Trends in Food Science & Technology 123 (2022) 310–321

Contents lists available at ScienceDirect

Trends in Food Science & Technology

journal homepage: www.elsevier.com/locate/tifs

Workflow for building interoperable food and nutrition security (FNS) data platforms

Yasmine Emara [a], Barbara Koroušić Seljak [b], Eileen R. Gibney [c], Gorjan Popovski [b,d], Igor Pravst [e], Peter Fantke [a,*]

[a] Quantitative Sustainability Assessment, Department of Environmental and Resource Engineering, Technical University of Denmark, Produktionstorvet 424, 2800, Kgs. Lyngby, Denmark
[b] Computer Systems Department, Jožef Stefan Institute, 1000, Ljubljana, Slovenia
[c] UCD Institute of Food and Health, School of Agriculture and Food Science, University College Dublin, Dublin 4, Ireland
[d] Jožef Stefan International Postgraduate School, Jožef Stefan Institute, 1000, Ljubljana, Slovenia
[e] Nutrition and Public Health Research Group, Nutrition Institute, Trzaska cesta 40, 1000, Ljubljana, Slovenia

ARTICLE INFO

Keywords:
Data integration
Interoperability criteria
FNS-Cloud
Ontology
Machine learning
Natural language processing
Branded food data

ABSTRACT

Background: In response to growing needs for the integration of heterogeneous data on food and nutrition security (FNS), and the current fragmentation of interoperability resources, the 'FNS-Cloud project' aims to develop a cross-domain, interoperable data platform that integrates diverse FNS data. Currently, there is insufficient guidance on how to develop such an FNS data platform and integrate a variety of FNS data types that differ in both their syntax and semantics.

Scope and approach: In the present study, we propose a generalizable workflow to guide data managers in building interoperable, cross-domain FNS data platforms, which centres around the definition of *interoperability criteria* that capture standardized data structures, terminologies and reporting formats for key variables across FNS data types. Information technology tools for automating different workflow steps are discussed. Finally, we include an illustrative case study, where we harmonize and link branded food datasets based on pre-defined interoperability criteria to answer an example research question.

Key findings and conclusions: Our work highlights the unique harmonization requirements within the FNS field. We provide two examples of how generic and domain-specific interoperability criteria addressing these requirements can be defined. Incoming FNS data must comply with defined criteria in order to enable their (semi-) automated integration into any data platform. Our case study reinforces the importance of semantic annotation of FNS data, and the need for clear mapping rules to be included into platform-internal semantic data models. The proposed workflow can be applied to any setting in which data managers strive towards harmonized and linked FNS data, and, thus, promotes an open-data and open-science environment.

1. Introduction

Generation of both scientific and industrial data related to food and nutrition security (FNS) has increased rapidly over the past decades. Simultaneously, growing digitalization of the FNS field has enhanced efforts to harmonize data and link information from heterogeneous data sources to help answer more complex research questions on food, diet and health, while capitalizing on advances in computational sciences for automated data processing and management (Bukhari, Klein, & Baker, 2013; Eftimov, Korošec, & Koroušić Seljak, 2017; Eftimov, Koroušić Seljak, & Korošec, 2017; Sansone et al., 2012).

Given its interdisciplinary nature, the FNS field produces highly diverse data types, from food composition, authenticity, toxicity and sustainability data, to food consumption, behaviour and socioeconomic data, and finally health biomarker and disease outcomes data. Consequently, FNS data differ in the way they are collected, structured, stored, queried, reported (e.g. vocabulary), analysed and visualised (Koroušić Seljak et al., 2018; Muljarto et al., 2017; Sansone et al., 2012).

To enable FNS data interoperability, i.e. the ability of data from different sources to 'communicate' and 'work together', a growing

* Corresponding author.
E-mail addresses: yasem@dtu.dk (Y. Emara), barbara.korousic@ijs.si (B. Koroušić Seljak), eileen.gibney@ucd.ie (E.R. Gibney), gorjan.popovski@ijs.si (G. Popovski), igor.pravst@nutris.org (I. Pravst), pefan@dtu.dk (P. Fantke).

