FNS - Cloud
Food Nutrition Security

Food Nutrition Security Cloud

Deliverable 3.1
Data requirements and applicability criteria

Due Date:	31.05.2020
Submission Date:	04.08.2020
Revision Date:	13.08.2021
Dissemination Level:	Public
Lead beneficiary:	DTU
Main contact:	Peter Fantke (pefan@dtu.dk)

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	*Deliverable 3.1 Data requirements and applicability criteria*		
Authors	*Yasmine Emara (DTU), Peter Fantke (DTU)*		
Reviewer(s)	*Eileen Gibney (UCD), Karl Presser (PMT)*		
Dissemination Level	□ CO Confidential ☒ PU Public		
Approved by	☒ RTDS (COO) ☒ QIB (SCO) ☒ JSI ☒ UCD ☒ PMT ☒ JDLC ☒ EuroFIR ☒ UWTSD ☒ DTU ☒ ENEA ☒ HYVE ☒ HYLO	☒ UM ☒ NUTRIS ☒ RIVM ☒ WUR ☒ UGent ☒ IMDEA ☒ HUA ☒ TUM ☒ GS1 ☒ SF ☒ UoR ☒ IFA	☒ ILSI ☒ BfR ☒ AUTH ☒ FEM ☒ CNR ☒ APRE ☒ CAP ☒ UNIFI ☒ LIFE ☒ Nutritics ☒ EFF
IPRs underlined	n/a (no confidential or licensed data are included in the present deliverable)		
Datasets underlined	n/a (no actual datasets are included in the present deliverable)		

Version/Date	*Change/Comment*
V0.1 2020-04-15	*Draft outline prepared by DTU*
V0.3 2020-04-17	*Final outline based on comments from JSI (WP3 lead)*
V0.4 2020-05-18	*Draft deliverable (version 1) prepared by DTU*
V0.5 2020-06-05	*Draft deliverable (version 2) prepared by DTU*
V0.6 2020-06-30	*Final deliverable version prepared by DTU*
V1.0 2020-07-30	*Final deliverable submitted by DTU incl. all reviewer comments addressed*
V2.0 2021-13-08	*Updated deliverable as per experts request for revision*

Table of Contents

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

List of Tables

List of Figures

Publishable Summary

To answer questions from the field of food and nutrition security (FNS), including food quality, sustainability, and the link between diet and health, various types of data must be gathered, consistently integrated, co-analysed and interpreted. This includes, for example, food composition data, food authenticity and traceability data, food consumption data and health biomarker data. Considering its multidisciplinary nature, the FNS field produces data that can significantly differ in the way the data are generated, structured, formatted, and in the way (same) concepts are expressed, which makes the integration and joint use of such data by humans as well as its incorporation into FNS cloud information platforms challenging.

The FNS-Cloud aims to provide a platform for collecting, sharing, structuring, harmonizing, matching and linking data within and across FNS-related research areas to support integrative data analysis and interpretation. As a starting point, a set of defined data requirements (i.e. what type of data and what variables are necessary to answer a given research question) and applicability criteria (i.e. when is data suitable for harmonization and integration) is needed.

The goal of Task 3.1 (T3.1) is to provide an operational workflow for the development of data requirements and applicability criteria, including an initial set of criteria. Such workflow and initial criteria will help to develop information technology tools (e.g. machine learning algorithms) that will facilitate (semi-)automated data harmonization and integration and enable data interoperability within the FNS-Cloud data platform.

As a first step, an 'FNS data map' is created to delineate the FNS field and define FNS-related 'research question spaces' (i.e. broad research interests or purposes) as well as FNS-relevant data domains and the different types of data they entail. Three data domains were found to capture knowledge and research foci of the FNS field: The 'Agrifood' domain covers all data related to food composition, quality, safety, production processes (including farming activities) and sustainability (e.g. food composition data, branded food data, agronomic performance data). The 'Intake of food and lifestyle' domain covers all data generated on the consumers, their food/lifestyle related behaviours and their nutrient/contaminant intake (e.g. consumption data, sociodemographic data, and fertilizer residue data). Finally, the 'Health, body function and disease risk' domain includes data on outcomes/effects of nutrition and of exposure to e.g. foodborne contaminants (e.g. health biomarker data, chemical toxicity data, and genomic data). The FNS data map provides a first level association between the diverse data available within the FNS field and can be used as reference to cluster, organize and tag data in the FNS Cloud.

In a second step, examples of FNS-relevant data standards, reporting guidelines and ontologies, which can aid in data harmonization and integration, are presented and linked to the FNS data domain(s) that they apply to. Similarly, all FNS datasets available to the project via the FNS-Cloud consortium are mapped onto their associated data domains and data types and linked to potential research questions. This manual mapping of FNS data to their respective data domains, data types and most importantly to potential research questions reflects processes that will take place 'in the background' (internal processing) of the FNS Cloud. The FNS data map will allow (a) to identify data requirements in response to inserted user questions and (b) to effectively retrieve information for users.

Based on the analysis of FNS data available via the project consortium, common data gaps, relating primarily to study metadata, are discussed and three main sources of inconsistencies between FNS datasets identified: (a) inconsistencies in the way concepts are expressed (nomenclature), (b) inconsistencies in the way values are measured (i.e. their units, data field types and variable type) and (c) inconsistencies in the way the data is saved (data file formats). To address such inconsistencies, applicability criteria targeting the harmonization of data nomenclature (nomenclature criteria) as well as criteria targeting the harmonization of actual data values (data value criteria) are needed.

Two illustrative research question case studies are presented within this report to demonstrate the development of data requirements and applicability criteria along two practical case examples. In the case studies, we use one single-domain and one cross-domain research question. Research questions are defined based on the availability and accessibility of datasets to the project consortium. For the single-domain question, branded food data from three European countries (NL, CH and SI) are used, whereas for the cross-domain question two datasets entailing food consumption and health biomarker data from two EU-funded projects ('Food4me' and 'Feel4Diabetes') are used. Except the Swiss data, datasets are not publicly available and were provided by the respective project partners within the FNS-Cloud consortium.

For each case study, a list of common variables or parameters included in the datasets is defined, which are necessary to answer the research questions. These variables are then considered targets of harmonization, and thus, are used as candidates for developing data requirements and applicability criteria.

Applicability criteria formalize the process of (a) checking the submitted data for any given parameter against standard, agreed upon formats, and (b) applying, where necessary, processing algorithms to harmonize the data and enable (semi-)automatic data integration and consolidation. Generic (domain-independent) criteria are developed to cover the harmonization of generic metadata parameters such as 'dataset title' or 'date', whereas domain-specific criteria cover study design, study population or sample and data type-specific parameters, such as 'physical activity' or 'food product group'.

Decision trees to represent the applicability criteria (Parameter (A) fits the defined standard format (Yes/No), if not – Parameter (A) can be transformed to defined standard format (Yes/No)) are designed for selected variables required to answer the research questions. The developed 'criteria catalogue' demonstrates the development of applicability criteria for different FNS data, providing a formalized framework for automatic data processing procedures that will take place 'in the background' to enable data integration and interoperability.

Recommendations regarding the development and application of data requirements and applicability criteria are finally provided for three stakeholder groups (target audience): for FNS-Cloud platform developers, FNS-Cloud data providers, and FNS-Cloud users. FNS-Cloud developers are advised to map FNS data to research questions, data domains and data types and use ontology-based keywords ('tags') to enable efficient conceptual matching between existing data on the Cloud and user queries. They are further advised to develop a comprehensive FNS-Cloud internal ontology for standardized annotation of incoming data, and to define a set of minimal requirements for metadata in each data domain and preferably even for each data type. Finally, a list of key variables in each dataset, that are required to answer research questions of scientific interest, shall be generated with the help of data providers and other experts in the consortium. A standard or reference format for each one of these variables for the Cloud shall be agreed upon and applicability criteria developed for them in the same way they are developed in the present report.

Data providers, on the other hand, can ease the integration of their data and enable its reuse for new questions by providing complete and if possible, ontology-based metadata, as well as following domain-specific reporting guidelines and checklists. Legal and ethical questions of data sharing must be resolved from both sides – data providers and Cloud developers – at the earliest convenience. Finally, data users are advised to clearly define their research questions, agree on their data needs to answer the question and use, where possible, ontology-based keywords to search for data on the Cloud and identify the most suitable datasets.

1. Introduction

1.1 Background and rationale

Addressing food and nutrition security (FNS), health and sustainability-related challenges involves answering complex research questions, which in turn require combining various types of data such as food composition data, food authenticity & traceability data, food consumption data and health biomarker data. It may further require the (re)use and combined analysis of data from comparable studies to validate statistical results, strengthen conclusions, improve generalizability and sometimes lead to new findings (Sansone et al., 2012). However, aligning and linking diverse data in a centralised repository of FNS-related knowledge faces several obstacles. For instance, each dataset may use its own coding systems (e.g. food classification coding). Data from different researchers, sources and laboratories are currently heterogeneous and not harmonized, which makes their conjoint use by humans as well as incorporation into information systems difficult (Eftimov et al., 2019; Muljarto et al., 2017).

The FNS-Cloud aims to provide a platform for collecting, sharing, structuring, matching and linking data within and across FNS-related research areas to support conjoint data analysis and interpretation. To achieve this, the data must be prepared following the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles for scientific data management and stewardship (Wilkinson et al., 2016). Interoperability means the ability of heterogeneous data from different data sources and laboratories to integrate or work together with minimal effort (Nogueira et al., 2016).

To make data integration (i.e. submission and consolidation of data from different data providers) and use (i.e. extraction and combined analysis of stored data by researchers or decision makers) within the FNS-Cloud possible, data stored in such a platform need to be **harmonized, structured, accessible in a machine-readable format, and made interoperable**. This can be achieved through (a) the availability of adequate metadata describing datasets and (b) the use of integrated domain ontologies (i.e. semantic data models) for standardized annotation (Ćwiek-Kupczyńska et al., 2016), which ultimately facilitates (semi-)automated data harmonization and integration.

As a starting point to systematically and semi-automatically harmonize, structure and annotate data within a data repository such as the FNS-Cloud, a set of defined data requirements (i.e. what type of data and what variables are necessary to answer a given research question) as well as a set of applicability criteria (i.e. when is data suitable for harmonization and integration) is required.

1.2 Objectives

The overarching goal of Task 3.1 (T3.1) is to provide an **operational workflow to define data requirements and applicability criteria, including an initial set of criteria.** Such workflow and initial criteria will help to develop information technology tools (e.g. machine learning algorithms) that will facilitate (semi-)automated data harmonization and integration, and enable data interoperability within the FNS-Cloud data platform.

To achieve this goal, a set of specific objectives are defined:

- To map the FNS field by identifying FNS-related 'research question spaces', research areas or data domains and types of data or datasets produced under each domain.
- To provide an overview of data standards, ontologies (or thesauri) and data reporting guidelines relevant for harmonization and standardization of FNS data.
- To identify gaps (e.g. in metadata descriptions), inconsistencies (e.g. in data nomenclature and units) and other data harmonization requirements and to provide guidance on mapping FNS-

related types of data/ datasets to identified data domains, research questions as well as related data standards, ontologies, and reporting guidelines.

- To develop a consistent initial set of data requirements and applicability criteria **along two illustrative case studies**, where, for two relevant research questions, required input data to answer these questions are systematically analysed and compared. Based on the selected research questions, data requirements to answer these questions as well as applicability criteria that will render the associated datasets interoperable will be developed and discussed. It is foreseen to include one question which requires data (datasets) from different research areas or data domains (cross-domain question), while the other question pertains to a single research area/ data domain (single domain question). This way, interoperability challenges within and across research areas and data domains will be highlighted. It is not intended to develop an exhaustive list of criteria, but to rather demonstrate the development of data requirements and applicability criteria along two practical case examples.
- To provide a set of recommendations, structured as a list of bullet points, for different stakeholder groups to facilitate data harmonization and integration and enable efficient use & reuse of available data for new research questions.

The outputs of this task build on and complement the work done in Task 2.1. A data map for the FNS-Cloud is first developed in collaboration between the two tasks' beneficiaries as a basis for structuring FNS data. While a comprehensive review of existing guidelines, data exchange models and thesauri was conducted in T2.1 to identify and recommend certain data models and advanced programme interfaces for different FNS data domains and types, a short review of data standards, reporting guidelines and ontologies was done in T3.1 to demonstrate the process of mapping such resources to different FNS data domains and types as the first step to defining data requirements and applicability criteria. The work done here is, thus, mainly aimed at proposing and demonstrating the whole workflow to define data requirements and applicability criteria for different FNS data domains and types as a prerequisite for achieving data interoperability and (semi)-automatically integrating diverse data into the FNS-Cloud.

1.3 Target audience

As target audience of the present document, we identified three stakeholder groups, which are targeted with our criteria catalogue:

- **FNS-Cloud platform developers**: this stakeholder group includes academic and technical experts developing FNS-Cloud approaches and technologies for handling and (pre-) processing any type of data in a (semi-)automated and structured manner. **Our data requirements and criteria catalogue provide a starting point for and formalize the process of** identifying the connections across FNS-relevant data types and data domains, mapping data(sets) to related data domains, specific user questions, ontologies, standards and guidelines and developing the required interfaces and algorithms to semi-automatically harmonize, structure and integrate data for joint analysis, interpretation and visualization.
- **FNS-Cloud data providers:** this stakeholder group includes academic, industry, regulatory and any other users who wish to make their data available for the FNS-Cloud platform under their respective licence agreements. **Our data requirements and criteria catalogue provide initial pre-processing recommendations (e.g. metadata descriptions)** that can be implemented by the data providers themselves, which will ease upload and subsequent machine processing and integration of their data in a consistent, standardized and automated way into the FNS-cloud.
- **FNS-Cloud platform users:** this stakeholder group includes mainly but is not limited to researchers, public and private decision makers, and regulatory agencies, aiming to use various

types of data from one or multiple studies to answer questions related to food safety and sustainability, nutrition and health. **Our data requirements and criteria catalogue help users to rapidly and consistently identify, access and interpret available data** that can help answer their specific research question.

2. Delineating the field of food and nutrition security

2.1 FNS-related research questions, data domains and types of data

As a starting point, FNS-relevant research questions, as exemplified by WP4 Use Cases and WP5 Demonstrators, were defined (for more details, see Section 3.1). Research questions were grouped into 'research question spaces', describing primary research themes and interests or research purposes within the FNS field. Based on the research question spaces and in close alliance with the 'data map' described in D2.1 (Presser et al., 2020), **three data domains, which capture knowledge areas and research foci of the FNS field, were defined** (see Figure 1):

- 'agrifood'
- 'intake of food and lifestyle'
- 'health, body function and disease risk'

Each data domain includes different data types generated and used to answer a variety of research questions. The 'agrifood' data domain covers all data types related to the food product itself, i.e. its quality & safety, its production (e.g. farming activities) and sustainability. The agrifood domain can thus include 'traditional' food nutrient composition data, agronomic performance data for crop varieties or branded food data, as well as more novel food omics data (e.g. proteomic and metabolomic profiles of foods). The 'intake of food and lifestyle' domain covers consumers and consumer behaviour which is captured in traditional food intake studies e.g. food consumption data, sociodemographic data, as well as food choice and lifestyle data. Finally, the 'health, body function and disease risk' domain covers the link between diet and health and includes for instance phenotypic data (biomarkers, anthropometry), genotypic data and hazard data (e.g. bioaccumulation and toxicity potency of food additives).

The resulting 'FNS data map' presented in Figure 1 provides a first-level association between different datasets and their link to the three main FNS data domains. It also serves as a basis for characterizing research questions as cross-domain or domain-specific questions and identifying data requirements (i.e. data types and datasets) to answer these questions (Figure 1).

Some data types are used for multiple purposes, i.e. they're linked to different questions/question spaces (i.e. questions related to the same research field, such as risk assessment, or aspect, such as human disease outcomes). For example, food composition data can be used for food quality related questions, calculations of nutrient intake, as well as exposure and risk assessments. Elemental data (a subset of food composition data) are used for nutrition analysis as well as food traceability. Food consumption data are used for defining nutrient intake and dietary recommendations, as well as determining the link between nutrition and health outcomes. With the help of e.g. natural language processing technologies, components of research questions can be analysed and datasets available within the FNS-Cloud that match to new user questions identified and retrieved for analysis. This does not only require that available datasets are correctly mapped to their respective data domains and types, but that they're 'tagged' with ontology-based keywords to help achieve a conceptual connection between the user's question and the available machine-readable information (Humphreys & Lindberg, 1993).

Some FNS-relevant research questions may also require or benefit from data lying outside of the FNS field. For example, in order to determine the geographical origin of food in food authenticity and traceability studies, soil isotope data can complement the stable isotope analysis of food samples (Katerinopoulou et al., 2020). The FNS data map can therefore be later expanded to include external domains linked to FNS and data mining and integration techniques explored to allow inclusion of external data within the FNS-Cloud.

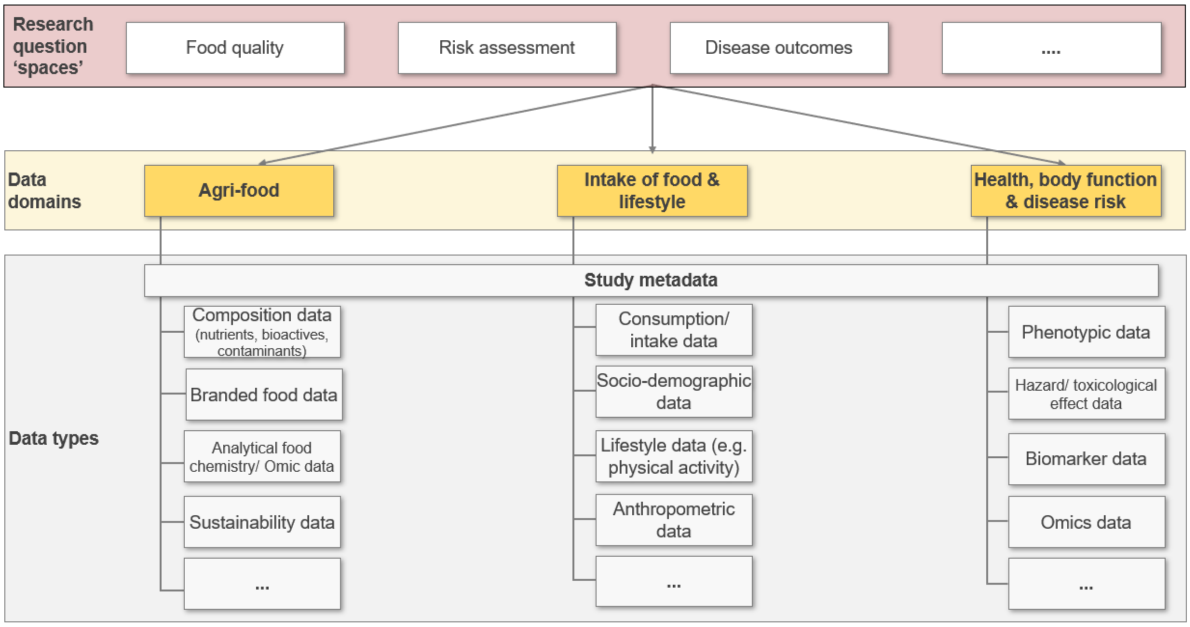

Research question 'spaces'
Food quality
Risk assessment
Disease outcomes
....
Data domains
Agri-food
Intake of food & lifestyle
Health, body function & disease risk
Data types
Study metadata
Composition data (nutrients, bioactives, contaminants)
Branded food data
Analytical food chemistry/ Omic data
Sustainability data
...
Consumption/ intake data
Socio-demographic data
Lifestyle data (e.g. physical activity)
Anthropometric data
...
Phenotypic data
Hazard/ toxicological effect data
Biomarker data
Omics data
...

Figure 1: FNS data map, representing FNS-relevant research question spaces, FNS data domains as well as associated data types. Three dots "…" indicate that further elements can be defined for questions and data types.

