset	CNR PARTNER	T4.1, T5.2	1. IBBA-CNR owns the dataset; 2. Yes is available for use in the project; 3. A formal communication is required Available for project with permission.	Authenticity 1. Food authentication, composition, traceability; 2. Service to food companies, at the moment; 3. Analytical labs interested in the use of the TBP method	We provide a company service for authentication of several raw material and food matrices by what we call the Foodcode platform. Data are also provided as a DNABarcode and translated into a QRcode for convenient reading by a smart device	Morello et al., Genes 2019, Mar 18 10(3) Giani et al., Food Control 2020, 110 article 107010. Braglia et al., American Journal of Plant Sciences 2017, doi 10.4236/ajps.2017.813234 . Braglia et al, J AOACInt 2018, Jan 1;101(1):227-234. Braglia et al, Anal Bioanal Chem 2016, Nov;408(29):8299-8316
26	CNR_ISPA_Salmon mass spectrometric dataset	CNR PARTNER	T4.1, T5.2	Dataset stored in EC repositories. Data obtained for the FP7 - Large scale integrated project FP7-KBBE-2013-7-single stage - Grant agreement number: 613688 Completed December 2018; It needs permission of all participating partners for dataset sharing of other food matrices Available for project with permissions.	Authenticity Food analysis, Fingerprint analysis	All data are also available in Excel files	Internal repository at CNR-ISPA of dataset of salmon samples; Fiorino et al., Food Res Intern 116 (2019) 1258–1265
27	CNR_ISPAAM_milk_species_peptides	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for milk speciation	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
28	CNR_ISPAAM_milk_species_proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for milk speciation	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
29	CNR_ISPAAM_milk_species_adulteration_peptides	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for milk species adulteration identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
30	CNR_ISPAAM_milk_species_adulteration_proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for milk species adulteration identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71

31	CNR_ISPAAM_bovine_milk_thermal_treatment_peptides	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for bovine milk thermal treatment identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
32	CNR_ISPAAM_bovine_milk_thermal_treatment_proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for bovine milk thermal treatment identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
33	CNR_ISPAAM_bovine_milk_thermal_treatment_adulteration_peptides	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for bovine milk thermal treatment adulteration identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
34	CNR_ISPAAM_bovine_milk_thermal_treatment_adulteration_proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for bovine milk thermal treatment adulteration identification	None	Sassi et al., J Agric Food Chem. 2015 Jul 15;63(27):6157-71
35	CNR_ISPAAM_buffalo_milk_freezing_overtime_proteins	CNR PARTNER	T4.1, T5.2	Yes, available for FNS Cloud; external users with permissions Available for project.	Authenticity Food fingerprinting for buffalo milk freezing identification	None	Arena et al. J Proteomics. 2016 Sep 16;147:56-65
36	LFCT-AUTH/quantum chemically calculated values of indices characterizing the radical scavenging activity of virgin olive oil phenols dataset	AUTH PARTNER	T4.1, T5.2	Yes, available for FNS Cloud Available for project.	Authenticity data related to radical scavenging activity (property), prioritization, Researchers		https://pubs.acs.org/doi/10.1021/jf048776x , https://www.sciencedirect.com/science/article/pii/S0963996915300776?via%3Dihub