https://doi.org/10.1016/j.tifs.2022.03.022
Received 26 February 2021; Received in revised form 17 March 2022; Accepted 21 March 2022
Available online 24 March 2022

number of efforts have committed themselves to developing various interoperability resources. This includes data standards (e.g. Becker, Unwin, Ireland, & Møller, 2008; FAO/INFOODS, 2012), minimum reporting guidelines (e.g. Field et al., 2008; Pinart et al., 2018; Taylor et al., 2007), data exchange schemas (e.g. EFSA, 2014), domain ontologies or controlled vocabularies (e.g. Dooley et al., 2018; EFSA, 2015), and research infrastructures (e.g. Bogaardt et al., 2018; Poppe, 2019; Rychlik et al., 2018). However, these efforts have evolved largely independently (Zeb, Soininen, & Sozer, 2021), have focused on the collection and/or harmonization of a specific FNS data type (e.g. food composition) or a single FNS domain (e.g. nutrition or exposure assessment), and have predominantly sought the harmonization of metadata (e.g. study descriptors) as opposed to harmonizing the full array of FNS data (e.g. typically measured health biomarkers). As a result, existing FNS interoperability resources are fragmented, lack widespread, cross-domain adoption and suffer from uneven accessibility by different users. The corollary to this fragmentation is that FNS data remain scattered, are not comparable and their meaningful interpretation, integration and (re-)use continue to be among the biggest challenges within the FNS field.

The 'Food and Nutrition Security Data Cloud (FNS-Cloud)' project (https://fns-cloud.eu) aims to make FNS data findable, accessible, interoperable and reusable (FAIR) (Wilkinson et al., 2016) by harmonizing and integrating diverse data linked to FNS into a federated, cross-domain platform that is interoperable with other solutions (e.g. other data repositories). In this context, the FNS-Cloud will offer a number of new ICT services that enable efficient and consistent FNS data curation, processing and annotation, data matching and linking, as well as data analysis and visualization.

To build an interoperable FNS data platform, such as the envisioned FNS-Cloud, platform developers must agree on which data types to include, on relevant harmonization requirements within and across FNS domains, how to address these requirements and how to guarantee platform-internal as well as external (i.e. between the platform and other solutions) interoperability. The FAIR guiding principles (Wilkinson et al., 2016) characterize interoperability as data/metadata using formal, accessible, shared, and broadly applicable language for knowledge representation, using vocabularies that are themselves FAIR, and including qualified references to other (meta)data (principles I1, I2 and I3). As they are intended to provide broadly applicable and general advice, these principles need to be complemented by specific, actionable steps towards achieving FAIR data in practice. In response, several FAIRification workflows have been proposed, which can be applied to any type of data (Beyan et al., 2021; Jacobsen et al., 2020). However, these workflows are generic in nature, and focus on FAIRifying datasets as opposed to providing guidance on how to build entire interoperable data platforms (i.e. platforms that host and consolidate FAIR data from various sources, and that are further able to interoperate with external repositories and other digital infrastructure). In addition to the general FAIR principles, specific steps are hence required to address a number of unique challenges related to integrating data across FNS research/data domains, especially those challenges associated with building cross-domain, centralized and interoperable data platforms (e.g. challenges related to legal/license interoperability given varying FNS data-specific ethical considerations and data protection/copyright requirements).

To address this aspect, we propose a generalizable workflow based on the FAIR guiding principles as a step-by-step action plan for building interoperable FNS data platforms, while considering the unique requirements for harmonizing and linking heterogeneous FNS data within and across FNS research domains. At the centre of our proposed workflow, we introduce the definition of *interoperability criteria* (IC) that formalize platform requirements for consistent file formats, terminology, and reporting formats, thereby turning the FAIR interoperability principles (I1–I3) into practice and leading the way towards (semi-) automated (i.e. involving human verification when required) FNS data integration and joint analysis.

After introducing interoperability challenges within the FNS field, we present our proposed workflow for achieving FNS platform interoperability. We then map a number of automation approaches based on e.g. machine learning (ML) and natural language processing (NLP) to different steps of the workflow and briefly outline their advantages and technological readiness for realizing FNS data normalization and integration. Finally, we present an illustrative case study, where we demonstrate the process of defining and implementing IC to harmonize and link selected heterogeneous FNS datasets and answer an example research question.

2. Interoperability challenges within the FNS field

In a Big Data-driven field such as FNS, the sheer volume of generated data, let alone their structural and semantic diversity, render their integration and linking a laborious, time-consuming and, thus far, predominantly manual endeavour that often requires extensive expert knowledge. To understand these interoperability challenges, we present two illustrative examples: one concerning domain-specific and the other cross-domain data interoperability.