2.2 FNS-related data standards, ontologies and data reporting guidelines

Considering its multidisciplinary nature, the FNS field produces data from a variety of scientific disciplines, research areas or data domains and, thus, contains heterogeneous data that can significantly differ in their structure, format and annotation. Several data standards, reporting guidelines (called hereafter standards and guidelines) and shared terminologies/ controlled vocabularies (thesauri) or ontologies have already been developed (e.g. Bodenreider, 2004; Dooley et al., 2018; Eftimov et al., 2019; Vitali et al., 2018) with the aim to unify data of the same type, harmonize reporting of study results in a given domain (e.g. Lachat et al., 2016; Morrison et al., 2007; Pinart et al., 2018; Taylor et al., 2007), standardize annotation and encourage good data stewardship and sharing (Sansone et al., 2012).

Existing standards and guidelines as well as available thesauri and ontologies can be used as a basis for harmonizing, structuring and annotating data within the FNS-Cloud. They can thus provide a starting point for fulfilling data requirements and applicability criteria, either at the stage of data generation (by the researchers or data providers themselves) or at the stage of semi-automated data integration and annotation (by the FNS-Cloud systems).

Table 1 includes examples of FNS-relevant standards and guidelines, while Table 2 presents ontologies and thesauri. The examples selected represent widely known or widely applied standards, guidelines and ontologies within the different data domains of the FNS field. All listed resources were linked to the respective FNS data domain that they apply to. This initial linking allows data (datasets) of a given type that are uploaded to the FNS-Cloud to be automatically mapped to the associated data domains and in turn to related guidelines, standards and ontologies.

Some resources cover more than one FNS data domain and include harmonization recommendations and provisions related to multiple data types such as "STROBE-nut", a guideline for reporting nutrition epidemiology and dietary assessment research (Lachat et al., 2016). Other resources are of an even more comprehensive nature, spanning over several research fields (e.g. apply to all the biomedical sciences). For example, the ISA-Tab format has been developed to address descriptions for many types of experiments and assays and constitutes a general experimental metadata description standard (Sansone et al., 2016). It consists of a set of tab-delimited text files, namely Investigation, Study, and Assay files, that are linked to each other to form a hierarchy, and describe different properties of a scientific undertaking (e.g. the title, goals, methods, participants, experimental design, environmental conditions and treatments).

Ontologies play a critical role in achieving semantic interoperability as they support automated integration and standardized annotation of study descriptors and later enable a more accurate search for required data and efficient analysis of data from multiple sources (Eftimov et al., 2018). Ontologies link knowledge by defining relations between key concepts and entities of given fields as well as allow the generation of new knowledge (Yang et al., 2019). Naturally, in any given ontology certain classes and instances, especially those relating to study metadata (e.g. 'Date', 'Title', 'Assay'), may be applicable to any FNS data domain (and beyond); yet the ontology is assigned only to the specific data domain (field of knowledge) it intends to map (see Table 2). Cross-domain ontologies, such as the ONS-ontology (Vitali et al., 2018) or the 'Unified medical language system'-ontology (Bodenreider, 2004; Lindberg et al., 1993), which incorporate unique vocabularies and axiomatic linkages from across different FNS-related data domains, can be assigned to several domains and constitute excellent candidates for use in the harmonization and structuring of data within the FNS-Cloud and integration (import) of concepts into an internal FNS-Cloud ontology.

Ontology look-up services such as the National Center for Biomedical Ontology BioPortal (https://bioportal.bioontology.org) and the EMBL-EBI Ontology Lookup Service (https://www.ebi.ac.uk/ols/) facilitate the search for and identification of existing ontologies. Additionally, the annotator tool (https://bioportal.bioontology.org/annotator) on the BioPortal enables users to search for ontologies which include classes and annotations for any 'concept' (e.g. nutrient, additive, glucose, apple). For example, searching for the concept 'ethnicity' provides 35 different ontologies (e.g. Neuroscience Information Framework Standard ontology, Medical Subject Headings, Clinical Study Ontology, Gender, Sex and Sexual Orientation ontology). FNS-relevant concepts can be imported from other ontologies into an FNS-Cloud internal ontology, which will later allow users to store ontology-based data. If no suitable matches in existing ontologies are found, new concepts and relationships can be defined for the new FNS-Cloud internal ontology.

Similar to the ontology look-up service, the EQUATOR Network was created as a central repository and organization to improve the quality of reporting guidelines. A comprehensive list of reporting guidelines is available at their website (http://www.equator-network.org/), together with toolkits and flow charts to choose the most suitable reporting guideline for a researcher's article.

Table 1: Examples of relevant standards and guidelines applicable within the FNS field mapped to the three main FNS data domains

Standard or guideline	Standardizing body	Description/ main purpose	Data domain[1]			Reference
			Agrifood	Food intake and lifestyle	Health, body function and disease risk	
ISA Framework	**ISA community**	Helps you to provide rich description of the experimental metadata (i.e. sample characteristics, technology and measurement types) for research related to life science, environmental and biomedical experiments	x^+	x^+	x^+	(Sansone et al., 2016)
ISO 80000-1:2009	**ISO**	A standard describing scientific and mathematical quantities and their units	x^+	x^+	x^+	(ISO 80000-1:2009)
EuroFIR Technical Standard	**EuroFIR**	Describes the framework for the standardisation of food composition data carried out by the EuroFIR Network of Excellence	x			(Becker et al., 2008)
European standard for food data (EN 16104:2012, Food data - structure and interchange format)	**BSI**	Specifies requirements on the structure and semantics of food datasets and of interchange of food data for various applications	x			BS EN 16104:2012
EU Menu Methodology (& data schema)	**EFSA**	Indicate methodological principles and protocols for the collection of high-quality, harmonized individual dietary information within a pan-European context for use in dietary exposure assessments		x		(EFSA, 2014)
STROBE-nut: guideline for reporting nutrition epidemiology and dietary assessment research	**University of Bern**	Recommendations for reporting nutritional epidemiology and dietary assessment research		x	x	(Lachat et al., 2016)
EFSA Standard Sample Description ver. 2.0	**EFSA**	Provides specifications aimed at harmonising the collection of analytical data on chemical substances and microbiological agents in different matrices of non-human nature (e.g. food, feed, animals, water, environmental samples and food contact materials).	x			(EFSA, 2013)
MIxS standard	**Genomic Standards Consortium**	Standard for reporting of minimum information about any (x) nucleotide sequence. It consists of three separate checklists; MIGS for genomes, MIMS for metagenomes, and MIMARKS for marker genes.	x		x	(Yilmaz et al., 2011)
Consolidated Standards of Reporting Trials (CONSORT)	**The CONSORT group**	Guidelines for the reporting of randomised controlled trials in healthcare interventions			x	(Schulz et al., 2010)

[1] if all three data domains are selected and marked with x^+, the resource applies to all FNS-relevant data domains **and beyond** (i.e. generic standards or guidelines, not specific to the FNS field, applying to e.g. the natural sciences as a whole)

Table 2: Examples of relevant ontologies and thesauri applicable within the FNS field mapped to the three main FNS data domains

Ontology	Acronym	Description/ Main purpose	Data domain[1]			Reference	URL
			Agrifood	Food intake and lifestyle	Health, body function and disease risk		
EuroFIR reference type thesaurus	**n.a.**	Details of bibliographical references describing documents that are sources of data for value, method, recipe, etc.	x	(x)[2]	(x)[2]	(Macháčková et al., 2017)	http://www.eurofir.org/our-resources/eurofir-thesauri/
EuroFIR value type thesaurus	**n.a.**	Description of the data values or a qualitative description of the value when no value can be given	x	(x)[2]	(x)[2]	(Macháčková et al., 2017)	http://www.eurofir.org/our-resources/eurofir-thesauri/
EuroFIR matrix unit thesaurus	**n.a.**	Terms for the amount of the matrix material that has quantity reported as the value, usually expressed using the preposition 'per'.	x			(Macháčková et al., 2017)	http://www.eurofir.org/our-resources/eurofir-thesauri/
Statistics Ontology	**STATO**	Describing key statistical measures, such as p value, mean, standard deviation	x^+	x^+	x^+	(ISA-tools, 2014)	https://github.com/ISA-tools/stato
Units of measurement ontology	**UO**	Metrical units for use in conjunction with the Phenotype and Trait Ontology (PATO) framework for describing qualitative and quantitative observations in biology	x^+	x^+	x^+	(Gkoutos et al., 2012)	https://github.com/bio-ontology-research-group/unit-ontology
EFSA Food classification and description system for exposure assessment (version 2)	**FoodEx2**	A standardised food classification and description system, consisting of descriptions of a large number of individual food items aggregated into food groups and broader food categories in a hierarchical parent-child relationship	x	x		(EFSA, 2014)	https://github.com/openefsa/catalogue-browser/wiki
Langua aLimentaria	**LanguaL**	A standardised language for describing foods, specifically for classifying food products for information retrieval.	x	x		(Møller & Ireland, 2017)	https://www.langual.org/Default.asp
Food Ontology	**FoodOn**	An ontology to represent entities which bear a "food role" and develop semantics for food safety, food security, agricultural and animal husbandry practices, nutrition and chemical ingredients and processes	x	x		(Dooley et al., 2018)	https://github.com/FoodOntology/foodon
ONS Ontology for Nutritional Studies	**ONS**	A formal ontology framework for the description of nutritional studies	x	x	x	(Vitali et al., 2018)	http://bioportal.bioontology.org/ontologies/ONS
ISO-FOOD	**-**	Ontology for describing isotopic data within Food Science	x			(Eftimov et al., 2019)	https://bioportal.bioontology.org/ontologies/ISO-FOOD
Systematized Nomenclature of Medicine - Clinical Terms	**SNOMEDCT**	Provides the core general terminology for electronic health records. It includes clinical findings, symptoms, diagnoses, procedures, body structures, organisms and other aetiologies, substances, pharmaceuticals, devices and specimens			x	(El-Sappagh et al. 2018)	https://bioportal.bioontology.org/ontologies/SNOMEDCT
OntoFood	**OF**	An ontology with of nutrition for diabetic patient		x	x	none	https://bioportal.bioontology.org/ontologies/OF
Unified medical language system	**UMLS**	A repository of biomedical vocabularies developed by the US National Library of Medicine	x	x	x	(Bodenreider, 2004; Lindberg et al., 1993)	https://www.nlm.nih.gov/research/umls/index.html
Quisper Ontology	**-**	An ontology to capture food and nutrition related data in e-health systems. It can be used to harmonize personalized dietary web services	x	x	x	(Eftimov et al. 2018)	-

Ontology	Acronym	Description/ Main purpose	Data domain[1]			Reference	URL
			Agrifood	Food intake and lifestyle	Health, body function and disease risk		
The Ontology for Biomedical Investigations	**OBI**	Provides terms with precisely defined meanings to describe all aspects of how investigations in the biological and medical domains are conducted			x	(Bandrowski et al., 2016)	http://obi-ontology.org/
Genomic Epidemiology Ontology	**GenEpiO**	Covers vocabulary necessary to identify, document and research foodborne pathogens and associated outbreaks of infectious diseases.	x		x	-	http://www.obofoundry.org/ontology/genepio.html
Chemical entities of biological interest	**ChEBI**	Dictionary of molecular entities focused on 'small' chemical compounds. The molecular entities in question are either natural products or synthetic products used to intervene in the processes of living organism	x			(de Matos et al., 2009)	https://www.ebi.ac.uk/chebi/
Environment Ontology	**ENVO**	Ontology of environmental features and habitats	x			(Buttigieg et al., 2013)	http://www.obofoundry.org/ontology/envo.html
AGROVOC	-	Controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations, including food, nutrition, agriculture, fisheries, forestry, environment etc.	x			(Soergel et al., 2004)	http://aims.fao.org/vest-registry/vocabularies/agrovoc
Open Food Facts Food Ontology	-	Models the Food domain. It allows to describe ingredients and food products. Ontology used by the Open Food Facts dataset	x			-	https://wiki.openfoodfacts.org/Main_Page

[1] if all three data domains are selected and marked with x^+, ontology applies to all FNS-relevant data domains **and goes beyond** (i.e. generic ontology, not specific to the FNS field, applying to e.g. the biomedical sciences as a whole)

[2] EuroFIR thesauri have been developed specifically for the standardization of terminology used in food composition datasets. However, some of them are of a generic nature and can be applied to other data domains (e.g. value type or reference thesaurus). They may however require further extensions to better cover other data domains (x). For example, the reference type thesaurus includes sources typical for food composition data (e.g. product label). It can be expanded to include e.g. medical records for health data.

3. FNS data gaps, inconsistencies, and needs for harmonization

3.1 Common data and data description gaps

FNS data or datasets can include hundreds of different measured or computed variables. Data of the same type (e.g. food composition data or health biomarker data) often entail a set of commonly measured variables, which usually fulfil the main purpose or research interest that the data is generated for. For example, food composition data will always – at least – contain a food name (and facet description if applicable) and at least one measured component (a nutrient, bioactive or contaminant) and the measured value(s). Given that the data collected or the variables measured in each dataset are always dependent on the research question and the approach selected by the researcher, there is no standardized or mandatory list of variables that must be measured and reported in each dataset of a given data type.

However, what are often missing (not reported) and for which (mandatory) harmonized checklists may exist, are metadata. Metadata is descriptive information about data, i.e. it describes source, content, collection methodology and other characteristics/ details about data (GovEx 2016). Scientific metadata can include for instance:

- dataset title (or name)
- author(s), publisher
- date created, date modified
- description
- keywords
- contact person
- access level (public, restricted, non-public) or license agreement
- written language (e.g. English)
- reference types (e.g. journal article, book/book chapter, software, webpage)
- value types (e.g. average, median, weighted, best estimate)
- experimental design
- characteristics of the samples and procedures applied

Metadata helps people find data (via e.g. internet searches) as well as enables correct interpretation of data, and thus, data comparability and interoperability (Ćwiek-Kupczyńska et al., 2016). Table 3 presents examples of often missing metadata (data gaps) specific to certain FNS data types, which were identified by analysing some of the FNS datasets available to the project consortium as well as through direct discussions with FNS data providers involved in the project. While often not reported along with the published results of experimental studies, this metadata usually relates to key input parameters in different types of models or provides relevant information for answering certain research questions (Sansone et al., 2016; Taylor et al., 2008). For example, Fantke et al. (2016) provide an extended overview of data (parameters) that are often missing in empirical studies on plant bioaccumulation of chemicals, yet they represent key parameters relevant for food intake analysis and risk assessment models.

If (FNS-related) data from multiple sources, studies and laboratories is to be made interoperable, and data reuse, interpretation, comparison and joint analysis to be made possible, **correct, harmonized and complete metadata is a prerequisite**. Several initiatives to develop a set of minimum (meta)data reporting requirements for studies in a given field follow the hierarchical structure developed by the Investigation, Study, and Assay (ISA) Commons (Sansone et al., 2016). These ISA metadata categories establish the minimum background information that may help with respect to interpreting results and having a better picture of the context of the study as well as the methods used, data collected, and conclusions drawn. For example, within the European Nutritional Phenotype Assessment and Data Sharing Initiative (ENPADASI), a list of forty-one descriptors of minimal requirements for study data from observational nutritional studies was identified adhering to the ISA structure (Pinart et al., 2018)**.** These requirements facilitate data exchange, data interpretation and data integration into infrastructures such as the FNS-Cloud. Similarly, the Dublin Core™ Metadata Initiative (DCMI) establishes a common, cross-domain metadata vocabulary and provides access to schemas (structure and syntax) defining DCMI term declarations (DCMI, 2020). The DCMI has published several standards to describe and standardize its metadata specifications (ISO 15836-1:2017; Z39.85-2012), which can also serve as a basis in developing domain-specific, minimal metadata requirements for different data types within the FNS-Cloud infrastructure.

Table 3: Examples of common (meta-) data gaps in different FNS-relevant data types

Data type	**Examples of data gaps in reported datasets**
Food intake data	• Dietary assessment method used (24-h recall, food frequency questionnaire ... etc.) • Physical activity questionnaire applied • Description of dietary supplements taken
Branded food data	• Source of data • Time point of data collection • Food classification system used
Toxicological hazard data	• Endpoint tested • Exposure route • Unit interpretation (e.g. ppm mass-based or volume-based)
Meta-genomic data	• Platform details • Analysis details • Diet of sample donors

3.2 Existing data nomenclature, units, and other inconsistencies

Given the diversity of data types generated and used within the FNS field, significant inconsistencies between datasets exist in multiple aspects, the most relevant of which are:

- the way the data (i.e. measured variables or concepts) are named (nomenclature)
- the units used for measured variables
- the variable type (continuous, categorical)
- the data field type for any given parameter (e.g. text/string, number or date)

In order to make FNS data from different sources interoperable, harmonized vocabularies and domain ontologies need to be applied (or developed) to reach a standardized nomenclature, and differences in e.g. units or data field type alleviated with the help of e.g. algorithms (Eftimov et al., 2018). These inconsistencies in terminologies and in values/value properties exist not only between different data domains and data types, but also within domains and same data types. For example, food consumption data can be collected using multiple tools: food frequency questionnaires (FFQs), 24-h dietary recalls, diet history or diet records. Each of these data collection tools covers different timeframes, asks different questions (e.g. recording of consumed food items or food groups) and returns similar data, yet at different aggregation levels and offering different information/content (Thompson & Subar, 2017). To what extent is then food consumption data comparable and when (how) can it be compared or even merged? To answer these questions, applicability criteria must be developed to help check available data (sets) against a standard format (i.e. standard nomenclature and meaning, structure and value type) and transform them into the desired standard format where necessary. Applicability criteria are thus a formal representation of the data (pre-) processing that will take place in the 'background' to enable (semi-)automated data harmonization and integration.

To address the different inconsistencies among datasets, **two types of applicability criteria were defined:**

A. **Nomenclature criteria**: these criteria relate to the terminology (vocabulary) used for measured parameters, as well as the meaning (semantics) of this vocabulary. Is it 'food name' or 'food item' (and is that the same thing?), is it 'participant_ID' or 'subject_ID', is it 'gender' or 'sex'?

B. **Data value criteria**: these criteria relate to all aspects or attributes of the data value itself and how it is measured/ given. This includes, the variable type (continuous/ categorical), the value type (e.g. mean, median, weighted, 95%-ile, larger than), the data field type (e.g. number, string, Boolean or picture) and finally the units used. For a list of possible data field types see D2.1 (Section 4.3.1) (Presser et al., 2020).

Criteria of both types (nomenclature and data value criteria) must be developed for all commonly measured and reported variables that require harmonization and standardization between datasets to enable interpretation, comparison, re-exploitation and joint data analysis (Doiron et al., 2013; Sansone et al., 2012). Some of these criteria will be **generic or domain-independent**, i.e. they apply to all FNS-related data types and data domains. Other criteria are **domain-specific**, meaning they only apply to datasets generated within one FNS data domain or even datasets of a specific data type.

Generic (domain-independent) criteria can be broadly clustered into the following groups:

A. Generic metadata related criteria: target metadata that is relevant for all FNS datasets, which include e.g. study title, country, description, funding body ... etc. These metadata parameters pertain mostly to the metadata category 'Investigation' of the ISA framework
B. Food and food description related criteria: capture all aspects of food identification and description, which are generic (equal) for all data types which include food entity related parameters (e.g. food name, food group, preparation method)

Domain-specific criteria, on the other hand, can be broadly clustered into the following groups:

A. Study methods related criteria
B. Sample or study population related criteria
C. Data type-specific criteria

Similar to the 'FNS data map', a map of the different criteria types and groups is presented in Figure 2.