37	Chemical dissipation half-lives in food crops and other plants	DTU PARTNER	T3.1	Published open access; full database available for FNS Cloud upon request from the main contact person Partially publicly available, full version available with permission.	Risk Assessment Tabular database containing dissipation half-lives of various chemicals (mostly agricultural pesticides) in food crops and other plants for use in human and ecological exposure and risk assessment	Data are stored in an Excel workbook	Key reference: Fantke, P., Juraske, R., 2013. Variability of pesticide dissipation half-lives in plants. Environmental Science and Technology 47, 3548-3562. http://doi.org/10.1021/es303525x The database can also be requested via http://dynamicrop.org/contact.php
Food intake & lifestyle							
38	EFSA Comprehensive European Food Consumption Database	Hylø EXTERNAL	T4.3	Publicly available	Consumption food consumption, Foodex2 categories		https://www.efsa.europa.eu/en/food-consumption/comprehensive-database
39	Belgian food consumption data 2014-2015	UGent EXTERNAL	T4.3	Data is owned by Sciensano, Brussels. Gent University bought this data and Used it for TDS study but UGent is not allowed to share this data with the third party. However, the data are available in the EFSA database now. Publicly available via EFSA.	Consumption 1) Food consumption data. 2) Assessment of food consumption, dietary habits and nutritional quality of the diet among adults and children in Belgium (3-64 years). Assessment of physical activity levels and sedentariness in Belgium among adults and children (3-64 years). Assessment of the adequacy of food and nutrient intakes and physical activity in the different subgroups of the population compared to the recommendations. Estimation of exposure of Belgian children and adults to contaminants, additives and other chemicals in food. Estimation of the effect of nutrition policies (e.g. Belgian National Food and Health Plan (2005-2010), salt reduction, fortification of foods with nutrients, ...) on food and nutrient intakes of the population. 3) see2.		http://www.efsa.europa.eu/en/microstrategy/food-consumption-survey

40	Dietary intake of people aged 65+	TUM, UoR PARTNER	T4.4/T5.3	Yes, available for FNS Cloud Available for project.	Consumption dietary intake data and anthropometric data	None	
41	eNutri (Quisperm 2019 EatwellUK2 study)	UoR PARTNER	T4.4/T5.3	EIT Food funded - Completed Dec 2019. Owned by UoR; Available for use; non-commercial purposes Available for project.	Consumption Food intake, portion size, frequency, nutrient intake, anthropometric, physical activity, demographic data, device (e.g. tablet, mobile), screen size, usability scores	eNutri app (Field Lab 2)	NA
42	ScARES - Seafood Study	UCD PARTNER	T5.3	Data collected as part of Dept of Agriculture Food and Marine (Ireland) funded project. Data is held within UCD. Available for project.	Consumption/Food Choice Dietary intake, Demographic and Food Choice data	None	http://www.ucd.ie/seafirstudy/
43	The Serbian National Food Consumption Survey	CAP PARTNER	T4.4/T5.3	Yes, available for FNS Cloud; external users with permissions Available for project.	Consumption/Socio-economic Dietary intake, anthropometric, socio-demographic, and physical activity data	DIET ASSESS & PLAN (DAP), innovative nutritional software tool for standardised and harmonised food consumption collection and comprehensive dietary intake assessment	
44	Folate intake among Serbian women of reproductive age	CAP PARTNER	T4.4/T5.3	Yes, available for FNS Cloud; external users with permissions Available for project.	Consumption/Socio-economic Dietary intake, anthropometric and socio-demographic data	DIET ASSESS & PLAN (DAP), innovative nutritional software tool for standardised and harmonised food consumption collection and comprehensive dietary intake assessment	
45	Tomappo garden plans dataset	Lifely EXTERNAL	T4.4.3	Owned by Tomappo, available for the needs of T4.4.3 Available for the specific project task.	Food choice & habits Data about vegetable gardeners in Europe (mostly Slovenia and Italy). What vegetables are planted in what amounts and when. The data is used internally by the Tomappo gardening assistant.	Tomappo	N/A
46	ESRC Cognitive Food Choice	UoR PARTNER	T5.3	Data collected as part of ESRC funded project. Data are held within UoR. Data are available for project. Available for project.	Food choice/socio-economic Food expenditure and cognition data. Main purpose is to assess household food spending behaviour, role of cognition in food purchasing		
Food intake & lifestyle / Nutrition & health							