In dietary intervention studies, information on the subjects' blood glucose levels are routinely measured to assess the efficacy of a given treatment or prevention strategy for type-2-diabetes. If such health biomarker data are to be compared or linked across studies or research institutions, the variable 'glucose' (data item) must be reported in a harmonized and standardized way. However, it may be named differently across datasets (e.g. 'blood glucose', 'fasting plasma glucose (FPG)' or 'fasting blood sugar level'), depending on the accuracy of vocabulary used and whether it is linked to a semantic data model (e.g. ontology) or not, as well as on the analytical method used to measure it. From the term 'blood glucose' alone, it is unclear whether this is fasting plasma glucose typically measured in the morning after 10 hours of fasting, or whether this is measured at a random point in time during the day (random glucose). Ambiguously described (non-annotated) data are not human- or machine-readable and cannot be (automatically) integrated into data platforms/cloud infrastructures. Similarly, the data value itself for FPG can be given in mmol/L or mg/dL, or even expressed in categorical terms (normal, prediabetes/insulin resistance or diabetes). Interoperability of glucose measurements generated within the health & nutrition domain requires harmonization at the level of *data item* (i.e. its name, meaning and relationship to other data items), as well as at the level of *data value*. The latter will include e.g. the variable type (e.g. continuous vs. discrete/polychotomous), the data field type (e.g. date, integer, string, Boolean), the value type (e.g. 95^{th} percentile, average) and the value reporting format (e.g. units or allowed standardized text) (Emara & Fantke, 2020; Rocca-Serra et al., 2020).

To illustrate a cross-domain example of FNS data interoperability challenges, exposure and risk assessment of chemical residues, such as pesticides in harvested food crops (Fantke, Friedrich, & Jolliet, 2012; Fantke & Jolliet, 2016), requires a combination of food consumption data and data on chemical concentrations in food. However, while the latter is typically done at the level of raw agricultural commodity (RAC), food consumption data are usually collected at the level of raw or prepared food (i.e. both ingredients and recipes). To be able to use both data types conjointly in exposure assessments, consumption data need to be first converted to the level of edible RAC (e-RAC). This can be done using a RAC conversion database that relies on e.g. recipes to break a cooked meal down into its raw ingredients. To then link consumption at e-RAC to chemical concentrations in RAC as analysed in the laboratory according to rules set in legislation, these chemical concentrations must first be converted to those in the edible part of the food, using processing factors that account for the change in chemical concentration due to processes such as peeling, cooking, washing, and juicing (Boon et al., 2009). Because the data were generated for different purposes and measured following different study protocols, guidelines and legislation,

both these data types must first be converted to one common denominator (e-RAC level) and then analysed for use in exposure and risk assessment. Consequently, for FNS data to be interoperable and reusable across research domains and data types, not only harmonized vocabulary and reporting formats are required, but our semantic resources must be enriched with mapping rules that define conversion/calculation processes (e.g. processing factors) to transform one data item into another.

The two described examples illustrate challenges of *semantic* interoperability, which is concerned with the harmonization of meaning and of the representation of shared data (e.g. units/scale, aggregation level). FNS data integration within and across research domains is additionally hampered by challenges in *syntactic* (structural) interoperability. The extraction of structured data and accurate information from the plethora of unstructured FNS data stored in different file/data formats (e.g. narrative sections in electronic health records or food images uploaded to food consumption monitoring apps) presents unique technical challenges that continue to require new and improved information technology tools to assist in data harmonization and integration (Sinaci et al., 2020).

To enable FNS data interoperability, heterogeneous data – be it metadata or other measured variables – must thus be harmonized in terms of:

(a) data storage/file format (data structure harmonization),
(b) terminology used and its meaning (data item harmonization), and
(c) reporting format for disclosing measured data (data value harmonization.

3. Workflow for building interoperable FNS data platforms

Fig. 1 illustrates the overall proposed workflow. While inspired by previous work on the FAIRification of data, our proposed workflow is especially targeted at guiding data managers towards establishing interoperable, cross-domain FNS data platforms, such as the currently