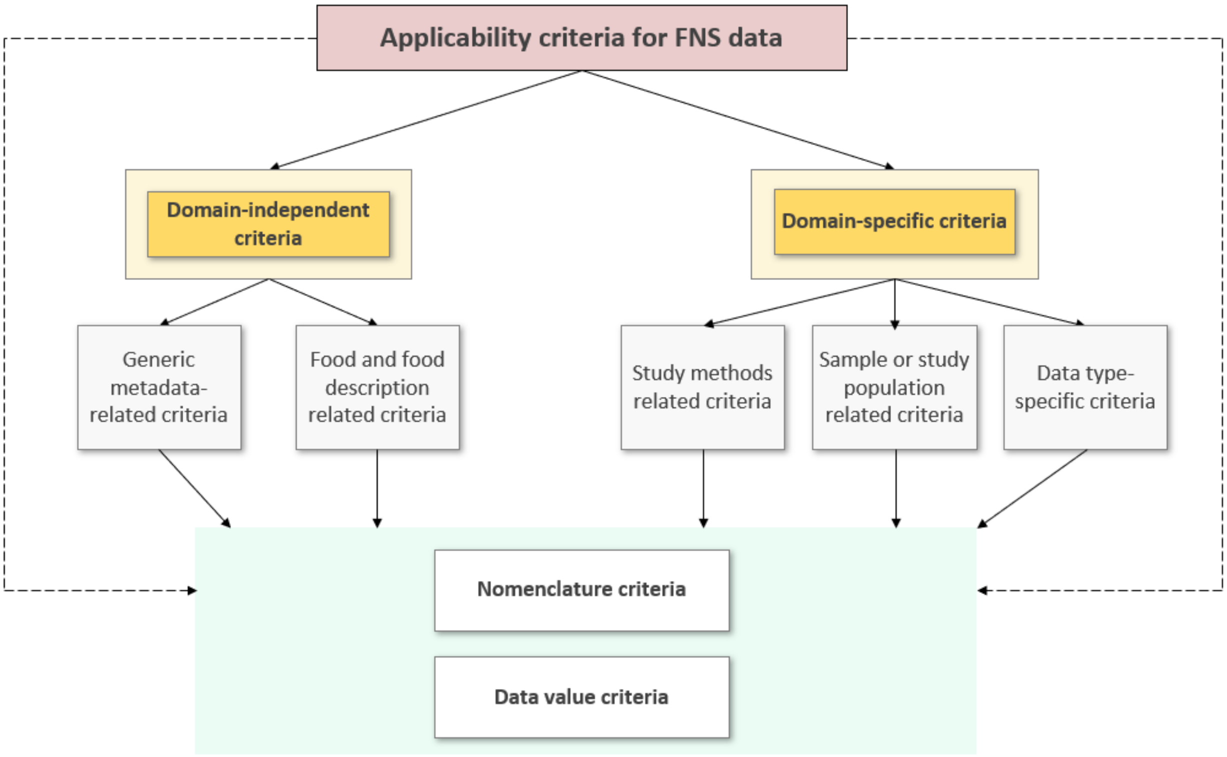

Applicability criteria for FNS data
Domain-independent criteria
Domain-specific criteria
Generic metadata-related criteria
Food and food description related criteria
Study methods related criteria
Sample or study population related criteria
Data type-specific criteria
Nomenclature criteria
Data value criteria

Figure 2: Applicability criteria for FNS data divided into different criteria types and groups

Even though study methods and sample or study population also refer to metadata aspects (mostly covered under the metadata categories ‘Study’ and ‘Assay’ of the ISA framework), **these aspects (i.e. which descriptors are relevant and which values can they assume) will – unlike food entity related descriptors – naturally depend on the specific study and data type.** For example, in food composition or food authenticity data the study methods parameters can be termed ‘analytical methods’ and the options given can include spectrometry (with different possibilities such atomic emission or near infrared spectroscopy), chromatography (with different options such as gas or high performance liquid chromatography) or calorimetry (e.g. enzymatic, dye binding). For food consumption data, the study methods descriptor can be termed ‘dietary assessment tool’ and include the following options: dietary record, 24-h recall, FFQ, diet history or other. The same applies to the sample or study population-related criteria. Therefore, criteria developed for these variables, which are also metadata descriptors, will nevertheless be domain-specific.

If metadata requirements (checklists) are developed for each individual data type (or for a group of related data types), many applicability criteria of the criteria groups ‘generic metadata’, ‘study population or sample’ and ‘study methods’ would be already harmonized; potentially in both their nomenclature and their value, depending on how elaborate the requirements/ checklists are.

Finally, data type-specific criteria relate to parameters that are unique to a given data type (‘non-metadata’). For example, the parameters ‘portion size’ or ‘frequency of consumption’ are routinely collected as part of food consumption assessments and are unique to this type of data. Developing criteria for data type-specific parameters, which provides a basis for direct data integration or conversion into a standard format using different IT approaches, is mostly important for making data of the same type collected in different studies, countries, laboratories etc., interoperable (Presser et al., 2018). This is particularly important in cases where the same data type is collected and used for different purposes.

Every variable in a given dataset, that is selected for harmonization, requires a nomenclature criterion and one or several data value criteria (see Figure 2). For example, the parameter ‘glucose’ is typically measured in dietary intervention studies to assess the efficacy of a given treatment or prevention strategy for type-2-diabetes. It thus belongs to the data type 'health biomarker data' within the data domain of 'health, body function and disease risk'. If datasets from different intervention studies are to be compared or linked, this parameter must be reported in a harmonized and standardized way. However, the parameter may be called differently in each dataset (e.g. 'blood glucose', 'fasting plasma glucose (FPG)' or 'fasting blood sugar level'), depending on the accuracy of vocabulary used (and whether it follows an existing ontology or not) as well as the analytical method to measure glucose that was used. From the term 'blood glucose' alone it is not clear if that is fasting plasma glucose, measured typically the morning after 10 hours of fasting, or whether this is random glucose, measured at a random point in time during the day. Inconsistencies in semantics make the comparison of that value with FPG from other datasets not possible. Similarly, the value measured can be given in mmol/L or mg/dL, or even expressed in categorical terms (normal, prediabetes/ insulin resistance or diabetes). Harmonization of the parameter 'glucose' will thus require several 'data value criteria', one related to the variable type (categorical vs. continuous) and one related to the data field format (number vs. string).

Units are unique in that they require a generic and a domain-specific criterion. On the one hand, units shall follow the International System of Units (SI) or be convertible into SI units (generic criterion). On the other hand, non-SI units are defined in specific domains and require domain-specific standardization. For example, the Units of Measurement ontology defines units for measurements within the field of biology/ biomedical sciences. Both the nomenclature and the meaning of such domain-specific units must be agreed upon. For any given measured variable from a specific data type, a reference unit including matrix units (e.g. per 100g, per package) will have to be defined in an FNS-Cloud internal ontology and selected as standard unit. Only then can values reported in different units be converted to the standard unit, where necessary (domain-specific criterion). Additionally, the 'meaning' of units will have to be harmonized, again with the help of an FNS-Cloud ontology. Is 'ppm' (parts per million) used for a given parameter mass-based or volume-based? Does % relate to dry weight or wet weight?

Inconsistencies in data format, nomenclature, variable type, units ...etc. for any given parameter hamper data integration and interoperability. Therefore, it is necessary to define applicability criteria for commonly reported parameters that are required to answer FNS-relevant research questions, while considering both sources of potential inconsistency (nomenclature and actual value). This will be later shown along two illustrative case studies to demonstrate the development and application of such criteria based on two selected research questions.

3.3 Other needs for data harmonization

Aside from data gaps and existing inconsistencies in data nomenclature, data field types and units, the format in which the data is saved may require harmonization to allow data transfer and merging without loss of information.

Unstructured data files often include **text** and **multimedia content**. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents, including container formats, such as the Portable Document Format (PDF) and onscreen visuals.

This type of data is most difficult to process because FNS entities can be specified in an unspecified way. Other tasks in WP3 deal with this type of data by performing so called **information extraction** on unstructured data.

Semi-structured data, on the other hand, does not reside in fixed fields or records, but does contain elements that can separate the data into various hierarchies. Examples of semi-structured data are:

- html page: a text-based format

- Fastq.gz: gz is the compressed version of a fastq file
    - Fastq: a text-based format for storing both a biological sequence and its corresponding quality scores
    - it can be parsed using specific software
- CSV: a simple file format used to store tabular data, such as a spreadsheet or database. It can be converted to the Excel format
- JavaScript Object Notation (JSON): is an open standard file format, and data interchange format, that uses human-readable text to store and transmit data objects consisting of attribute-value pairs and array data types (or any other serializable value).
    - It can be converted to the Excel format
    - It can be converted to the XML format
- Extensible Markup Language (XML): format that is both human-readable and machine-readable
    - It can be converted to the Excel format
    - It can be converted to the JSON format

This type of semi-structured data is less complex and can be structured. Here, the main problem is to match data being described using different description and classification systems. For example, vitamin C can be named in one system as VITC, in another as ascorbic acid and in the third one as Vitamin C. To solve this problem, we can use an **ontology to normalize data**, where FNS terms and relations between them are described.

Structured data is highly organized and easily understood by machine language. Examples include:

- Excel: a file format developed by Microsoft to present spreadsheets. It can be converted to CSV format
- Structured Query Language (SQL) file format: is a format written in ASCII and used by database products. An SQL file typically contains queries with data to modify the structure of a relational database. Tables from the database can be converted into the Excel format
- SPSS Statistics File Format: is a proprietary binary format, developed and maintained as the native format for the SPSS statistical software application. It can be converted to the Excel format
- sdf file: a compact relational database saved in the Microsoft SQL Server Compact (SQL CE) format. It can be converted to the SQL format

Here, the problem is the same as with the semi-structured data, i.e. data may be described using any description and classification system.

4. Data requirements and applicability criteria along an illustrative case study

4.1 Identifying data requirements for research questions

As seen in Figure 1 FNS data cover a variety of data types characterized by a high degree of heterogeneity. Each of the data types can represent hundreds of different datasets with multiple (commonly) measured and reported variables or parameters, which are required to answer a multitude of research questions related to food, nutrition and health. Even within the same data type there can be significant differences in the way the data is represented, documented and saved, which hampers data interpretation, extraction (by humans or machines) as well as data integration and interoperability (Sansone et al., 2012; Taylor et al., 2008).

Data requirements define which data types and datasets are needed to answer a given research question, while applicability criteria ensure that data (or a selected subset of it) are suitable for (semi-)automated data harmonization and integration into the FNS-Cloud, as well as for conjoint use and analysis to answer a given research question. Research questions thereby also include any type of user query to find, match, combine or compare data. **Hence, data requirements and applicability criteria are primarily dictated by the research questions addressed within the FNS domain**. We therefore analysed datasets available to the FNS-Cloud project consortium via WP4 and WP5 (or will be developed during the project) in order to map them to (1) the data domains they are linked to as well as the data types they represent and (2) the research questions they can be used for to answer, following intentions of WP4 Use Cases and WP5 Demonstrators (see Table 4 in Appendix 1). WP4 aims to generate novel FNS data through proof-of-principal Use Cases in all three FNS data domains (Agri-Food, Nutrition and lifestyle, Health, body function and disease risk). Existing and new FNS data will then be used for WP5 Demonstrators to test the FNS-Cloud tools, services and infrastructure, as well as showcase the potential for re-use of data by user communities. Data management methodologies to be developed in WP3 to support data upload, integration, (co-)analysis and exploitation are guided by the needs of WP4 Use Cases and the feedback from WP5 Demonstrators.

Figure 3 shows one example of datasets from the FNS-Cloud project consortium linked (mapped) to related data domains and data types, and, in turn, to a specific research question. It thus formally represents the process of determining data requirements for any given user question. When a user types in a research question in natural language, with the help of NLP technologies, the respective data domains and data types can be determined and related datasets available on the FNS-Cloud retrieved for analysis. Figure 3 illustrates this process using several examples. For instance, the use of the words ‘high-bioactive’ and ‘low-bioactive’ within the question leads to the identification of eplantLIBRA and eBASIS as available datasets, as these are composition and biological activity databases for bioactive (non-nutrient) compounds in plant foods (eBASIS) and food supplements (eplantlibra). The use of the word “microbiome” leads to the identification of gut microbiome data and metabolomics data. Complete and properly annotated metadata about available datasets on the FNS-Cloud, as well as (semi-)automatic pre-processing of the data (i.e. mapping datasets to their associated data domains and types), will help match user requirements to the existing data. Additionally, available FNS datasets could be tagged with several other keywords that help describe their content and purpose and associate them with respective research questions.

For each research question, specific parameters or variables within the retrieved/selected datasets (i.e. only a subset of the available data) will actually be required to answer it. This ‘deeper level’ of data requirements will be determined by each researcher or decision maker (FNS-Cloud user) and will be dependent on their unique research question. **Key variables that are required to answer specific research questions constitute targets of harmonization and standardization (i.e. applicability criteria)** in order to enable data interoperability, reuse and repurposing as well as integrative analysis (see Figure 3). Examples of such key variables requiring harmonization and the applicability criteria developed for them are presented along the two illustrative case studies in the following sections.

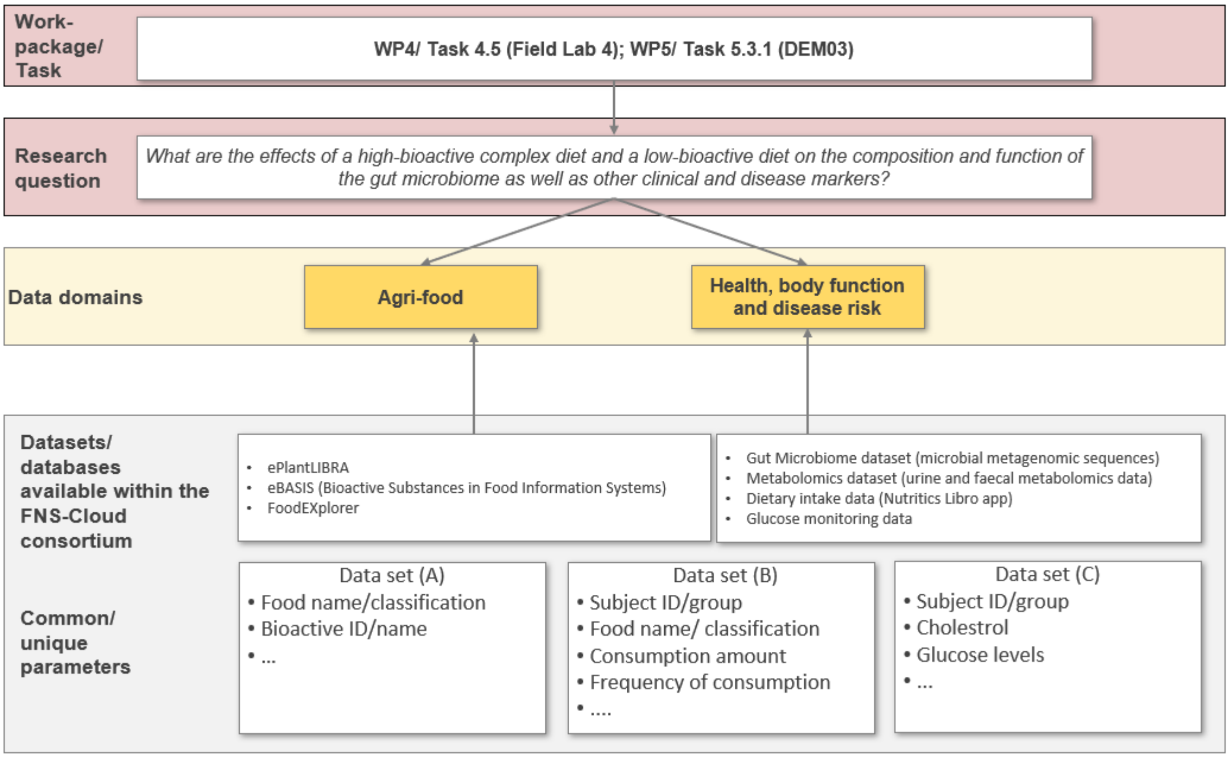

Work-package/ Task
WP4/ Task 4.5 (Field Lab 4); WP5/ Task 5.3.1 (DEM03)
Research question
What are the effects of a high-bioactive complex diet and a low-bioactive diet on the composition and function of the gut microbiome as well as other clinical and disease markers?
Data domains
Agri-food
Health, body function and disease risk
Datasets/ databases available within the FNS-Cloud consortium
• ePlantLIBRA
• eBASIS (Bioactive Substances in Food Information Systems)
• FoodEXplorer
• Gut Microbiome dataset (microbial metagenomic sequences)
• Metabolomics dataset (urine and faecal metabolomics data)
• Dietary intake data (Nutritics Libro app)
• Glucose monitoring data
Common/ unique parameters
Data set (A)
• Food name/classification
• Bioactive ID/name
• ...
Data set (B)
• Subject ID/group
• Food name/ classification
• Consumption amount
• Frequency of consumption
• ....
Data set (C)
• Subject ID/group
• Cholestrol
• Glucose levels
• ...

Figure 3: Example FNS data (can be raw and/or pre-analysed data) mapped to related data domains and in turn to a specific research question from WP4, Task 4.3 and WP5, Task 5.1.3

4.2 Case study definition (along two selected research questions)

The selection of two research questions to serve as case studies in T3.1 was primarily dictated by data availability and accessibility to the FNS-cloud consortium at the time of deliverable preparation. Datasets that are either used or will be generated during the project in WP4 and WP5 were first mapped onto their respective data domains and data types (see Table 4 in Appendix 1) and their accessibility/availability status determined. This helped identify data types of which several datasets are available and accessible, as well as determine potential research questions that can be answered using the available datasets.

Based on the results of this preliminary, two research questions were selected to demonstrate how applicability criteria can be developed to formalize the process of (1) checking submitted data against standard, agreed upon formats and (2) applying, where necessary, processing algorithms to harmonize the data and enable (semi-)automatic data integration and consolidation.

The two case studies and results of their respective dataset analysis are presented in the following sections. Not all datasets used for the case studies were publicly available, which inhibited the use of any specific data examples (i.e. actual values) in this deliverable. The analysis results will be therefore presented on a conceptual level with no values presented.

4.2.1 Single-domain question

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

The first selected research question relates to pre-packaged food products and their labelling information (e.g. ingredients, nutritional declarations, allergen information or dietary claims). The following research question was addressed:

How does the average total sugar content of selected branded food products within specific food product groups differ among European countries?

To answer this question, branded food data from the 'Agrifood' domain is necessary. **Using three branded food datasets from Slovenia (SI), the Netherlands (NL) and Switzerland (CH)**, we focused on those parameters necessary to answer the research question and identified challenges in data integration and co-analysis. Figure 4 summarizes characteristics of the first, single-domain case study.

In a 'real-life' use case, NLP tools can be used to break down components of the research question and map it onto associated data domain(s) and data types (i.e. identify data requirements). Datasets within the selected domain and of the mapped data type are then extracted/ retrieved from the FNS-Cloud for the user to choose the datasets he/ she can use to answer their respective research question (see Figure 4).

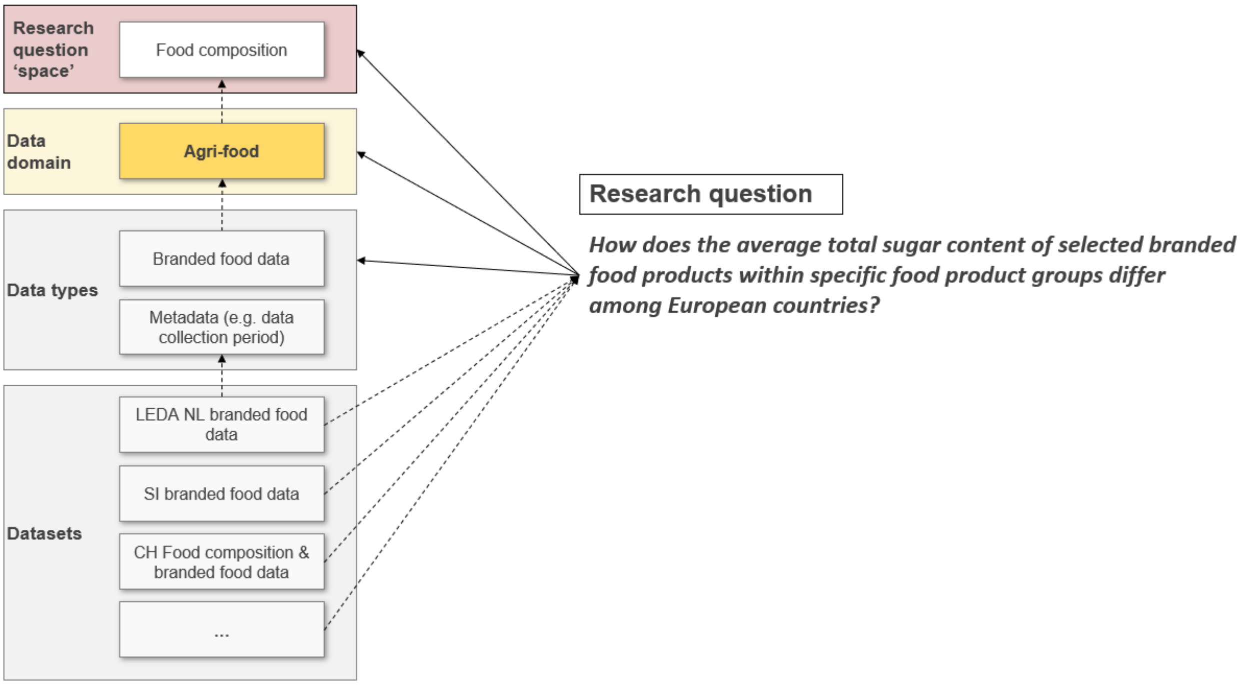

Research question 'space'
Food composition
Data domain
Agri-food
Data types
Branded food data
Metadata (e.g. data collection period)
Datasets
LEDA NL branded food data
SI branded food data
CH Food composition & branded food data
...
Research question
How does the average total sugar content of selected branded food products within specific food product groups differ among European countries?