47	Food4me	UCD, UoR, TUM, HUA PARTNER, EXTERNAL	T4.2/5.4.2	FP7 - EU funded - Grant agreement ID: 265494. Completed March 2015 Project use needs permission of all participating partners from food4me (Newcastle Uni, Uni of Navarra, INSTYTUT ZYWNOSCI I ZYWIENIA Poland) Available with permissions.	Consumption/Biomarker/Genomic Multiple .csv files for differing datasets - food intake, nutrient intake, dietary supplement intake, demographic data, genetic data, biochemistry (hosted by Uni of Newcastle, Dietary change data)	None	N/A
48	food and associated glucose data	QIB PARTNER	T4.1, T5.4.1	Yes, available for FNS Cloud; external users with permissions Generated in project.	Consumption/Biomarker dietary assessment and glucose levels after Nutritics and CGM integration	tools to be developed in WP3	
49	Feel4Diabetes study_High risk adults_baseline_GR	HUA PARTNER, EXTERNAL	T4.2/5.4.2	1. HUA (only the data for Greece) 2. Yes 3. In case we want to use the full dataset of this study (that contains data collected in five more European countries), then we should ask for the permission of the relevant Pis from the other countries beyond Greece Full dataset available with permissions.	Consumption/Biomarker 1. Dietary intake, physical activity, blood data, anthropometric and demographic data 2. Data collected in the Feel4Diabetes study http://feel4diabetes-study.eu/ 3. researchers	N/A	N/A
Nutrition & health							
50	metabolomics data	QIB PARTNER	T4.1, T5.4.1	Yes, available for FNS Cloud; external users with permissions Generated in project.	Biomarker urine and faecal metabolomics data	tools to be developed in WP3	
51	WikiPathways	UM PARTNER	T4.4, 4.5, 5.4	The set of biological pathways can be used by everyone Publicly available.	Biomarker 1. Biological processes 2. Functional analysis 3. Researchers and clinician	The data collection of WikiPathways is used in many tools, among which is PathVisio	www.wikipathways.org DOI: 10.1093/nar/gkx1064
52	Gut microbiome dataset	QIB PARTNER	T4.1, T5.4.1	Yes, available for FNS Cloud; external users with permissions Generated in project.	Microbiome Microbial metagenomic sequences	CLIMB (to be implemented in FNS Cloud)	
53	European public assessment reports (EPAR)	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1. European Medicine Agency. 2. Publicly available.	Food-drug interactions UE List of approved drugs for human use.	None	https://www.ema.europa.eu/en/medicines/download-medicine-data

54	Drug Bank	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1. Genome Canada, Genome Alberta, and Genome British Columbia. 2. Creating a free account is mandatory to download the complete database. Publicly available.	Food-drug interactions Comprehensive, freely accessible, online database containing information on drugs and drug targets. Freely accessible but approval for registering is required.	None	https://www.drugbank.ca/releases/latest
55	Drugs.com	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) Drugsite trust 2) dataset is publicly available 3) No limitations or special permissions required to access the data Publicly available.	Food-drug interactions 1) Various information about drugs and medications, including interactions 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://www.drugs.com/
56	Medscape.com	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) WebMD LLC 2) Yes 3) Free registration is required Publicly available.	Food-drug interactions 1) Contains clinical information for use by physicians and healthcare professionals, including drug interactions 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://www.medscape.com/
57	Rxisk.org	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) Data Based Medicine Americas 2) Yes 3) None Publicly available.	Food-drug interactions 1) Drug side effect and interaction data. 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://rxisk.org/
58	NutriChem 2.0	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) University of Hong Kong 2) Yes 3) None Publicly available.	Food-drug interactions 1) Contains Food-Disease associations and food-drug interactions data 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	http://147.8.185.62/services/NutriChem-2.0/
59	Webmd	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) WebMD LLC 2) Yes 3) Certain services require registration Publicly available.	Food-drug interactions 1) Health news and drug information 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://www.webmd.com/