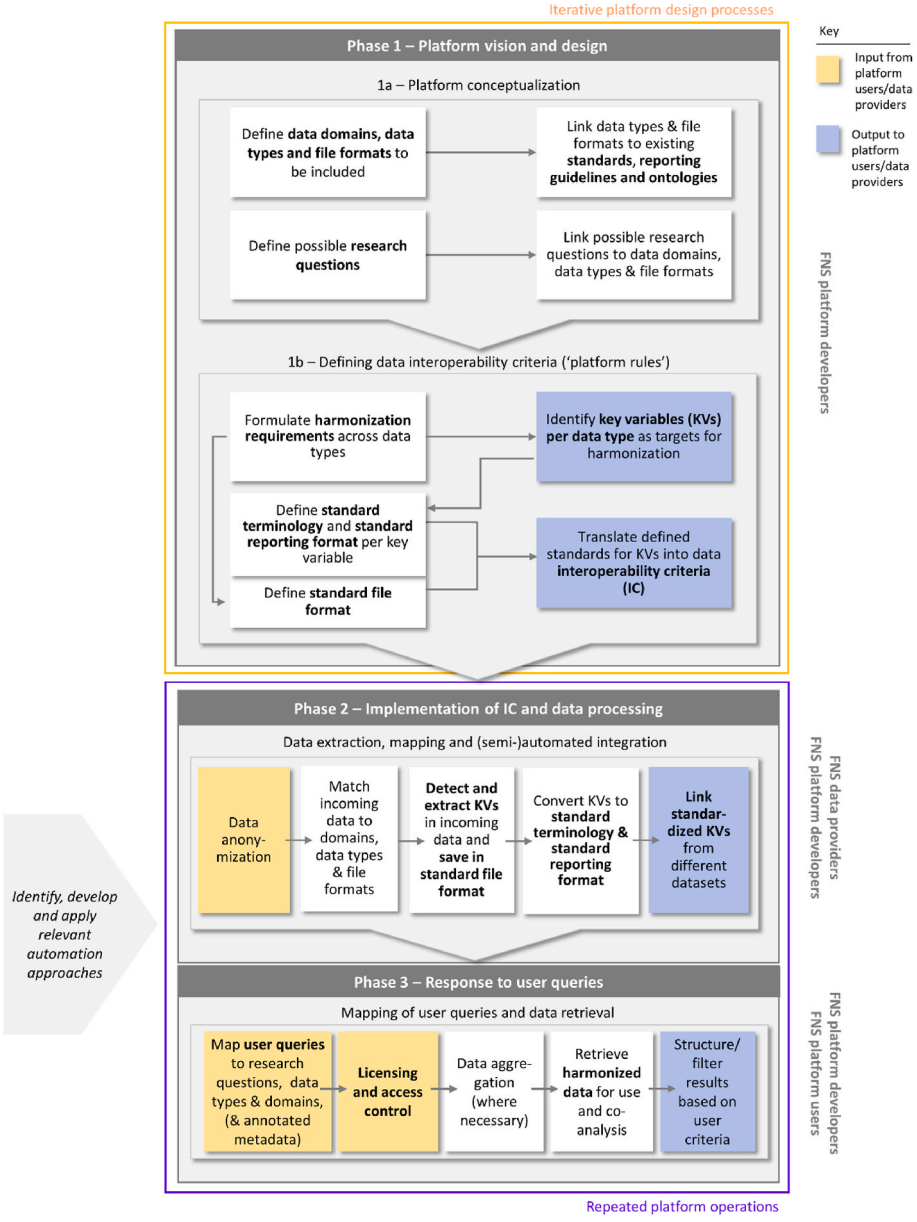

Fig. 1. Workflow for building interoperable FNS data platforms/clouds, integrating and consolidating heterogeneous, cross-domain data based on defined interoperability criteria.

developed FNS-Cloud, integrating and linking diverse FNS data types in an unambiguous, human- and machine-readable format. Our workflow, therefore, includes actionable steps for three consecutive phases in the life cycle of a data platform, namely *Phase 1* – platform vision and design, *Phase 2* – implementation of interoperability criteria and data processing, and *Phase 3* – response to user queries. In each phase, FNS platform developers, FNS data providers or FNS platform users are involved (see Fig. 1).

3.1. Platform vision and design

During platform conceptualization (*Phase 1a* in Fig. 1), we propose starting with the definition of the different FNS *data types* (e.g. food composition, food production, food authenticity, food consumption, sociodemographic, omics and disease outcome data), related FNS *data domains* (e.g. the 'Agri-food' domain, 'Food intake and lifestyle' or 'Health, body function and disease risk') (Presser, Roe, Matuszczak, & Finglas, 2020; Snoek et al., 2018), and respective *file formats* (e.g. plain text, XML, TIFF) that will be included in the platform. Existing data standards, reporting guidelines and domain ontologies (including thesauri) that govern selected FNS data types and stipulate common vocabulary are then researched and linked to respective data type(s) and data domain(s). In parallel, platform developers define possible research questions that typically utilise the included FNS data types. Research questions are then mapped to FNS data domains, data types, and file formats. This step is crucial to help data managers identify those data and variables that are often jointly required to answer FNS-relevant questions and, in turn, where harmonization is needed within and across domains and data types.