Figure 4: Characteristics of the first case study related to a single-domain question. Datasets within the FNS-Cloud are mapped to their respective data domain and type. Question component analysis helps map questions to question spaces, data domains and types. Available datasets can thus be retrieved for analysis in response to a user query.

The Slovenian branded food data originates from the 'Composition and Labelling Information System' (CLAS), created and managed by the Nutrition Institute in Ljubljana, Slovenia (Nutrition Institute, 2018). A CLAS application was created in 2015 to enable the collection of photographs of food labelling, incorporation of data into the CLAS database, and digital recognition of EAN codes to speed up the database formation and to avoid duplicate entries. Information on a total of 10,674 unique food items were collected, including the product name, company, brand, list of ingredients, nutritional values, data on allergens and health claims, packaging volume, price, and EAN barcode (Korošec & Pravst, 2014; Zupanič et al., 2018). An updated version of the Slovenian branded food data will be made publicly available via the FNS-Cloud project.

The Dutch branded food data comes from the Dutch Food Label Database (LEDA), created in 2007 as commissioned by the Ministry of Health Welfare and Sport and maintained by the Netherlands Nutrition

Centre (NNC) and the National Institute for Public Health and the Environment (RIVM). This database is intended to be used for consumer information by the NNC and for health policy supporting research by RIVM. It is thus not made publicly available. The following 5 data providers provided label data to the NNC using APIs: GS1 NL, Brandbank NL, PS in Foodservice, Albert Heijn (the largest supermarket in NL) and SIM (umbrella organization for a large group of supermarkets in NL). Private label (= supermarket brands) are covered by Albert Heijn and SIM, manufacturers brand (e.g. Kellogg's, Coca cola) are covered by GS1 and Brandbank. PS in Foodservice covers catering, hospitality etc. The LEDA currently contains around 100.000 foods and covers around 75% of the food supply in supermarkets in NL (Westenbrink, 2020). Data have been mainly uploaded between 2011 and 2017 using application programming interfaces (APIs), while data collection and manual data entry on a smaller scale already started in 2007.

The Swiss branded food data was part of the Swiss Food Composition Database (Federal Food Safety and Veterinary Office, 2020) and was recently removed to have only generic foods in the database. It contained information on the composition of approximately 10,500 foods available in Switzerland that have been classified into 19 main and 105 sub categories. The data, which are continuously updated and extended, are available in four languages (English, German, French, and Italian) and accessible free of charge to all interested parties. A subset of the food data comes from companies, i.e. constitutes branded food data.

To analyse the data in more detail and take a closer look at the relevant parameters requiring harmonization, two food product groups were selected as examples: breakfast cereals and soft drinks. These two groups constitute common food product categories consumed in almost every food culture in Europe, and both are/have been targets for sugar reformulation in several member states. Soft drinks in particular are also known to be a major source of sugar in people's diets (Amoutzopoulos et al., 2020; Ruiz et al., 2017; Zupanič et al., 2018). Data focused on these two food product groups were extracted from CLAS (SI) and LEDA (NL), whereas the full Swiss food database was provided by the partner (V5.3).

4.2.2 Cross-domain question

The following cross-domain research question was selected for the second illustrative case study:

How high is an individual's risk of developing type-2-diabetes (T2D) given his or her personal diet, lifestyle, sociodemographic and anthropometric characteristics?

Research questions targeted at identifying links between nutrition and disease outcomes continue to be at the heart of investigations within the FNS field. As a major cardio-metabolic risk factor, the importance of early detection of T2D and prevention (via e.g. healthier diets) is well-documented (Ekoe et al., 2018; Gilmer & O'Connor, 2010). There are many risk scores and indices in the literature, which attempt to predict the chances of developing T2D via the incorporation of various risk factors such as age, family history, sex, body mass index (BMI), medication, etc.; however only some of them integrate nutrition and diet as an explanatory factor (Gray et al., 2010; Hippisley-Cox & Coupland, 2017; Kanellakis et al., 2020; Lindström & Tuomilehto, 2003; Schmidt et al., 2005).

For the purpose of this specific research question within the FNS-Cloud project, two available datasets from the **'Feel4Diabetes'** (Manios et al., 2018) and the **'Food4me'** (Celis-Morales et al., 2017) study, where not only dietary intake but also anthropometric, sociodemographic, lifestyle and biomarker data was collected, will be used and jointly analysed to develop a new model (index) which helps predict the risk of T2D, and provide personalised feedback for change (WP5, Demonstrator-DE03). Characteristics of the second case study are summarized in Figure 5.

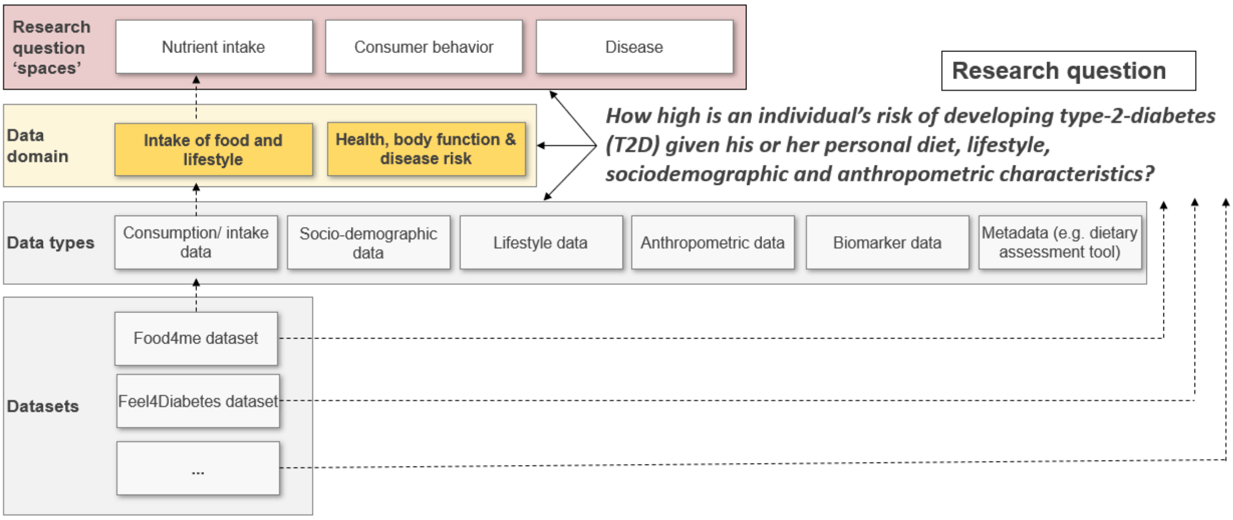

Research question 'spaces'
Nutrient intake
Consumer behavior
Disease
Data domain
Intake of food and lifestyle
Health, body function & disease risk
Data types
Consumption/ intake data
Socio-demographic data
Lifestyle data
Anthropometric data
Biomarker data
Metadata (e.g. dietary assessment tool)
Datasets
Food4me dataset
Feel4Diabetes dataset
...
Research question
How high is an individual's risk of developing type-2-diabetes (T2D) given his or her personal diet, lifestyle, sociodemographic and anthropometric characteristics?

Figure 5: Characteristics of the second case study related to a cross-domain question. Datasets within the FNS-Cloud are mapped to their respective data domain and type. Question component analysis helps map questions to question spaces, data domains and types. Available datasets can thus be retrieved for analysis in response to a user query.

The Feel4Diabetes study was an EU-funded project with a duration of 5 years (2015-2019). Its aim was to develop, implement and evaluate a school- and community-based intervention to promote healthy lifestyle and tackle obesity and obesity-related metabolic risk factors for the prevention of T2D among families from low- and middle-income countries and vulnerable populations in high-income countries in Europe (http://feel4diabetes- study.eu). The intervention was applied in six European countries (Belgium, Bulgaria, Finland, Greece, Hungary and Spain) and targeted dietary and physical activity behaviours. To evaluate the impact of the intervention, behavioural and lifestyle indices on drinking, eating, physical activity and sedentary behaviours as well as their determinants were self-reported by the participating families in standardized questionnaires developed specifically for the Feel4Diabetes study (Anastasiou et al., 2020). A detailed description of the methodology of the Feel4Diabetes-study can be found in Manios et al. (2018). Variables collected in two stages from study subjects (ca. 20,500 parents in stage 1 and ca. 3,150 parents in stage two) included the following:

- **Sociodemographic data**: e.g. sex, age, educational level, marital status
- **Behavioural indices** regarding dietary habits, physical activity and sedentary behaviours: e.g. portions of sugary drinks per week, number of meals and snacks during a day, minutes of daily vigorous physical activity, time spent in front of computers and television, etc.
- **Anthropometry:** height, weight and waist circumference
- **Health biomarker data**: blood pressure and fasting plasma glucose (FPG)

Not all parameters were collected for all study subjects in both stages.

The data sample used here for the illustrative case study contains a sample (two participants) of the Feel4Diabetes data collected in Greece.

The Food4me study was also an EU-funded project from 2011-2015 with the aim to conduct a multi-centre, web-based, proof-of-principle study of personalised nutrition (PN) to determine whether providing more personalised dietary advice (and at a higher or lower frequency) leads to greater improvements in eating patterns and health outcomes compared to conventional population-based advice. The intervention took place across seven European countries (Germany, Greece, Ireland, the Netherlands, Poland, Spain and UK). Participants were randomly assigned to one of the following intervention groups for a 6-month period: Level 0—control group—receiving conventional, non-PN advice; Level 1—receiving PN advice based on dietary intake data alone; Level 2—receiving PN advice based on dietary intake and phenotypic data; and Level 3—receiving PN advice based on dietary intake, phenotypic and genotypic data. Details of the study methodology are described elsewhere (Celis-Morales et al., 2015). To compare the effect of different levels or types of PN advice, dietary intake, phenotypic and genotypic data were collected on study participants (> 1,500 participants). Similar to the Feel4Diabetes study, this included:

- **Lifestyle indices** related to dietary habits (food frequency questionnaire and physical activity.
- **Phenotypic data**: body weight, body mass index (BMI) and waist circumference.
- **Blood-based biomarkers**: blood glucose, total cholesterol, vitamins, carotenoids and fatty acids.
- **Genetic data**: information on genetic loci that are linked to specific dietary or phenotypic outcomes (e.g. fat intake).

Integrating the data (on common parameters) from both studies would allow a bigger sample size and including more ethnicities and various socioeconomic backgrounds, which would help improve model accurateness, statistical validity and generalizability. In WP5 (Demonstrator-DE03) a predictive model for cardio-metabolic risk based on these two datasets will be developed. Parameters, which are common between both datasets (i.e. were gathered in both studies) and will go into the model, will have to be harmonized to enable data integration/ merging and joint data analysis.

4.3 Generic (domain-independent) data criteria

Generic (domain-independent) data criteria relate primarily to general study metadata but also to food identification and description (see Section 3.2). They equally apply to all identified FNS data domains and represent a set of minimal requirements that should be fulfilled by a dataset (or otherwise achieved with the help of advanced IT technologies) in order to integrate the data into the envisioned Cloud infrastructure and achieve data interoperability.

As mentioned in Section 3.2, every single parameter will require a nomenclature criterion (i.e. standard terminology/ annotation) and one or more data value criteria (see Figure 2). To take an example of a generic (domain-independent) criterion, 'title of dataset' is an important general metadata parameter. Unless users can directly enter this metadata information into a predefined submission form on the Cloud, the descriptor name used within a given dataset for its title will have to match an existing concept within the FNS-Cloud internal ontology or be linked to it (nomenclature criterion). For instance, if the dataset title is described as 'dataset name' this may easily be mapped to the FNS ontology concept 'title'. Checking submitted data against the FNS-Cloud ontology and determining whether data processing is necessary to transform incoming data into standard terminology or whether data cannot be integrated automatically at all, represents a standard procedure for all nomenclature criteria, regardless which parameter is addressed and whether it relates to generic metadata, food identification and description, study methods, study population/ sample or any other data type-specific measure. Therefore, a generic decision tree, which formally represents the applicability check and harmonization processes for all nomenclature criteria, is presented in Figure 6. Instead of 'dataset title', any other parameter included in a dataset will have to match standard vocabulary (FNS-Cloud ontology) or be otherwise transformed to be integrated (taken up) into the FNS-Cloud infrastructure. The decision-making process for checking nomenclature will always follow the procedure highlighted in Figure 6.

Two modes of action have been differentiated: FNS-Cloud development mode and static mode. During the construction process of the FNS Cloud infrastructure, as many foreign concepts as possible can be identified from available FNS datasets and integrated into an internal FNS-Cloud ontology (FNS-Cloud development mode). This activity may be extended well beyond the project end, as ontologies benefit from an active community contributing to its continuous growth. However, at any given point in time, a new dataset may include concepts which are not yet defined within (and cannot be matched to) the FNS-Cloud ontology (static mode). Such parameters will have to be directed to manual integration or rejection (see Figure 6). To ensure as many parameters in a dataset as possible are machine-readable and can be incorporated into the FNS-Cloud, the FNS-Cloud developers shall strive towards as comprehensive an ontology as possible, borrowing concepts from various general and domain-specific ontologies.

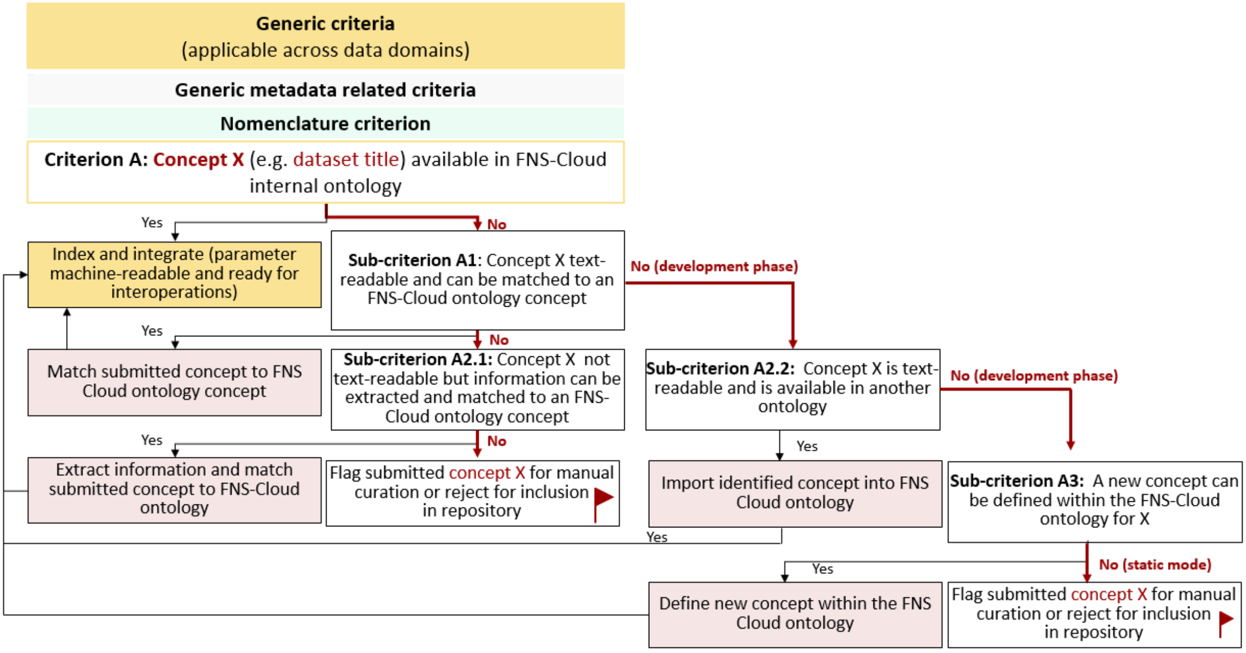

Generic criteria
(applicable across data domains)
Generic metadata related criteria
Nomenclature criterion
Criterion A: Concept X (e.g. dataset title) available in FNS-Cloud internal ontology
Yes
No
Index and integrate (parameter machine-readable and ready for interoperations)
Sub-criterion A1: Concept X text-readable and can be matched to an FNS-Cloud ontology concept
No (development phase)
Yes
No
Match submitted concept to FNS Cloud ontology concept
Sub-criterion A2.1: Concept X not text-readable but information can be extracted and matched to an FNS-Cloud ontology concept
Sub-criterion A2.2: Concept X is text-readable and is available in another ontology
No (development phase)
Yes
No
Yes
Extract information and match submitted concept to FNS-Cloud ontology
Flag submitted concept X for manual curation or reject for inclusion in repository
Import identified concept into FNS Cloud ontology
Sub-criterion A3: A new concept can be defined within the FNS-Cloud ontology for X
Yes
Yes
No (static mode)
Define new concept within the FNS Cloud ontology
Flag submitted concept X for manual curation or reject for inclusion in repository

Figure 6: Generic decision tree for all nomenclature criteria

Next to nomenclature criteria, for each parameter requiring harmonization, data value criteria will be necessary. To take the example from above, the ‘title of dataset’ may be stipulated to have the data field type ‘string’ (text). If it is a number instead or a date or an image, it cannot be automatically incorporated into the Cloud. Some parameters will not only require data value criteria to cross-check and harmonize their data field type to the standard format, but also their variable type and the categories (or values) they may assume. For example, another generic metadata parameter is ‘data sharing policy’. It may be specified as a categorical variable within the FNS-Cloud (variable type) and follow strict multiple-choice options (e.g. publicly accessible, available upon request, not publicly accessible).

Data value criteria thus help check submitted values against a pre-defined (e.g. agreed upon between work package beneficiaries) reference value format. If the data does not match the reference format, a process to check whether it matches any other existing standard format, or whether it is machine-readable and can be transformed into the reference format, must take place. If there is no way to match a given parameter value to the standard format in the background and thus integrate it into the FNS-Cloud, the value must be flagged for manual curation or rejected. Again, each step of this integration or rejection process represents an applicability criterion: Parameter X shall match [standard reference]. If not, what next? Figure 7, Figure 8 and Figure 9 show examples for generic (domain-independent) criteria that address the harmonization of the values assumed by selected parameters.