60	Pubmed	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) United States National Library of Medicine 2) Yes 3) None Publicly available.	Food-drug interactions 1) Database of scientific paper's references and abstracts 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://www.ncbi.nlm.nih.gov/pubmed/
61	Arxiv	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) Cornell University 2) Yes 3) No Publicly available.	Food-drug interactions 1) Open access archive for scholarly archives 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://arxiv.org/
62	clinicalTrials	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) U.S. National Institutes of Health, Department of Health and Human Services 2) Yes 3) None Publicly available.	Food-drug interactions 1) Clinical studies 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://clinicaltrials.gov/
63	dailyMed	IMDEA Food Institute EXTERNAL	T4.5.2, T5.3.3	1) United States National Library of Medicine 2) Yes 3) None Publicly available.	Food-drug interactions 1) Information about marketed drugs in the United States 2) Main purpose is to extract information about drug nutrient interactions 3) Physicians/healthcare professionals or patients	None	https://dailymed.nlm.nih.gov/dailymed/

Appendix 2: Summary results of the information collection exercise relating to tools to be utilised in the FNS-Cloud project. Full details can be accessed [here](#)

	Tool name	Project partner(s)	Task No.	Tool description	Data map classification Primary data type, purpose and users	URL for tool and/or key references
1	FoodCASE	Premotec GmbH PARTNER	T4.2	FoodCASE is a data management system to manage food composition, food consumption, TDS and branded food data. In the food composition module it is possible to calculate and aggregation nutrient values and to perform recipe calculations.	Food Classification & Description, Composition, Labeling, Food Safety (Total Diet Study)	https://www.foodcase.org/
2	Food composition web app (single or multiple sources)	Premotec GmbH PARTNER	T2.5	The web app offers search and compare functionalities for food composition data. The web apps retrieves data from a single data management system and caches it or retrieves data several data management systems in real time.	Food Classification & Description, Composition	https://naehrwertdate.n.ch/en/
3	FoodCASE Research	Premotec GmbH PARTNER	T2.5	A light version of FoodCASE only containing the food composition module and without the initial data level.	Food Classification & Description, Composition	N/A
4	Search engine	Premotec GmbH PARTNER	T2.2	A search engine to give an overview over available datasets in the food science data. It will also be experimented if the search engine can access data from these datasets.	Food Classification & Description, Composition	N/A
5	Food shopping advisor	Premotec GmbH PARTNER	T4.3	A web or mobile app for consumers to help making food product choices based on the TDS data in Germany.	Food Safety (Total Diet Study)	N/A
6	Data visualisation app	Premotec GmbH PARTNER	T3.6	A tool will be developed that is able to visualise different food data in different forms. Depending on the food data and goal of the visualisation, a different graph will be used.	TBD	N/A

7	FoodExplorer	EuroFIR PARTNER	T2.5	Allows users to search information from most European Union (EU) Member States (MS) and an increasing number of countries outside Europe, simultaneously. FoodExplorer includes options to search for foods by name, food groups, and the most common LanguaL food descriptors. FoodEXplorer has the unique ability to compare component values amongst foods from different countries' datasets. Results can be downloaded as a food data transport package (FDTP) or in Excel.	Food Classification & Description, Composition	http://www.eurofir.org/our-tools/foodexplorer/
8	eBASIS	EuroFIR/QIB PARTNER		eBASIS (Bioactive Substances in Food Information Systems) is an Internet deployed food composition and biological effects database for plant-based bioactive compounds with putative health benefits. Over 300 major European plant foods are listed and information on 17 compound classes is provided covering multiple bioactive compound classes and plant foods, with data sourced from peer-reviewed literature. Search outputs can be downloaded as spreadsheets, allowing the user to perform calculations, create graphs and manage the data as required. The database represents a unique comprehensive resource on bioactive compounds for researchers, health professionals, health educators, the food industry and policy makers.	Composition (Bioactive components)	http://www.eurofir.org/our-tools/ebasis/
9	ePlantLIBRA	EuroFIR/QIB PARTNER		ePlantLIBRA is a database containing information about plant- and plant-food supplements specifically bioactive compounds in botanicals and herbal extracts with putative health benefits and adverse effects. ePlantLIBRA is based on three existing databases: eBASIS (Bioactive Substances in Food Information System), developed by EuroFIR; the MoniQA contaminants database (FP6 Monitoring and Quality Assurance in the total food supply chain); and FERA's HorizonScan database.	Composition (Bioactive components, Contaminants)	http://www.eurofir.org/our-tools/ebasis/