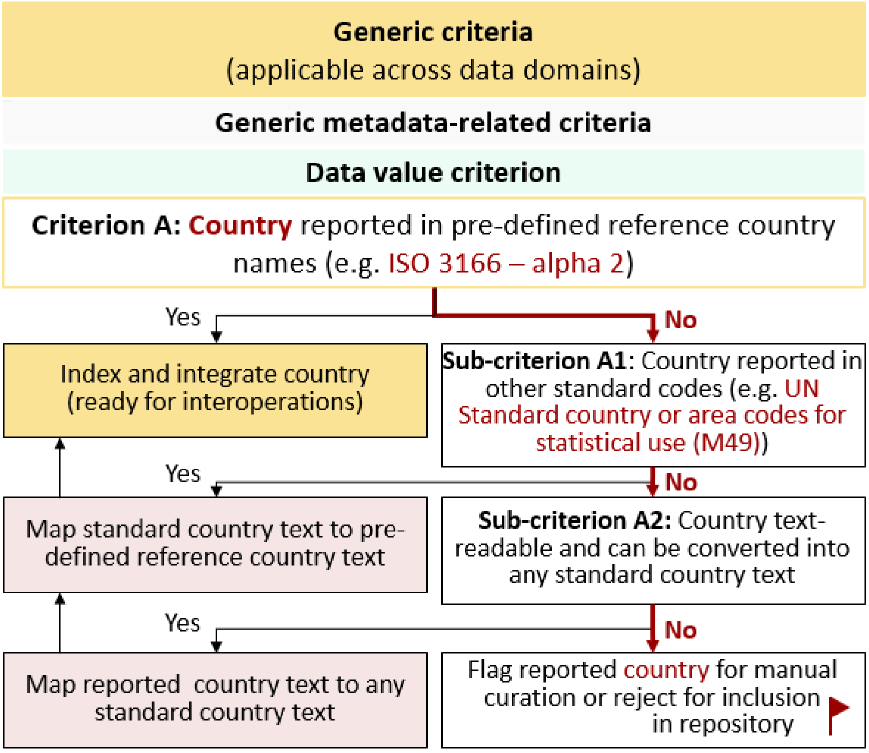

Generic criteria
(applicable across data domains)
Generic metadata-related criteria
Data value criterion
Criterion A: Country reported in pre-defined reference country names (e.g. ISO 3166 – alpha 2)
Yes
No
Index and integrate country (ready for interoperations)
Sub-criterion A1: Country reported in other standard codes (e.g. UN Standard country or area codes for statistical use (M49))
Yes
No
Map standard country text to pre-defined reference country text
Sub-criterion A2: Country text-readable and can be converted into any standard country text
Yes
No
Map reported country text to any standard country text
Flag reported country for manual curation or reject for inclusion in repository

Figure 7: Applicability criteria for the parameter 'country', which are generic (domain-independent) and pertain to the data value

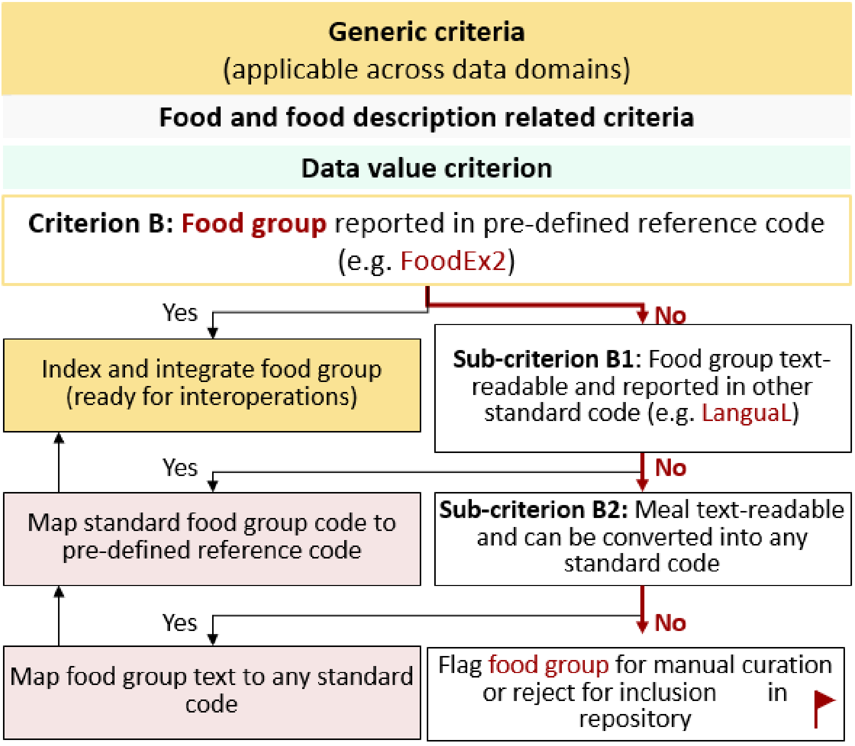

Generic criteria
(applicable across data domains)
Food and food description related criteria
Data value criterion
Criterion B: Food group reported in pre-defined reference code (e.g. FoodEx2)
Yes
No
Index and integrate food group (ready for interoperations)
Sub-criterion B1: Food group text-readable and reported in other standard code (e.g. Langual)
Yes
No
Map standard food group code to pre-defined reference code
Sub-criterion B2: Meal text-readable and can be converted into any standard code
Yes
No
Map food group text to any standard code
Flag food group for manual curation or reject for inclusion in repository

Figure 8: Applicability criteria for the parameter 'food group', which are generic (domain-independent) and pertain to the data value

The 'food group' constitutes a particularly important parameter for the cross-domain study, as certain food groups (e.g. vegetables, fruits, legumes, berries and sugary drinks) were found to be linked with a risk of T2D (Kanellakis et al., 2020; Lindström & Tuomilehto, 2003) and their mean daily intake (in e.g.

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

g/day) represents a candidate parameter to include in the model for linking diet with T2D. However, the FFQ applied in the Feel4Diabetes vs. the Food4me study used different food group names. The FFQ of the Food4me study included both food groups (e.g. cream crackers, cheese biscuits, rusks) and food items (e.g. white rice), whereas in the Feel4Diabetes FFQ only the consumption of certain food groups was asked. A food coding system like LanguaL or FoodEx2 was not used in either of the studies, which hinders data merging and interoperability. Translation of food items/groups into the groups of interest (vegetables, fruits or e.g. legumes) is one of the challenges, which the FNS Cloud project aims to address and offer guidance for. Possible starting points are manual matching (which is very time consuming) or using nascent machine learning algorithms (which allows for automated matching workflows) (Eftimov et al., 2017; Koroušić Seljak et al., 2018).

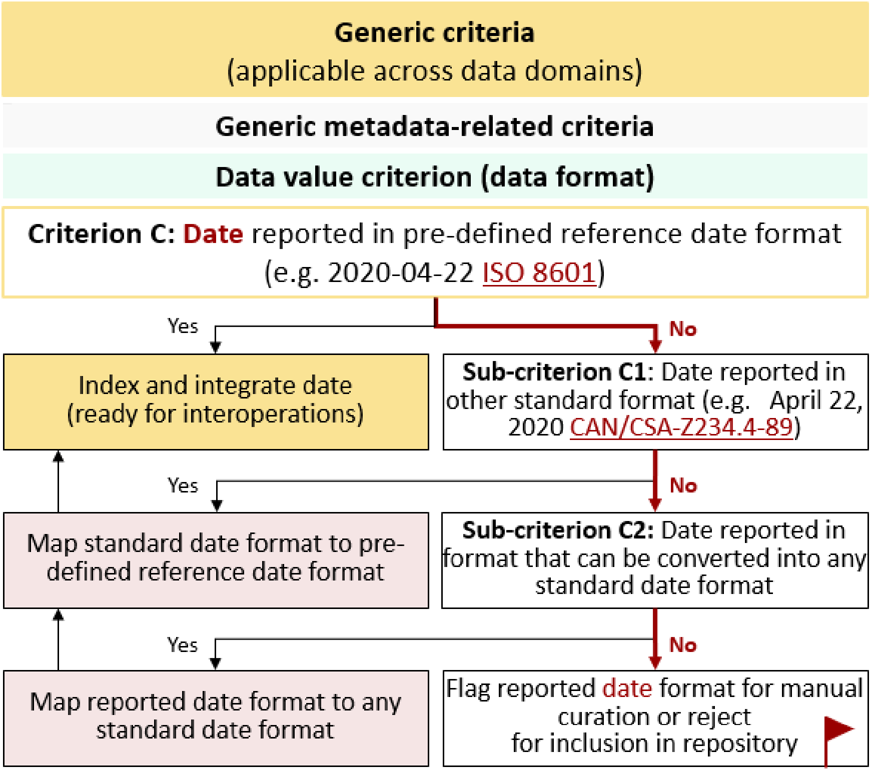

Generic criteria
(applicable across data domains)
Generic metadata-related criteria
Data value criterion (data format)
Criterion C: Date reported in pre-defined reference date format
(e.g. 2020-04-22 ISO 8601)
Yes
No
Index and integrate date
(ready for interoperations)
Sub-criterion C1: Date reported in other standard format (e.g. April 22, 2020 CAN/CSA-Z234.4-89)
Yes
No
Map standard date format to pre-defined reference date format
Sub-criterion C2: Date reported in format that can be converted into any standard date format
Yes
No
Map reported date format to any standard date format
Flag reported date format for manual curation or reject
for inclusion in repository

Figure 9: Applicability criteria for the parameter 'date', which are generic (domain-independent) and pertain to the data value

Applicability criteria for the parameter 'date' may also apply to several other (data type-specific) parameters. For instance, the date a food dietary survey was filled out would also have to be in the same reference format (or be convertible into the reference format), if this information is to be taken up into the FNS-Cloud.

4.4 Additional, domain-specific data criteria

While generic criteria apply to the various data relevant in both single-domain and cross-domain questions, domain-specific criteria only apply to data from the respective domain(s) involved in a given single- or cross-domain question. Domain-specific criteria aim to make datasets from a given domain interoperable among themselves as well as with datasets from other domains, which may include common parameters. In the following, we outline how these criteria are defined and applied for our two example questions.

4.4.1 Single-domain question: Criteria for the 'agrifood' domain

To determine the average total sugar content (i.e. added and naturally occurring sugars) of selected food products within specific food product groups (here: cereals and soft drinks) across different European countries, a list of parameters, which are required to answer the research question, was first determined. These parameters are:

- Food name: food item (product) name and identification
- Food product group
- Sugar, total (nutrient name or nomenclature)
- Unit
- Matrix unit

The above parameters require harmonization and standardization to enable data integration, interoperability and co-analysis of all three branded food datasets (SI, CH and NL).

The food product group first allows selecting all food products within the chosen categories (soft drinks and breakfast cereals). While this was a fairly easy task in the SI and NL datasets, as the data related to these two specific categories were extracted and made available by partners, the case with the CH dataset is not as straightforward. Food product groups in the Swiss dataset represent bigger 'clusters' of products (e.g. breakfast cereals, bread and bread products). This means that allocating products to the group 'breakfast cereals' requires a data pre-processing step with regards to the Swiss data to filter products belonging to breakfast cereals specifically.

Again, the situation is fairly easy (even with the necessary data pre-processing step) because the products 'breakfast cereals' and 'soft drinks' are quite standard categories. However, given that the three datasets use their own food product categorization (e.g. national food classification systems), other product groups may not actually mean the same thing (represent the same category of food) across the board. For example, 'sausages and cold meats/Boiled sausage products' will not be equal to the food product category 'processed meats' in another data set. Using a standardized food product classification system (e.g. Brick codes from the Global Product Classification system) or transforming all data to match a selected classification system (e.g. LanguaL (Møller & Ireland, 2017) or FoodEx2 (EFSA, 2015)) is the only way to ensure interoperability of different branded food datasets. Applicability criteria addressing the food product group as an example of a domain-specific data criterion are illustrated in Figure 10.

As a cross-check (validation) for the selection of food products of a given category, one could check how many of the selected products are equivalent in the different markets. It would therefore be beneficial to match selected products by their Global Trade Item Number (GTIN). The NL and SI datasets include the GTIN, while the Swiss dataset doesn't. Of course, whenever the GTIN is available matching of same products is a straightforward task. However, as exemplified by the Swiss dataset, it is not always included. Additionally, some products are quite similar (almost identical) in different markets, yet have diverging GTIN/ EAN numbers. Matching food products in different countries via their name, ingredients and other product label information may present an option on how to tackle this task, yet it will require data processing algorithms to do the matching. However, the Swiss food dataset does not include ingredients, which again complicates this matching task.

Next to harmonized food product groups, nutrient names, as well as nutrient values (data field type, unit and matrix unit) need to be harmonized. For the nutrient name (i.e. 'sugar'), nomenclature criteria again follow the general pattern depicted in Figure 6. The concept 'sugar' is naturally part of several ontologies and thesauri (e.g. FoodOn (Dooley et al., 2018), EuroFIR component thesaurus (Becker et al., 2008)) and will inevitably be part of the FNS-Cloud ontology. However, the units used for indicating the nutrient content can differ among different datasets. Therefore, next to a unit criterion (unit shall be an SI unit or other standardized/ ontology-based domain-specific unit), a unit for reporting nutrient content (e.g.

g/100g or 100 mL) shall be agreed upon. Non-matching units will have to be transformed and values converted (see Figure 11).

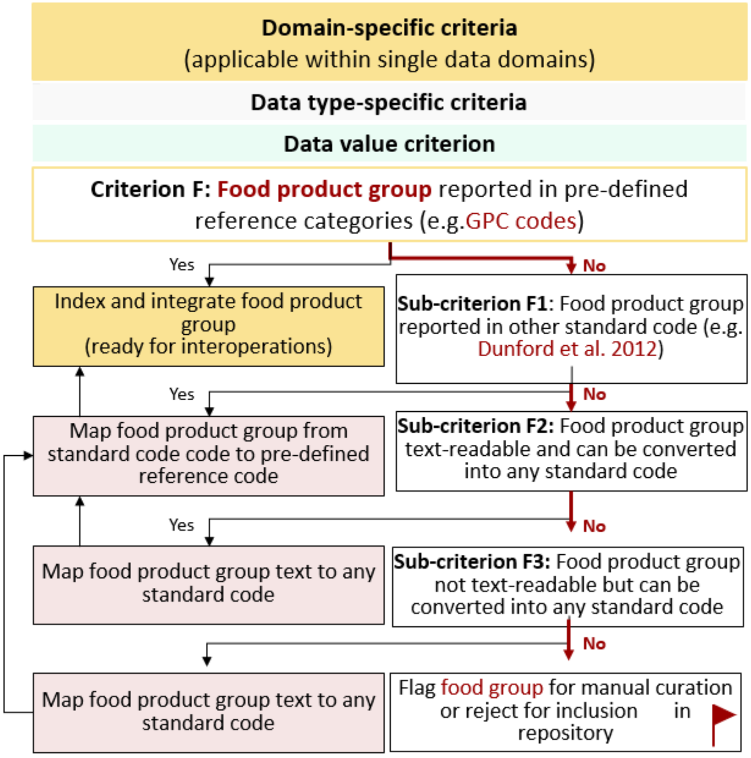

Domain-specific criteria
(applicable within single data domains)
Data type-specific criteria
Data value criterion
Criterion F: Food product group reported in pre-defined reference categories (e.g.GPC codes)
Yes
No
Index and integrate food product group (ready for interoperations)
Sub-criterion F1: Food product group reported in other standard code (e.g. Dunford et al. 2012)
Yes
No
Map food product group from standard code code to pre-defined reference code
Sub-criterion F2: Food product group text-readable and can be converted into any standard code
Yes
No
Map food product group text to any standard code
Sub-criterion F3: Food product group not text-readable but can be converted into any standard code
No
Map food product group text to any standard code
Flag food group for manual curation or reject for inclusion in repository

Figure 10: Applicability criteria for the parameter 'food product group', which are domain-specific and pertain to the data value

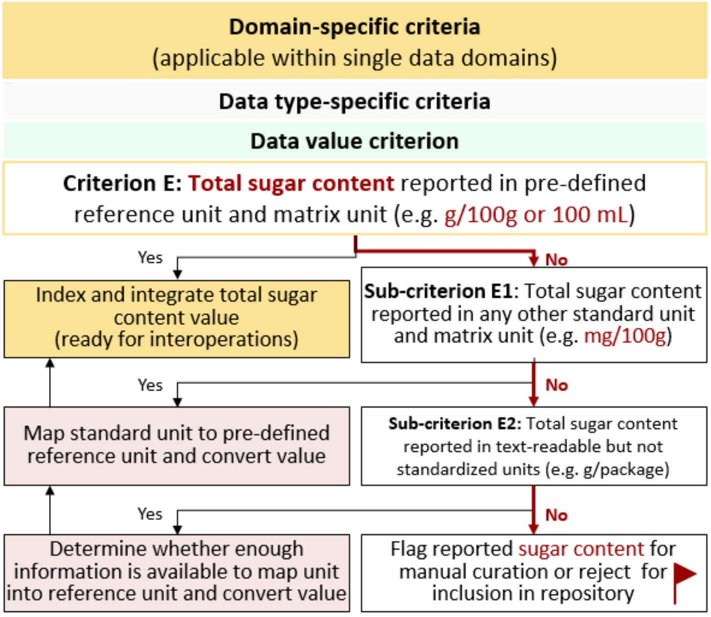

Domain-specific criteria
(applicable within single data domains)
Data type-specific criteria
Data value criterion
Criterion E: Total sugar content reported in pre-defined reference unit and matrix unit (e.g. g/100g or 100 mL)
Yes
No
Index and integrate total sugar content value (ready for interoperations)
Sub-criterion E1: Total sugar content reported in any other standard unit and matrix unit (e.g. mg/100g)
Yes
No
Map standard unit to pre-defined reference unit and convert value
Sub-criterion E2: Total sugar content reported in text-readable but not standardized units (e.g. g/package)
Yes
No
Determine whether enough information is available to map unit into reference unit and convert value
Flag reported sugar content for manual curation or reject for inclusion in repository

Figure 11: Applicability criteria for the unit and matrix unit, which are domain-specific and pertain to the data value

4.4.2 Cross-domain question: Criteria for the 'food intake and lifestyle' & 'health, body function and disease risk' domains

The second selected research question related to the link between diet and diabetes risk. At this point in time, the exact parameters (variables) from each dataset that will finally be included as explanatory variables in the predictive model to estimate an individual's diabetes risk (DE03) are not known. However, previous studies, which have attempted to establish a link between lifestyle and sociodemographic indices and cardio-metabolic risk factors (e.g. Gray et al., 2010; Kanellakis et al., 2020), already provide a hint as to which parameters will have explanatory power. These studies were therefore used as a basis to identify the relevant parameters within the available datasets for which applicability criteria were necessary.

Parameters which have shown statistical significance in constructing the European Insulin Resistance Risk (Kanellakis et al., 2020) as well as the FINDISC T2D risk assessment tool (Lindström & Tuomilehto, 2003), i.e. parameters identified as relevant in both publications, were used as a basis. These were:

- Age
- BMI (calculated from weight and height)
- Waist circumference (WC)
- Sex
- An indicator of physical activity (e.g. walking or vigorous physical activity)
- An indicator of certain food/food group intakes (mean daily intake of vegetables, fruits and berries)
- An indicator of certain drinks consumption (daily consumption of sugary beverages)

For the dependent variable in the model, health biomarker data is necessary. Parameters collected in both studies and of relevance for the diagnosis of T2D are:

- Fasting blood glucose (FPG) and
- Total cholesterol
- Blood pressure

Applicability criteria are now necessary to harmonize the data on these parameters from the two datasets in order to build the model and answer the research question. Age, BMI and waist circumference are quite standardized parameters. They were reported in the same way in both datasets (in years, kg/m^2 and cm). Integrating food intake of certain foods and drinks requires primarily a harmonized food classification system to be applied in the different datasets. However, the FFQs used in each study differ in the way they classify foods into food groups in the collection phase, and cluster food groups in their analysis and hence there may be some difficulty in matching intakes of specific food/food groups that will have to be resolved (following a similar decision making process as in Figure 8). Additionally, units for mean daily intake of certain foods (e.g. g/day) as well as for FPG and total cholesterol will have to be harmonized. The applicability criteria formalizing the harmonization process of these parameters' units can be designed to mirror the applicability criteria for 'unit of sugar content' shown in Figure 11.

Furthermore, categories for daily consumption of drinks will also have to be defined (e.g. 1 cup a day, 2 cups a day or <250 mL a day, 250-500 mL a day, >500 mL a day). Which categories are to be used for any categorical variable will have to be pre-defined (e.g. by data providers or from guidelines and standards). For example, for education-related variables categories can be defined using UNESCO's International Standard Classification of Education (UNESCO-UIS, 2012), while the 'current occupation' variable can be developed using the International Labour Organization's International Standard Classification of Occupations (ILO, 2012).

The same goes for the variables gender or sex and physical activity (PA). To demonstrate examples of applicability criteria for categorical variables (data value criteria) Figure 12 and Figure 13Figure 12 illustrate decision trees for gender or sex and PA.

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

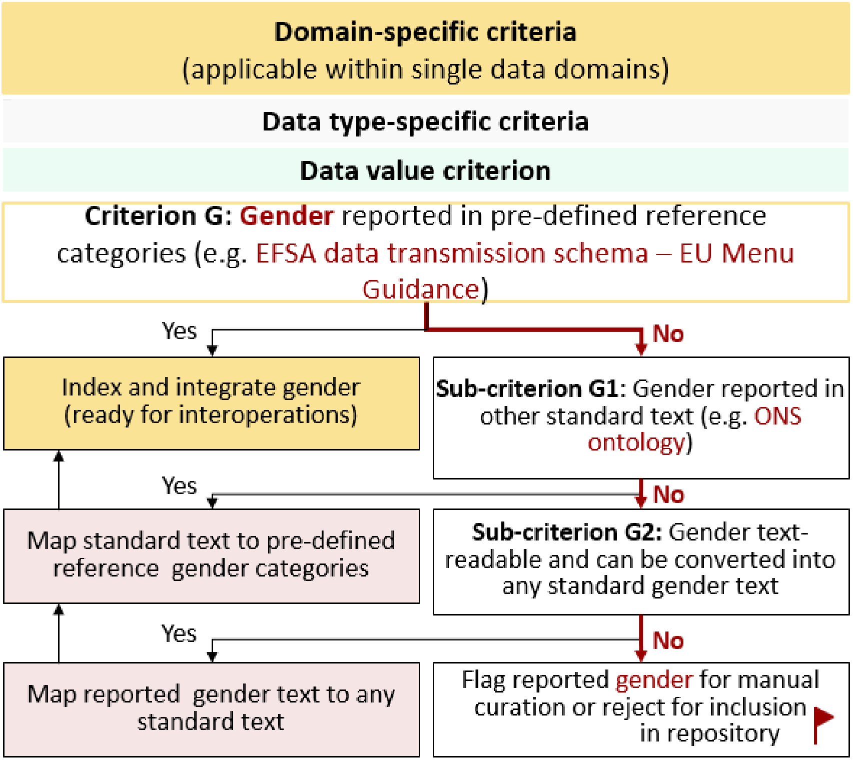

Domain-specific criteria
(applicable within single data domains)
Data type-specific criteria
Data value criterion
Criterion G: Gender reported in pre-defined reference categories (e.g. EFSA data transmission schema – EU Menu Guidance)
Yes
No
Index and integrate gender (ready for interoperations)
Sub-criterion G1: Gender reported in other standard text (e.g. ONS ontology)
Yes
No
Map standard text to pre-defined reference gender categories
Sub-criterion G2: Gender text-readable and can be converted into any standard gender text
Yes
No
Map reported gender text to any standard text
Flag reported gender for manual curation or reject for inclusion in repository

Figure 12: Applicability criteria for the parameter 'gender' (or 'sex'), which are domain-specific and pertain to the data value

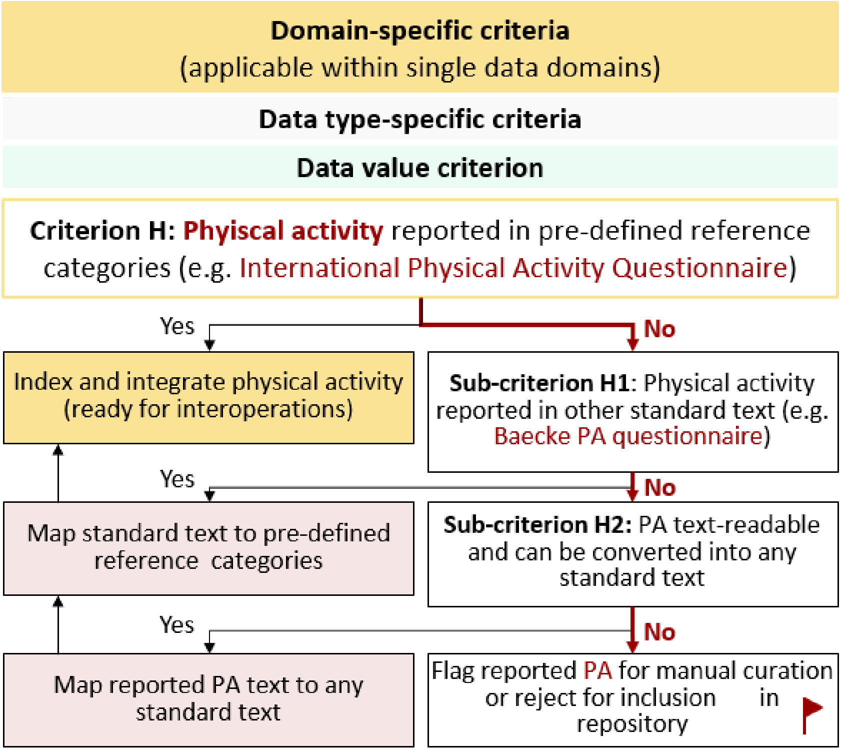

Domain-specific criteria
(applicable within single data domains)
Data type-specific criteria
Data value criterion
Criterion H: Phyiscal activity reported in pre-defined reference categories (e.g. International Physical Activity Questionnaire)
Yes
No
Index and integrate physical activity (ready for interoperations)
Sub-criterion H1: Physical activity reported in other standard text (e.g. Baecke PA questionnaire)
Yes
No
Map standard text to pre-defined reference categories
Sub-criterion H2: PA text-readable and can be converted into any standard text
Yes
No
Map reported PA text to any standard text
Flag reported PA for manual curation or reject for inclusion in repository

Figure 13: Applicability criteria for the parameter 'physical activity' , which are domain-specific and pertain to the data value

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

Like food intake, alignment of measurements of PA are also challenging. There is no commonly agreed best practice method to assess physical activity. In Feel4Diabetes the hours per day and days per week of vigorous physical activity, moderate activity and walking was recorded, using a questionnaire developed for this study. In Food4me, the Baecke questionnaire for PA (Baecke et al., 1982) was used, which classifies sport activity into low, medium and high intensity and then asks for the hours per week and months per year. Can vigorous, moderate and walking be translated 1:1 into low, medium and high intensity activity, using e.g. Metabolic Equivalent Tasks (METs)? While this example may still enable data integration, if we assume that the three categories from one questionnaire are equivalent to the categories of the other (yet integration rather by a human expert and not automated), other datasets which categorize PA in e.g. a Boolean way (more or less than e.g. 30 minutes a day of PA' (Yes/No)) would make such merging impossible.

5. Recommendations for different stakeholder groups (target audience)

5.1 FNS-Cloud platform developers

As primary target group of D3.1, the FNS-Cloud platform developers will rely on several advanced IT solutions to fulfil the developed applicability criteria, render submitted data interoperable and integrate datasets in a (semi-)automated way into the FNS-Cloud.

Based on the analysis of FNS data(sets) available through the project consortium, as well as the identified data inconsistencies and harmonization needs and the two performed case studies, a list of recommendations for FNS-Cloud developers was abstracted. These recommendations represent a kind of 'action list', as they include necessary steps towards data interoperability and establishing the envisioned FNS-Cloud infrastructure.

Our recommendations for FNS-Cloud developers are:

- All FNS data(sets) to be uploaded on the Cloud shall be linked to their respective data types (e.g. 'tagged') and mapped onto FNS data domains and FNS-relevant research questions or research question spaces (i.e. themes or research interests). Linking or assigning datasets to their respective data types can be done manually by asking data providers when uploading their data to select suitable data types from a pre-defined list of FNS-relevant data types. Alternatively, it can be done 'in the background' in an automated way by analysing dataset title, content ... etc. and assigning a suitable data type and ultimately data domain. Additionally, datasets can be tagged with several ontology-based keywords (again, either selected manually by the data provider or assigned automatically) which will facilitate achieving a conceptual connection between users' questions and available machine-readable information.
- **A set of minimum requirements for metadata descriptors** (e.g. information on study design, data ownership, study measurements, ethical issues) **shall be developed for the FNS-Cloud** or be adapted from previous efforts, such as ENPADASI, to enable data integration and interoperability, as well as efficient data querying, interpretation, comparability, (re)use and repurposing. Accurate representation of the content of available information sources is key to making data findable, interoperable and reusable.

  It is advisable to create a pre-defined data submission form where data owners fill in such metadata in an already standardized/ harmonized way. **It is further recommended to specify minimum metadata requirements for each individual FNS-relevant data type or group of closely connected data types** (e.g. for food composition data, health biomarker data or omics data), as several metadata descriptors – and the values they assume – are specific to a certain data type or research field (see for instance Pinart et al., 2018 for minimum requirements for nutritional studies). In developing such data type-specific metadata requirements, the ISA-Tab format (http://isa-tab.sf.net) offers a promising framework. Already existing domain-specific or data type-specific minimal reporting requirements and 'checklists' shall be adopted (for examples of such reporting requirements or checklists see Section 5.2).

- The process of developing applicability criteria demonstrated on the two illustrative case studies in Chapter 4 shall be repeated for multiple use cases:

  1. Using a set of example FNS-relevant research questions (e.g. as defined in WP4 Use cases and WP5 demonstrators, see Table 4, Appendix 1) **a list of variables, which are required for answering these research questions, must be defined**. Data providers within the consortium may also be asked to agree on a list of key variables (variable catalogue), required to answer research questions of current scientific relevance in their respective research domain, which will then be targeted for applicability criteria and harmonization ('target variables').
  2. Reference or standard formats for reporting values of the target variables shall be decided upon in agreement with data providers and other experts within the consortium. Ontologies, which provide a community vetted language, shall also be selected or developed for standardized annotation of all selected variables, thereby fulfilling all nomenclature criteria. It is necessary to arrive at a consensus on relevant variables (e.g. weight), their 'meaning' and definitions (e.g. measured weight), and format (weight in kg) (Doiron et al., 2013).
     International and commonly used guidelines and data standards (see Table 1) can be used to decide on a reference or standard format for any of the target variables. **Where applicable, recommended ontologies, controlled vocabularies and data exchange models shall be adopted from D2.1** (Presser et al., 2020).

  With the help of processing algorithms, submitted FNS data(sets) collected by different researchers can be checked against the selected reference or standard formats and transformed into the FNS-Cloud standard format where necessary.

- An FNS-Cloud internal ontology, covering all concepts related to FNS data domains, FNS metadata (study descriptors) as well as FNS target variables (i.e. variables required to answer FNS-relevant research questions or variables that allow integration of a dataset with another), shall be developed. The ontology would provide a coherent means of data annotation to achieve semantic interoperability. Concepts which already exist in available ontologies (see Table 2) can be imported.

- Once an FNS-Cloud design (pooled vs. federated) is clear/ agreed upon, its technical and other boundary conditions should be clarified that will influence harmonization needs and standardization efforts. This goes hand in hand with clarifying ethical and legal data sharing aspects and data access rights. Datasets already available to the FNS-Cloud project consortium as well as datasets to be later uploaded/submitted to the Cloud should undergo a 'clearance' process, addressing issues such as:

  - Is study metadata going to be made publicly available?
  - Is raw or aggregated study data going to be shared? Who has the rights to share this data?
  - Are there any ethical-legal constraints to sharing and reusing the data?

5.2 FNS-Cloud data providers

For FNS-Cloud data providers, the following list of recommendations was defined:

- Existing reporting guidelines and data standards, as well as ontologies or thesauri (see Table 1 and Table 2) , which can help produce data in an already standardized/ harmonized way, and in turn a "Cloud-compatible"/ machine-readable or easy to transform way, shall be identified and used early on in the study period. Software tools can also present useful resources if they e.g. provide a template to fill out, containing already standardized parameters (e.g. dietary intake survey software) and can be used when performing studies.

- Correct, complete and, if possible, ontology-based documentation of study metadata is paramount if data is to be made interoperable. We recommend using the above-mentioned ISA-Tab structure for experimental metadata collection and exchange. It is further recommended to search for data type-specific minimal requirements or checklists developed for a respective research field, containing assay-specific recommendations such as:
  - Minimal requirements for joint data analysis in nutritional epidemiology (Pinart et al., 2018)
  - Minimum information for biobank data sharing (Merino-Martinez et al., 2016)
  - Minimum information required for a glycomics experiment (Kolarich et al., 2013)
  - Minimum Information about a Genome/Metagenome Sequence checklist (MIGS/MIMS) (Field et al., 2008). Furthermore, to cover the description of phylogenetic and functional marker genes an extended standard, the Minimum Information about a MARKer gene Sequence (MIMARKS) checklist (http://gensc.org/gc_wiki/index.php/MIMARKS) has been developed (Yilmaz et al., 2011)
  - Reporting recommendations for protein data (MIAPE) (Taylor et al., 2007)
  - Requirements for the description of transcriptomic data (MIAME/Plant) (Zimmermann et al., 2006)
  - Rules concerning metabolomics observations (Fiehn et al., 2007; Goodacre et al., 2007; Morrison et al., 2007)

- Data sharing rights and legal and ethical issues shall be clarified at the earliest convenience and included in the study metadata.

5.3 FNS-Cloud platform users

Regardless of the specific user group (researchers, policy makers, industry... etc.), the following recommendations were defined:

- The research questions shall be clearly defined (reduce ambiguity, use keywords), as well as data requirements (data types and specific variables) in order to perform efficient queries and identify suitable data on the FNS-Cloud. Ideally ontology-based language is used to express research questions to facilitate matching and connecting concepts between available information sources and user information needs
- Data licensing agreements shall be read carefully and respected
- Metadata associated with any given dataset shall be read carefully at the start to avoid data misinterpretations or misuse
- ‘True’ comparability of studies as well as data quality shall be checked

6. Conclusions

The overarching goal of T3.1 was to demonstrate the process of identifying data requirements and developing applicability criteria as the basis for developing information technology tools that will enable semi-automated data integration and interoperability within the FNS-Cloud data platform. As a starting point, an ‘FNS data map’ that represents FNS-related ‘research question spaces’ as well FNS-relevant data domains and data types, was developed. The data map can be used to predetermine the categories (types) of data that will included on the Cloud and subsequently cluster and ‘tag’ submitted FNS data(sets) to enable efficient identification of suitable datasets in response to user queries.

In a second step, typical inconsistencies between different FNS data (sets) in e.g. nomenclature (vocabulary used), units, data field types, file formats etc. were identified and discussed. Two illustrative case studies, using one cross-domain and one single-domain research question, were then used to demonstrate the development of generic (domain-independent) and domain-specific applicability criteria. With the help of existing FNS datasets used to answer the research questions, it was possible to identify all variables requiring harmonization in order to co-analyse the different datasets and answer the research question at hand. Applicability criteria were then developed for these variables to help formalize data processing procedures, which will have to take place ‘in the background’ to enable semi-automated data integration and harmonization within the Cloud.

Platform developers can follow the demonstrated procedure and the example criteria provided to define further data requirements and develop more applicability criteria for all types of data that will be included on the FNS-Cloud. This will directly feed into their efforts in other tasks of WP3 to develop automation algorithms as well as other data processing tools to enable automated data harmonization and integration. Additionally, identified harmonization requirements and developed criteria help data providers (e.g. WP4 use cases and WP5 demonstrators) in generating already standardized (machine-readable) data consistent with existing data standards and reporting guidelines.

The two case studies have further illustrated the difficulties and challenges of harmonizing datasets from different sources or research labs and achieving data interoperability. To answer complex FNS-relevant research questions, it is necessary to combine and conjointly analyse data of different types and from different data (or research) domains. The process of developing applicability criteria for key variables that require harmonization will have to be repeated for multiple other datasets to clearly determine what data processing steps are required and what IT solutions are needed for. It remains to be seen how well-developed solutions (e.g. machine learning algorithms) will allow (semi-)automatic data integration and consolidation into the Cloud.

7. References

Amoutzopoulos, B., Steer, T., Roberts, C., Collins, D., & Page, P. (2020). Free and Added Sugar Consumption and Adherence to Guidelines: The UK National Diet and Nutrition Survey (2014/15–2015/16). *Nutrients*, *12*(2), 393. https://doi.org/10.3390/nu12020393

Anastasiou, C. A., Fappa, E., Zachari, K., Mavrogianni, C., Van Stappen, V., Kivelä, J., Virtanen, E., González-Gil, E. M., Flores-Barrantes, P., Nánási, A., Semánová, C., Dimova, R., Usheva, N., Iotova, V., Cardon, G., Manios, Y., & Makrilakis, K. (2020). Development and reliability of questionnaires for the assessment of diet and physical activity behaviors in a multi-country sample in Europe the Feel4Diabetes Study. *BMC Endocrine Disorders*, *20*. https://doi.org/10.1186/s12902-019-0469-x

Baecke, J. A. H., Burema, J., & Frijters, J. E. R. (1982). A short questionnaire for the measurement of habitual physical activity in epidemiological studies. *American Journal of Clinical Nutrition*, *36*(5), 936–942. https://doi.org/10.1093/ajcn/36.5.936

Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M. H., Bug, B., Chibucos, M. C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M. A., He, Y., Heiskanen, M., Hernandez-Boussard, T., … Zheng, J. (2016). The Ontology for Biomedical Investigations. *PLoS ONE*, *11*(4). https://doi.org/10.1371/journal.pone.0154556

Becker, W., Moller, 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. EuroFIR Technical Report.* http://www.eurofir.org/wp-content/uploads/2014/05/2.-Proposal-for-structure-and-detail-of-a-EuroFIR-standard-on-food-composition-data-II-Technical-Annex.pdf

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, *32*(DATABASE ISS.), 267–270. https://doi.org/10.1093/nar/gkh061

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., & Lewis, S. E. (2013). The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, *4*(1), 43. https://doi.org/10.1186/2041-1480-4-43

Celis-Morales, C., Livingstone, K. M., Marsaux, C. F. M., Forster, H., O'Donovan, C. B., Woolhead, C., Macready, A. L., Fallaize, R., Navas-Carretero, S., San-Cristobal, R., Kolossa, S., Hartwig, K., Tsirigoti, L., Lambrinou, C. P., Moschonis, G., Godlewska, M., Surwiłło, A., Grimaldi, K., Bouwman, J., … Mathers, J. C. (2015). Design and baseline characteristics of the Food4Me study: a web-based randomised controlled trial of personalised nutrition in seven European countries. *Genes and Nutrition*, *10*(1). https://doi.org/10.1007/s12263-014-0450-2

Celis-Morales, C., Livingstone, K. M., Marsaux, C. F. M., Macready, A. L., Fallaize, R., O'Donovan, C. B., Woolhead, C., Forster, H., Walsh, M. C., Navas-Carretero, S., San-Cristobal, R., Tsirigoti, L., Lambrinou, C. P., Mavrogianni, C., Moschonis, G., Kolossa, S., Hallmann, J., Godlewska, M., Surwiłło, A., … Mathers, J. C. (2017). Effect of personalized nutrition on health-related behaviour change: Evidence from the Food4Me European randomized controlled trial. *International Journal of Epidemiology*, *46*(2), 578–588. https://doi.org/10.1093/ije/dyw186

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., Fiorani, F., Frohmberg, W., Junker, A., Klukas, C., Lange, M., Mazurek, C., Nafissi, A., Neveu, P., Oeveren, J., Pommier, C., Poorter, H., Rocca-Serra, P., Sansone, S. A., … Krajewski, P. (2016). Measures for interoperability of phenotypic data: Minimum information requirements and formatting. *Plant Methods*, *12*(1), 44. https://doi.org/10.1186/s13007-016-0144-4

DCMI. (2020). *Dublin Core^{TM} Metadata Initiative*. https://dublincore.org/

de Matos, P., Alcántara, R., Dekker, A., Ennis, M., Hastings, J., Haug, K., Spiteri, I., Turner, S., & Steinbeck, C. (2009). Chemical entities of biological interest: An update. *Nucleic Acids Research*, *38*(SUPPL.1). https://doi.org/10.1093/nar/gkp886

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., Stolk, R. P., Foco, L., Minelli, C., Waldenberger, M., Holle, R., Kvaløy, K., Hillege, H. L., Tassé, A. M., Ferretti, V., & Fortier, I. (2013). Data harmonization and federated analysis of population-based studies: The BioSHaRE project. *Emerging Themes in Epidemiology*, *10*(1). https://doi.org/10.1186/1742-7622-10-12

Dooley, D. M., Griffiths, E. J., Gosal, G. S., Buttigieg, P. L., Hoehndorf, R., Lange, M. C., Schriml, L. M., Brinkman, F. S. L., & Hsiao, W. W. L. (2018). FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration. *Npj Science of Food*, *2*(1). https://doi.org/10.1038/s41538-018-0032-6

EFSA. (2013). Standard Sample Description ver. 2.0. *EFSA Journal*, *11*(10). https://doi.org/10.2903/j.efsa.2013.3424