10	Food Basket	EuroFIR PARTNER		FoodBasket supports all users, especially dieticians, with the calculation of composite and prepared foods. It runs on mobile devices (e.g. smart phones and tablets) as well as desktop computers, and the user-friendly, multi-lingual interface enables users to select any food composition dataset linked with FoodExplorer. The results can be exported as text as well as XML-files. Recipes calculated in advance can also be used as ingredients in new recipes using FoodBasket, which fully integrated with the EuroFIR Interchange Platform resources.	Composition	http://www.eurofir.org/our-tools/foodbasket/
11	FoodWasteExplorer	EuroFIR/QIB/JSI PARTNER		A database of data for food waste streams allowing users to identify potential valorisation opportunities. The web interface tool allows users to find and explore data in a number of ways.	Food Waste	https://ws.eurofir.org/foodwasteexplorer/home
12	PathVisio	UM PARTNER		Open-source and free pathway desing and analysis tool to functionally analyse omics data	Nutrition and Health Data	www.pathvisio.org
13	WikiPathways app	UM PARTNER		Open-source app to visualize and analyse biological processes from WikiPathways in the network analysis tool Cytoscape	Nutrition and Health Data	http://apps.cytoscape.org/apps/wikipathways
14	CyTargetLinker app	UM PARTNER		Open-source app to extend a biological network in Cytoscape with regulatory (protein-protein interactions, microRNA-target interaction, TF-gene interactions or drug-related information).	Nutrition and Health Data	https://cytargetlinker.github.io/
15	BridgeDb	UM PARTNER		Identifier mapping service for genes, proteins, metabolites and biological reactions	Nutrition and Health Data	https://bridgedb.github.io/

16	Foodbook24	UCD PARTNER		An online dietary assessment tool consisting of a 24-hour dietary recall and food frequency questionnaire alongside supplementary questionnaires	Food Intake and Lifestyle, Composition	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5445234/
17	TBP, Tubulin-Based-Polymorphism	CNR PARTNER		Dataset of multiple DNA polymorphisms found in different animal and plant raw material and derived products		N/A
18	eNutri65+ FFQ	UoR PARTNER		An online dietary assessment tool consisting of a food frequency questionnaire alongside supplementary questionnaires. TUM (Partner) is developing a German equivalent.	Food Intake and Lifestyle	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6447217/pdf/pone.0214931.pdf https://www.cambridge.org/core/services/aop-cambridge-core/content/view/64598CD4B17D603EEF901379AC9DA18E/S0029665118002707a.pdf/strategies_for_online_personalised_nutrition_advice_employed_in_the_development_of_the_e_nutri_web_app.pdf
19	Libro	Nutritics PARTNER		Smartphone app for consumers to track nutrition, exercise and sleep and other data points. Connects to Nutritics Professional platform for review and feedback.	Food Intake and Lifestyle	https://www.nutritics.com/p/userguide&c=libro&unlock=1
20	Nutritics Professional	Nutritics PARTNER		Software intended for use by nutritional professionals to perform database nutritional analysis of food diaries, recipes, menus, meal plans, Nutrition label panels, and provide feedback through digital channels and Libro app.	Food Intake and Lifestyle	www.nutritics.com/app