EFSA. (2014). Guidance on the EU Menu methodology. *EFSA Journal*, *12*(12). https://doi.org/10.2903/j.efsa.2014.3944

EFSA. (2015). The food classification and description system FoodEx 2 (revision 2). In *EFSA Supporting Publication* (Vols. 2015:EN-80). Wiley. https://doi.org/10.2903/sp.efsa.2015.en-804

Eftimov, T., Ispirova, G., Korošec, P., & Seljak, B. K. (2018). The RICHFIELDS framework for semantic interoperability of food information across heterogenous information systems. *IC3K 2018 - Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, *1*(September), 315–322. https://doi.org/10.5220/0006951703150322

Eftimov, T., Ispirova, G., Potočnik, 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*, *277*, 382–390. https://doi.org/10.1016/j.foodchem.2018.10.118

Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2017). Standfood: Standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients*, *9*(6). https://doi.org/10.3390/nu9060542

Ekoe, J. M., Goldenberg, R., & Katz, P. (2018). Screening for Diabetes in Adults. *Canadian Journal of Diabetes*, *42*, S16–S19. https://doi.org/10.1016/j.jcjd.2017.10.004

Fantke, P., Arnot, J. A., & Doucette, W. J. (2016). Improving plant bioaccumulation science through consistent reporting of experimental data. *Journal of Environmental Management*, *181*, 374–384. https://doi.org/10.1016/j.jenvman.2016.06.065

Federal Food Safety and Veterinary Office. (2020). *The Swiss Food Composition Database*. https://www.naehrwertdaten.ch/en/

Gilmer, T. P., & O'Connor, P. J. (2010). The growing importance of diabetes screening. In *Diabetes Care* (Vol. 33, Issue 7, pp. 1695–1697). https://doi.org/10.2337/dc10-0855

Gkoutos, G. V., Schofield, P. N., & Hoehndorf, R. (2012). The Units Ontology: a tool for integrating units of measurement in science. *Database : The Journal of Biological Databases and Curation*, *2012*, 1–7. https://doi.org/10.1093/database/bas033

GovEx. (n.d.). Getting meta with metadata. A City's guide to high quality, discoverable, and understandable open data. In *Open Data Guidebooks*. https://centerforgov.gitbooks.io/open-data-metadata-guide/

Gray, L. J., Taub, N. A., Khunti, K., Gardiner, E., Hiles, S., Webb, D. R., Srinivasan, B. T., & Davies, M. J. (2010). The Leicester Risk Assessment score for detecting undiagnosed Type 2 diabetes and impaired glucose regulation for use in a multiethnic UK setting. *Diabetic Medicine*, *27*(8), 887–895. https://doi.org/10.1111/j.1464-5491.2010.03037.x

Hippisley-Cox, J., & Coupland, C. (2017). Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study. *BMJ (Clinical Research Ed.)*, *359*, j5019. https://doi.org/10.1136/bmj.j5019

Humphreys, B. L., & Lindberg, D. A. (1993). The UMLS project: making the conceptual connection between users and the information they need. *Bulletin of the Medical Library Association*, *81*(2), 170–177. http://www.ncbi.nlm.nih.gov/pubmed/8472002

ILO. (2012). *International Standard Classifi cation of Occupations ISCO-08*. International Labour Office.

ISA-tools. (2014). *STATO: an Ontology of Statistical Methods*. http://stato-ontology.org/index.jsp

ISO 15836-1:2017. (2017). *Information and documentation — The Dublin Core metadata element set — Part 1: Core elements*. International Standards Organization. https://www.iso.org/standard/71339.html

ISO 80000-1:2009. (2009). *Quantities and units — Part 1: General*. https://www.iso.org/standard/30669.html

Kanellakis, S., Mavrogianni, C., Karatzi, K., Lindstrom, J., Cardon, G., Iotova, V., Wikström, K., Shadid, S., Moreno, L. A., Tsochev, K., Bíró, É., Dimova, R., Antal, E., Liatis, S., Makrilakis, K., & Manios, Y. (2020). Development and validation of two self-reported tools for insulin resistance and hypertension risk assessment in a european cohort: The feel4diabetes-study. *Nutrients*, *12*(4). https://doi.org/10.3390/nu12040960

Katerinopoulou, K., Kontogeorgos, A., Salmas, C. E., Patakas, A., & Ladavos, A. (2020). Geographical Origin Authentication of Agri-Food Products: A Review. *Foods 2020, Vol. 9, Page 489*, *9*(4), 489. https://doi.org/10.3390/FOODS9040489

Kolarich, D., Rapp, E., Struwe, W. B., Haslam, S. M., Zaia, J., McBride, R., Agravat, S., Campbell, M. P., Kato, M., Ranzinger, R., Kettner, C., & York, W. S. (2013). The minimum information required for a glycomics experiment (MIRAGE) project: Improving the standards for reporting mass-spectrometry-based glycoanalytic data. *Molecular and Cellular Proteomics*, *12*(4), 991–995. https://doi.org/10.1074/mcp.O112.026492

Korošec, Ž., & Pravst, I. (2014). Assessing the Average Sodium Content of Prepacked Foods with Nutrition Declarations: The Importance of Sales Data. *Nutrients*, *6*(9), 3501–3515. https://doi.org/10.3390/nu6093501

Koroušić Seljak, B., Korošec, P., Eftimov, T., Ocke, M., van der Laan, J., Roe, M., Berry, R., Crispim, S., Turrini, A., Krems, C., Slimani, N., & Finglas, P. (2018). Identification of Requirements for Computer-Supported Matching of Food Consumption Data with Food Composition Data. *Nutrients*, *10*(4), 433. https://doi.org/10.3390/nu10040433

Lachat, C., Hawwash, D., Ocké, M. C., Berg, C., Forsum, E., Hörnell, A., Larsson, C., Sonestedt, E., Wirfält, E., Åkesson, A., Kolsteren, P., Byrnes, G., De Keyzer, W., Van Camp, J., Cade, J. E., Slimani, N., Cevallos, M., Egger, M., & Huybrechts, I. (2016). Strengthening the Reporting of Observational Studies in Epidemiology—Nutritional Epidemiology (STROBE-nut): An Extension of the STROBE Statement. *PLoS Medicine*, *13*(6), 1–15. https://doi.org/10.1371/journal.pmed.1002036

Lindberg, D. A. B., Humphreys, B. L., & McCray, A. T. (1993). The unified medical language system. In *Methods of Information in Medicine* (Vol. 32, Issue 4, pp. 281–291). Methods Inf Med. https://doi.org/10.1055/s-0038-1634945

Lindström, J., & Tuomilehto, J. (2003). The diabetes risk score: A practical tool to predict type 2 diabetes risk. *Diabetes Care*, *26*(3), 725–731. https://doi.org/10.2337/diacare.26.3.725

Macháčková, M., Møller, A., & Ireland, J. (2017). *The EuroFIR Thesauri - Update wave 2016 – A report*. http://www.eurofir.org/wp-content/uploads/2017/06/Update-wave-2016-FINAL-170525.pdf

Manios, Y., Androutsos, O., Lambrinou, C. P., Cardon, G., Lindstrom, J., Annemans, L., Mateo-Gallego, R., De Sabata, M. S., Iotova, V., Kivela, J., Martinez, R., Moreno, L. A., Rurik, I., Schwarz, P., Tankova, T., Liatis, S., & Makrilakis, K. (2018). A school- and community-based intervention to promote healthy lifestyle and prevent type 2 diabetes in vulnerable families across Europe: Design and implementation of the Feel4Diabetes-study. In *Public Health Nutrition* (Vol. 21, Issue 17, pp. 3281–3290). Cambridge University Press. https://doi.org/10.1017/S1368980018002136

Merino-Martinez, R., Norlin, L., Van Enckevort, D., Anton, G., Schuffenhauer, S., Silander, K., Mook, L., Holub, P., Bild, R., Swertz, M., & Litton, J. E. (2016). Toward Global Biobank Integration by Implementation of the Minimum Information about BIobank Data Sharing (MIABIS 2.0 Core). *Biopreservation and Biobanking*, *14*(4), 298–306. https://doi.org/10.1089/bio.2015.0070

Møller, A., & Ireland, J. (2017). *LanguaL^{TM} 2017 The LanguaL^{TM} Thesaurus*. Danish Food Informatics 2018.

Morrison, N., Bearden, D., Bundy, J. G., Collette, T., Currie, F., Davey, M. P., Haigh, N. S., Hancock, D., Jones, O. A. H., Rochfort, S., Sansone, S. A., Štys, D., Teng, Q., Field, D., & Viant, M. R. M. R. (2007). Standard reporting requirements for biological samples in metabolomics experiments: Environmental context. *Metabolomics*, *3*(3), 203–210. https://doi.org/10.1007/s11306-007-0067-1

Muljarto, A. R., Salmon, J. M., Charnomordic, B., Buche, P., Tireau, A., & Neveu, P. (2017). A generic ontological network for Agri-food experiment integration – Application to viticulture and winemaking. *Computers and Electronics in Agriculture*, *140*(August), 433–442. https://doi.org/10.1016/j.compag.2017.06.020

Nogueira, E., Moreira, A., Lucrédio, D., Garcia, V., & Fortes, R. (2016). Issues on developing interoperable cloud applications: definitions, concepts, approaches, requirements, characteristics and evaluation models. *Journal of Software Engineering Research and Development*, *4*(1), 1–23. https://doi.org/10.1186/s40411-016-0033-6

Nutrition Institute. (2018). *Composition and Labelling Information System as a tool for monitoring of the food supply*. https://www.nutris.org/en/composition-and-labelling-information-system

Pinart, M., Nimptsch, K., Bouwman, J., Dragsted, L. O., Yang, C., De Cock, N., Lachat, C., Perozzi, G., Canali, R., Lombardo, R., D'Archivio, M., Guillaume, M., Donneau, A. F., Jeran, S., Linseisen, J., Kleiser, C., Nöthlings, U., Barbaresko, J., Boeing, H., … Pischon, T. (2018). Joint data analysis in nutritional epidemiology: Identification of observational studies and minimal requirements. *Journal of Nutrition*, *148*(2), 285–297. https://doi.org/10.1093/jn/nxx037

Presser, K., Roe, M., Matuszczak, A., & Finglas, P. (2020). *Food Nutrition Security Cloud. D2.1 Definition of data models and APIs. FNS-Cloud (Grant Agreement No. 863059)*.

Presser, K., Weber, D., & Norrie, M. (2018). FoodCASE: A system to manage food composition, consumption and TDS data. *Food Chemistry*, *238*(November), 166–172. https://doi.org/10.1016/j.foodchem.2016.09.124

Ruiz, E., Rodriguez, P., Valero, T., Ávila, J., Aranceta-Bartrina, J., Gil, Á., González-Gross, M., Ortega, R., Serra-Majem, L., & Varela-Moreiras, G. (2017). Dietary Intake of Individual (Free and Intrinsic) Sugars and Food Sources in the Spanish Population: Findings from the ANIBES Study. *Nutrients*, *9*(3), 275. https://doi.org/10.3390/nu9030275

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L. A., Copeland, J., Das, S., … Hide, W. (2012). Toward interoperable bioscience data. In *Nature Genetics* (Vol. 44, Issue 2, pp. 121–126). Nature Publishing Group. https://doi.org/10.1038/ng.1054

Sansone, S.-A., Rocca-Serra, P., Gonzalez-Beltran, A., & Johnson, D. (2016). *Isa Model And Serialization Specifications 1.0*. https://doi.org/10.5281/ZENODO.163640

Schmidt, M. I., Duncan, B. B., Bang, H., Pankow, J. S., Ballantyne, C. M., Golden, S. H., Folsom, A. R., & Chambless, L. E. (2005). Identifying individuals at high risk for diabetes: The Atherosclerosis Risk in Communities study. *Diabetes Care*, *28*(8), 2013–2018. https://doi.org/10.2337/diacare.28.8.2013

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 Statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine*, *8*. https://doi.org/10.1186/1741-7015-8-18

Soergel, D., Lauser, B., Liang, A., Fisseha, F., Keizer, J., & Katz, S. (2004). Reengineering thesauri for new applications: The AGROVOC example. *Journal of Digital Information*, *4*(4).

Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., Brazma, A., Brinkman, R. R., Clark, A. M., Deutsch, E. W., Fiehn, O., Fostel, J., Ghazal, P., Gibson, F., Gray, T., … Wiemann, S. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature Biotechnology 2008 26:8*, *26*(8), 889–896. https://doi.org/10.1038/nbt.1411

Taylor, C. F., Paton, N. W., Lilley, K. S., Binz, P. A., Julian, R. K., Jones, A. R., Zhu, W., Apweiler, R., Aebersold, R., Deutsch, E. W., Dunn, M. J., Heck, A. J. R., Leitner, A., Macht, M., Mann, M., Martens, L., Neubert, T. A., Patterson, S. D., Ping, P., … Hermjakob, H. (2007). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology*, *25*(8), 887–893. https://doi.org/10.1038/nbt1329

Thompson, F. E., & Subar, A. F. (2017). Dietary Assessment Methodology. In *Nutrition in the Prevention and Treatment of Disease* (pp. 5–48). Elsevier. https://doi.org/10.1016/B978-0-12-802928-2.00001-1

UNESCO-UIS. (2012). *International Standard Classification of Education*. UNESCO Institute for Statistics. http://www.uis.unesco.org

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. (2018). ONS: An ontology for a standardized description of interventions and observational studies in nutrition. *Genes and Nutrition*, *13*(1), 1–9. https://doi.org/10.1186/s12263-018-0601-y

Westenbrink, S. (2020). *Personal communication (E-mail 4-6-2020)*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(1), 1–9. https://doi.org/10.1038/sdata.2016.18

Yang, C., Ambayo, H., De Baets, B., Kolsteren, P., Thanintorn, N., Hawwash, D., Bouwman, J., Bronselaer, A., Pattyn, F., & Lachat, C. (2019). An ontology to standardize research output of nutritional epidemiology: From paper-based standards to linked content. *Nutrients*, *11*(6). https://doi.org/10.3390/nu11061300

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., … Glöckner, F. O. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. In *Nature Biotechnology* (Vol. 29, Issue 5, pp. 415–420). Nature Publishing Group. https://doi.org/10.1038/nbt.1823

Z39.85-2012. (2013). *The Dublin Core Metadata Element Set*. American National Standards Institute, National Information Standards Organization.

Zupanič, N., Miklavec, K., Kušar, A., Žmitek, K., Fidler Mis, N., & Pravst, I. (2018). Total and Free Sugar Content of Pre-Packaged Foods and Non-Alcoholic Beverages in Slovenia. *Nutrients*, *10*(2), 151. https://doi.org/10.3390/nu10020151

8. Appendix

Appendix 1: Overview of FNS datasets mapped to data domains and potential research questions

Table 4: FNS data[1] available to project consortium mapped to the associated data domains, data types and potential research questions addressed in WP4 and WP5

Dataset	Data owner/ Project partner	Data domain															Potential research questions[2]
		Agri-food					Food intake and lifestyle					Health, body function and disease risk					
		Composition data	Branded food data	Analytical food chemistry/ omics data	Sustainability data	Others	Consumption data	Socio-demographic data	Lifestyle data	Anthropometric data	Others	Phenotypic data	Biomarker data	Hazard & toxicological data	Omics data	Others	
FoodEXplorer	EuroFIR	x															*What are differences in the nutritional quality of the food supply in selected EU MS across food categories and for whole-country samples?*
FoodWasteEXplorer	EuroFIR					x											
ePlantLIBRA	EuroFIR	x												x			
eBASIS (Bioactive Substances in Food Information Systems)	EuroFIR	x												x			
SCaRES - Seafood Study Ireland	UCD						x	x	x								*How are consumer behaviors linked to dietary intake and how do such associatons in UK, IE and DE compare to Western Balkan countries?*
Food4me	UCD						x	x	x				x			x	*How high is the cardiometabolic disease risk given a certain personal diet/ dietary intake and lifestyle?*
Gut microbiome dataset	QIB														x		*What are the effects of a high-bioactive complex diet and a low-bioactive diet on the composition and function of the gut microbiome as well as other clinical and disease markers?*
Food and associated glucose data	QIB												x				
Metabolomics data	QIB														x		
EFSA Comprehensive European Food Consumption Database	Hylo						x										
LFCT-AUTH/ATR-FTIR spectroscopic dataset	AUTH	x	x														
NEVO Dutch food composition database	RIVM	x															

Dataset	Data owner/ Project partner	Data domain															Potential research questions[2]
		Agri-food					Food intake and lifestyle					Health, body function and disease risk					
		Composition data	Branded food data	Analytical food chemistry/ omics data	Sustainability data	Others	Consumption data	Socio-demographic data	Lifestyle data	Anthropometric data	Others	Phenotypic data	Biomarker data	Hazard & toxicological data	Omics data	Others	
LEDA Dutch branded food database (not publicly available)	RIVM		x														
Slovenian branded food database	Nutris, JSI		x														*How does a branded food composition/ labelling dataset collected via a standard Food Monitoring Survey in one EU MS (SI) compare to a dataset developed with the support of crowdsourcing?*
Tomappo garden plans dataset	Lifely					x											
Contaminants_simplified_SSD2_BE-FPS_Nickel_FNS	UGent	x															*What is the total diet combined exposure to multiple chemicals & what is the associated risk?* *For what substances are exposures (a) over the legislated limits, and (b) close to the legislated limits?*
Belgian food consumption data 2014-2015	UGent						x		x								
Feel4Diabetes study_High risk adults_baseline_GR	HUA						x	x		x			x				
Isotopic data	FEM			x													
WikiPathways	UM															x	
CNR_IBBA_TBP-based dataset	CNR					x											
CNR_ISPA_Salmon mass spectrometric dataset	CNR			x													
CoFID	IMDEA Food	x															
FooDB	IMDEA Food	x				x								x		x	
Phenol-Explorer	IMDEA Food	x															
European public assessment reports (EPAR)	IMDEA Food															x	
Drug Bank	IMDEA Food															x	
Drugs.com	IMDEA Food															x	
Medscape. com	IMDEA Food															x	

Dataset	Data owner/ Project partner	Data domain															Potential research questions[2]
		Agri-food					Food intake and lifestyle					Health, body function and disease risk					
		Composition data	Branded food data	Analytical food chemistry/ omics data	Sustainability data	Others	Consumption data	Socio-demographic data	Lifestyle data	Anthropometric data	Others	Phenotypic data	Biomarker data	Hazard & toxicological data	Omics data	Others	
Rxisk.org	IMDEA Food															x	
NutriChem 2.0	IMDEA Food															x	*How does a given drug interact with specific food/ a specific diet?*
Webmd	IMDEA Food										x					x	
ClinicalTrials	IMDEA Food															x	
dailyMed	IMDEA Food															x	
CNR_ISPAAM_bovine milk_thermal treatment_peptides	CNR			x													*How can the authenticity and quality of milk be judged?*
CNR_ISPAAM_bovine milk_thermal treatment_proteins	CNR			x													
CNR_ISPAAM_bovine milk_thermal treatment adulteration_peptides	CNR			x													
CNR_ISPAAM_bovine milk_thermal treatment adulteration_proteins	CNR			x													
CNR_ISPAAM_buffalo milk_freezing overtime_proteins	CNR			x													
CNR_ISPAAM_BoMiProt	CNR			x													
CNR_ISPAAM_bovine milk bioactive peptides_MBPDB	CNR			x													
CNR_ISPAAM_bovine milk bioactive peptides_MBPDB	CNR			x													
CNR_ISPAAM_caprine milk bioactive peptides_MBPDB	CNR			x													
CNR_ISPAAM_human milk bioactive peptides_MBPDB	CNR			x													
CNR_ISPAAM_AGEs-containing proteins	CNR			x													
CNR_ISPAAM_lactosylated proteins	CNR			x													

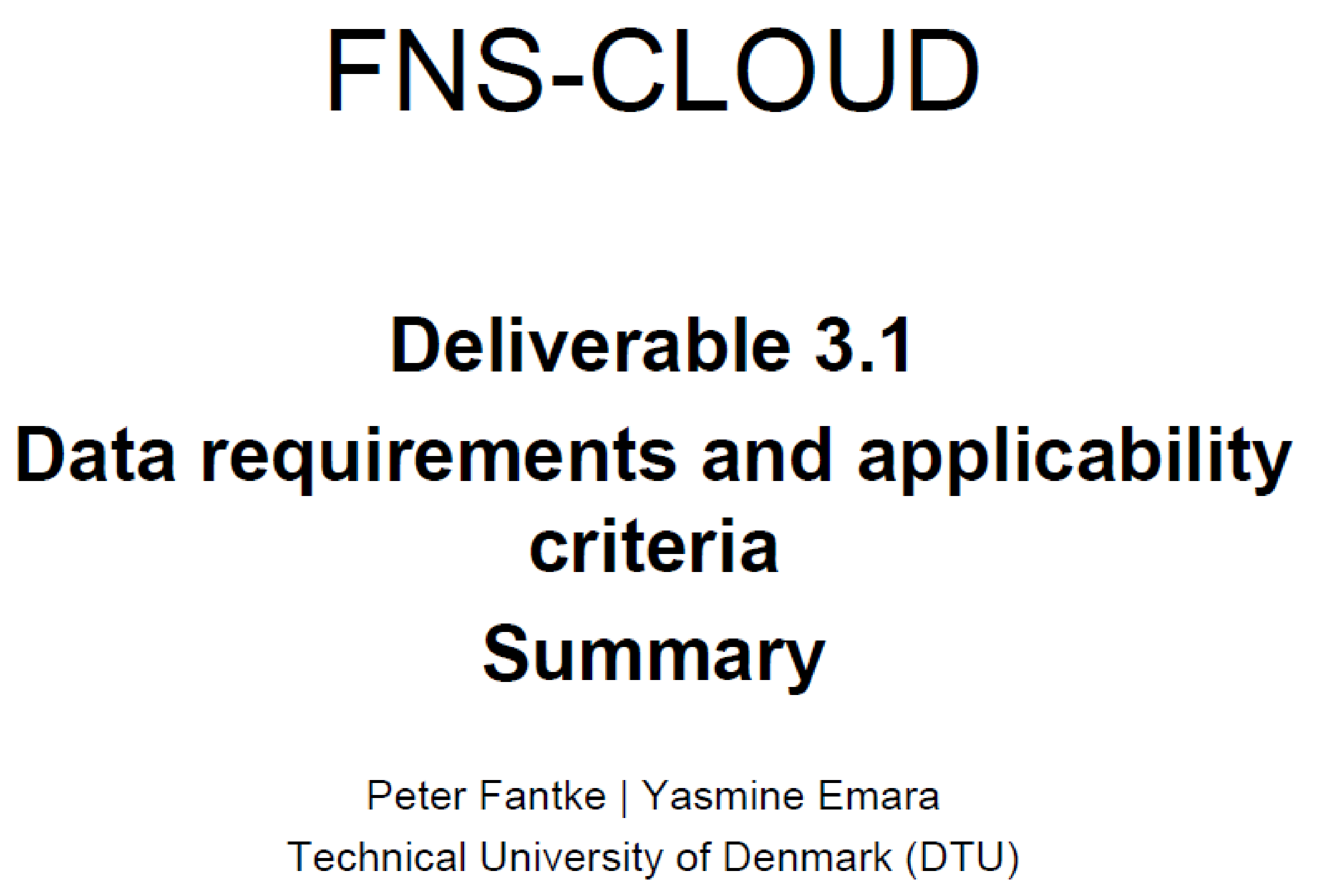
Food Nutrition Security Cloud (FNS-Cloud) has received funding from the European Union's Horizon 2020 Research and Innovation programme (H2020-EU.3.2.2.3. – A sustainable and competitive agri-food industry) under Grant Agreement No. 863059 – www.fns-cloud.eu

Dataset	Data owner/ Project partner	Data domain															Potential research questions[2]
		Agri-food					Food intake and lifestyle					Health, body function and disease risk					
		Composition data	Branded food data	Analytical food chemistry/ omics data	Sustainability data	Others	Consumption data	Socio-demographic data	Lifestyle data	Anthropometric data	Others	Phenotypic data	Biomarker data	Hazard & toxicological data	Omics data	Others	
CNR_ISPAAM_carbonylated proteins	CNR			x													
CNR_ISPAAM_milk metabolites_MCDB	CNR			x													
The Serbian National Food Consumption Survey	CAP						x	x		x	x						
Folate intake among Serbian women of reproductive age	CAP						x	x		x							
Swiss food composition database V6.1, 2019 and previous versions	PMT	x	x														
LFCT-AUTH/virgin olive oil phenol composition dataset	AUTH			x													*What is the geographic origin of a given olive oil sample (product)?*
LFCT-AUTH/quantum chemically calculated values of indices characterizing the radical scavenging activity of virgin olive oil phenols dataset	AUTH			x													
Chemical dissipation half-lives in food crops and other plants	DTU					x											
Dietary intake of people aged 65+	TUM, UoR						x	x									*What are dietary habits of vulnerable populations (65+ years) and how do they affect health outcomes?*
ESRC Cognitive Food Choice	UoR								x		x						
eNutri (Quisper 2019 EatwellUK2 study)	UoR						x	x			x						

[1] Some resources such as Pubmed and Arxiv were not included in the table, as they are repositories/ search engines for literature but not data. They can nevertheless be searched for published literature/ data.

[2] Example of questions addressed in WP4 and WP5 of the FNS-Cloud project (see description of work for more details). The questions will be answered solely on the basis of the dataset they've been assigned to in this table or will require the combination of multiple datasets to be answered (but were only mentioned once in relation to one of those dataset

Appendix 2: Summary slides



FNS-CLOUD

Deliverable 3.1

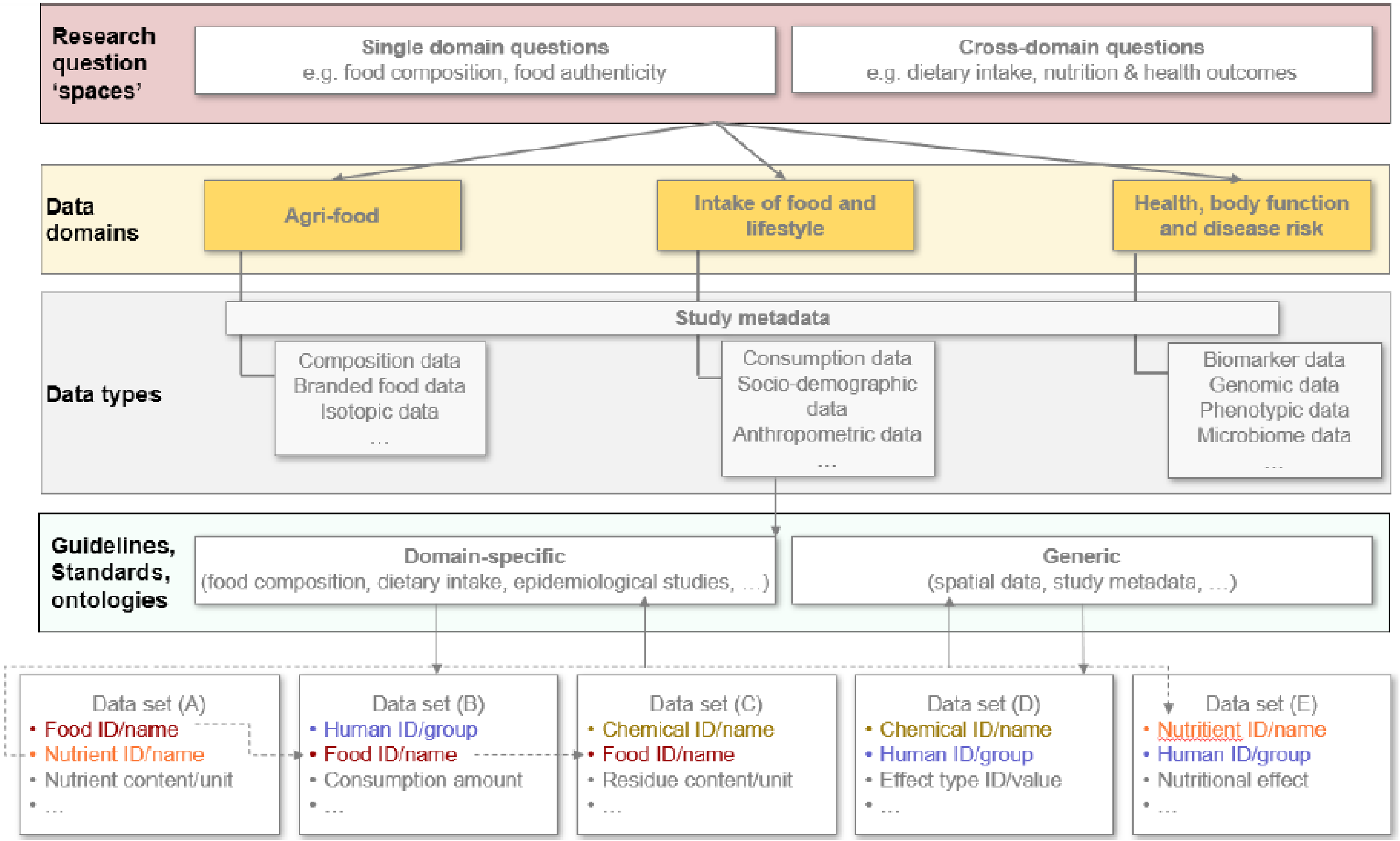
Data requirements and applicability criteria

Summary

Peter Fantke | Yasmine Emara

Technical University of Denmark (DTU)

Map FNS data to research questions, standards, ontologies and guidelines



Research question 'spaces'

Single domain questions
e.g. food composition, food authenticity

Cross-domain questions
e.g. dietary intake, nutrition & health outcomes

Data domains

Agri-food

Intake of food and lifestyle

Health, body function and disease risk

Data types

Study metadata

Composition data
Branded food data
Isotopic data
…

Consumption data
Socio-demographic data
Anthropometric data
…

Biomarker data
Genomic data
Phenotypic data
Microbiome data
…

Guidelines, Standards, ontologies

Domain-specific
(food composition, dietary intake, epidemiological studies, …)

Generic
(spatial data, study metadata, …)

Data set (A)
- Food ID/name
- Nutrient ID/name
- Nutrient content/unit
- …

Data set (B)
- Human ID/group
- Food ID/name
- Consumption amount
- …

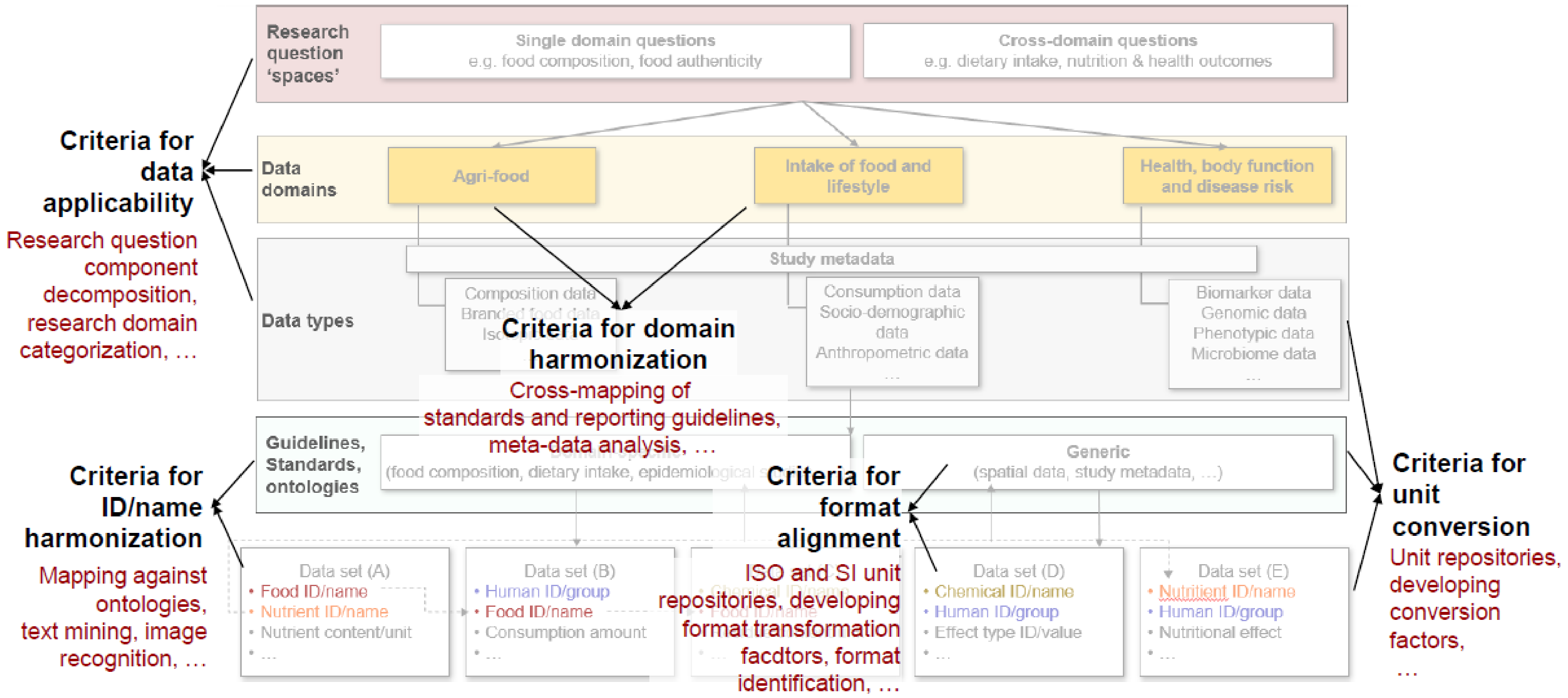
Data set (C)
- Chemical ID/name
- Food ID/name
- Residue content/unit
- …

Data set (D)
- Chemical ID/name
- Human ID/group
- Effect type ID/value
- …

Data set (E)
- Nutritient ID/name
- Human ID/group
- Nutritional effect
- …

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

Data interoperability criteria definition and implementation in FNS cloud



Research question 'spaces'

Single domain questions
e.g. food composition, food authenticity

Cross-domain questions
e.g. dietary intake, nutrition & health outcomes

Data domains

Agri-food

Intake of food and lifestyle

Health, body function and disease risk

Data types

Study metadata

Composition data
Branded food data
Isotopic data

Consumption data
Socio-demographic data
Anthropometric data
…

Biomarker data
Genomic data
Phenotypic data
Microbiome data
…

Guidelines, Standards, ontologies

Domain specific
(food composition, dietary intake, epidemiological s…

Generic
(spatial data, study metadata, …)

Data set (A)
- Food ID/name
- Nutrient ID/name
- Nutrient content/unit
- …

Data set (B)
- Human ID/group
- Food ID/name
- Consumption amount
- …

Data set (C)
- Chemical ID/name
- Food ID/name
- …

Data set (D)
- Chemical ID/name
- Human ID/group
- Effect type ID/value
- …

Data set (E)
- Nutritient ID/name
- Human ID/group
- Nutritional effect
- …

Criteria for data applicability

Research question component decomposition, research domain categorization, …

Criteria for domain harmonization

Cross-mapping of standards and reporting guidelines, meta-data analysis, …

Criteria for ID/name harmonization

Mapping against ontologies, text mining, image recognition, …

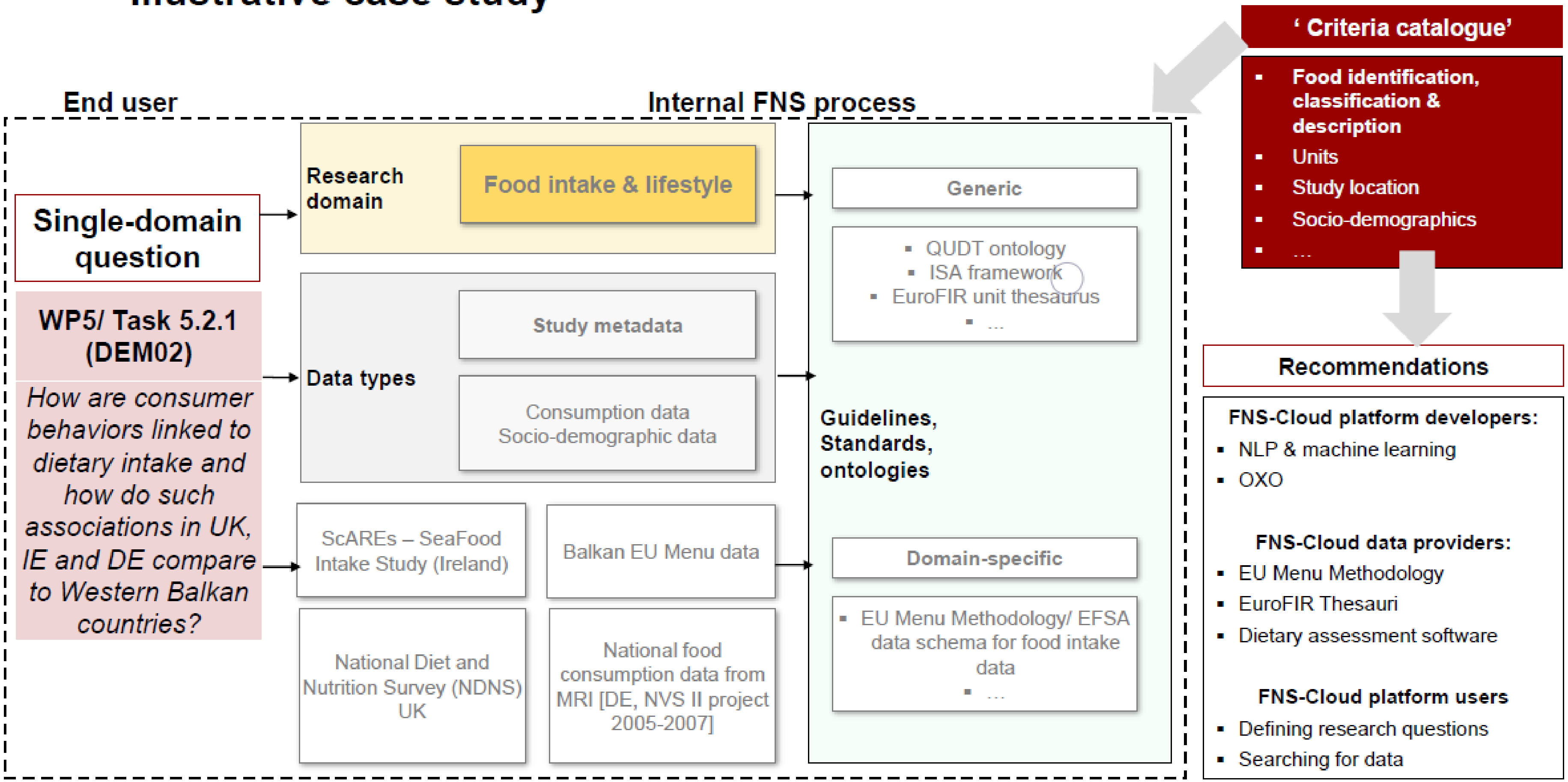
Criteria for format alignment

ISO and SI unit repositories, developing format transformation facdtors, format identification, …

Criteria for unit conversion

Unit repositories, developing conversion factors, …

Illustrative case study



End user

Internal FNS process

Single-domain question

WP5/ Task 5.2.1 (DEM02)

How are consumer behaviors linked to dietary intake and how do such associations in UK, IE and DE compare to Western Balkan countries?

Research domain

Food intake & lifestyle

Data types

Study metadata

Consumption data
Socio-demographic data

ScAREs – SeaFood Intake Study (Ireland)

Balkan EU Menu data

National Diet and Nutrition Survey (NDNS) UK

National food consumption data from MRI [DE, NVS II project 2005-2007]

Guidelines, Standards, ontologies

Generic

- QUDT ontology
- ISA framework
- EuroFIR unit thesaurus
- …

Domain-specific

- EU Menu Methodology/ EFSA data schema for food intake data
- …

' Criteria catalogue'

- Food identification, classification & description
- Units
- Study location
- Socio-demographics
- …

Recommendations

FNS-Cloud platform developers:

- NLP & machine learning
- OXO

FNS-Cloud data providers:

- EU Menu Methodology
- EuroFIR Thesauri
- Dietary assessment software

FNS-Cloud platform users

- Defining research questions
- Searching for data

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