21	Tool for Type 2 Diabetes (T2D) and Hypertension (HTN) risk assessment and personalised dietary and lifestyle feedback	HUA PARTNER		Following WP4, a web-based app, allowing communities to deploy the T2D and HTN tools for performing risk calculations and receiving personalised recommendations will be developed in WP5 and form part of DEM03.	Food Intake and Lifestyle	N/A
22	GEPIR	GS1 Slovenija PARTNER		N/A	Food Labeling	https://gepir.gs1.org/
23	GPC Browser	GS1 Slovenija PARTNER		N/A	Food Labeling	https://www.gs1.org/services/gpc-browser
24	GS1 GDSN Atrify GS1 Slovenija Data Pool	GS1 Slovenija PARTNER		GDSN - Global Data Synchronization Network - is a globally operating standardised network that enables exchange of master product data through one single point in B2B between brand owners, manufacturers, suppliers, distributors and their retailers. In Slovenia our members do synchronization via using Data Pool: Atrify (DataSyncEngine v2), https://www.atrify.com/ Certified data pool list: https://www.gs1.org/services/gdsn/certified-data-pools-list	Food Labeling	https://pool.atrify.com/publishing https://global-catalog.atrify.com
25	GS1 Slovenia GTIN Registry	GS1 Slovenija PARTNER		It is a registry of GTINs® (Global Trade Item Numbers®) for Slovenian brand owners, members of GS1 Slovenia for issuing and capture 7 basic attributes of a product: GTIN, Product Description, Language of description, Brand Name, Target Market / Country of Sale, Global Product Classification (GPC), Product Image URL. GS1 Slovenia GTIN Registry is part of the Global GS1 GTIN Registry Platform. In development it is the service Verified by GS1 to check a product data by querying the GS1 Registry Platform and/or GS1 Slovenia GTIN Registry.	Food Labeling	https://www.gs1si.org/Prijava

26	ONS	UNIFI PARTNER		ONS is an ontology aimed at standardizing terms and formalizing knowledge in the nutritional field	Nutrition and Health Data	https://github.com/enp-adasi/Ontology-for-Nutritional-Studies
27	CLIMB	QIB PARTNER		cyber-infrastructure for microbial bioinformatics, providing cloud-based compute, storage, and analysis tools	Microbiome data	N/A
28	MATAFILER	QIB PARTNER		bioinformatics pipeline for primary analysis of metagenomics data (taxonomic and functional analysis)	Microbiome data	https://github.com/hildebra/MATAFILER (version 1); locally stored in QIB for updated version
29	MGnify	QIB PARTNER		resource for archiving and analyses of microbiome data. We probably want workflow access, not a physical implementation of the resource	Microbiome data	https://www.ebi.ac.uk/metagenomics/about
30	R/Rstudio/Bioconductor	QIB PARTNER		R is a free software environment for statistical computing and graphics	Microbiome data (for the purpose of the project, but can be used for any data)	https://www.r-project.org/ https://www.bioconductor.org/ https://rstudio.com/
31	Irida	QIB PARTNER		IRIDA provides a fully featured system for the storage, management, sharing, and analysis of sequencing data and its associated metadata. Sequence data can be imported directly from Illumina MiSeq and NextSeq instruments. Sequencers into IRIDA's data storage and management system. The sequence data is organized into projects and access to the data can be shared with project collaborators. Data can also be shared with other IRIDA instances across the internet. Data can be analyzed directly with IRIDA, exported to file or to a companion Galaxy workflow system.	Genomic data	N/A

32	Fairspace	HYVE PARTNER		Fairspace provides a transparent layer on top of data sources to enable finding data, managing FAIR data governance processes as well as storing data access activity for regulatory compliance purposes. At the heart of Fairspace transparent layer is a semantic metadata store that stores metadata information about all linked data sources in the workspace, as well as context information such as metadata on subjects, samples, data consents etc.	Any data can be stored	https://thehyve.nl/cases/fair-vre-institut-curie/
33	Quisper Server Platform	Quisper ASBL 3rd PARTY (To EuroFIR)		Aims to deliver a digital platform (Quisper®) as support for creating and delivering personalised nutrition services in Europe. Connect individuals and companies to deliver and share personalised nutrition services. Connect individuals and companies to create and improve their personalised nutrition offerings through access to standards, scientific validated data and knowledge rules.	Food Intake and Lifestyle	https://quisper.eu/connecting-to-quisper/