Based on this delineation of the FNS field and the insight gained into prominent interoperability resources (e.g. data standards, ontologies) within and across research domains, platform developers make their first informed decision on an Open File format (e.g. XML) that is selected to become the platform-internal standard file format (per data type). Next, a list of *key variables* per data type that extend beyond metadata is defined, and a platform-internal *standard terminology* as well as *standard reporting format* for each key variable are decided upon (Doiron et al., 2013; Pinart et al., 2018). In this context, key variables can be metadata- or provenance data-related items that support correct data interpretation, but should further include (a) all variables that are required to answer FNS-relevant questions as identified in *Phase 1a*, and (b) variables that help link two datasets, such as the food entity, connecting information about the same food item from different datasets. The list of key variables is the first essential output of the this phase of the workflow, as it gives an overview of relevant variables within the different FNS research domains (e.g. nutrition, food authenticity, exposure and risk assessment, sustainability) and their interrelations (e.g. which variables are semantically equivalent but represented differently in different data types/research domains and why), as well as helps focus harmonization efforts in the subsequent steps.

Ideally, selected standard terminologies and standard reporting formats for key variables are drawn from already existing and widely adopted reference standards and ontologies to avoid 're-inventing the wheel' and increase interoperability with external information systems and applications. In this context, ontologies, such as the ONS-ontology (Vitali et al., 2018), the FoodOn ontology (Dooley et al., 2018) or the 'Unified medical language system' ontology (Bodenreider, 2004), which already incorporate unique vocabularies and axiomatic linkages from across different FNS-related data domains, constitute excellent candidates for use in the standardization and structuring of data for achieving cross-domain FNS data interoperability. Additionally, data platform developers can consult a number of initiatives that collect interoperability resources and provide interoperability advice for a variety of (FNS) data types (e.g. the Elixir distributed research infrastructure (Crosswell & Thornton, 2012), the FAIRsharing initiative (S. A. Sansone et al., 2019) and the FAIR Cookbook (Rocca-Serra et al., 2020).

In the next step, standard file formats, standard terminology and standard reporting formats for key variables are *translated into IC*, serving as FNS platform rules. IC provide the basis for the platform's semantic data or representation model, in which the selected key variables across FNS data types and research domains are represented as semantic objects (e.g. macronutrients or nutritional biomarkers). These objects are further connected to a finite set of other semantic objects describing all additional characteristics of the data item (i.e. its semantic context) (Bornhövd, 2000). The semantic context will include all metadata important for interpretation of the semantic object (e.g. which ontology or controlled vocabulary it was adopted from), the unit of measurement, the calculation approach, etc., as well as mapping rules/functions for semantic conversion of different representations of a given variable to the standard target format. Naturally, standard reporting formats and terminologies not adhering to a given data standard or ontology will require new concepts to be defined within the internal platform data model.

IC will include generic (domain-independent) as well as domain-specific criteria. Domain-independent criteria cover the harmonization of generic data descriptors e.g. 'file title', 'data owner', or 'date', whereas domain-specific criteria address data type- or domain-specific variables such as 'meal' and 'physical activity' or 'food product group'. Incoming data must either meet these criteria, enabling their (semi-)automated integration, or they must be reported in any other (standard) format that can be *detected, extracted and converted* into the target format, or otherwise be flagged for manual curation (see Fig. 2).

Fig. 2 presents two examples of IC and how an internal background compatibility check against them could look like as data are being uploaded, (pre-)processed and prepared for uptake into an FNS platform. The more IC are fulfilled by a given dataset, the higher its 'interoperability potential', which can be assessed at the start or end of data processing and curation (e.g. how many data items could be immediately integrated without mapping efforts vs. how many data items await manual curation).

As can be seen in Fig. 2, the successful implementation of IC will depend on the detectability and interpretability of incoming data as well as the ability to map it to the target standard terminology and reporting format. This, in turn, is a function of (1) how complete our platform-internal semantic data model is, (2) how well incoming data are already structured and annotated/already linked to external ontologies, and (3) how well we defined conversion functions and mapping rules in our semantic knowledge network. For example, focusing on the variable 'meal' and presuming it is reported in incoming data according to the EFSA controlled vocabulary as "M3", then this data item has to be correctly annotated (i.e. include this information) so the algorithm can understand the value "M3". If it is reported as < Meal, "12:00 p.m.">, then again the algorithm needs to first be able to know that "12:00 p.m." indicates time in the format <"HH:MM AM/PM"> and then, we would apply underlying mapping rules for how to make sense of a time in terms of 'meal'. For instance, we could define in an internal platform ontology that lunch can be from 11:00 a.m. to 2:00 p.m. and map any meal indicated in time format to the respective EFSA meal code.

The comprehensive list of cross-domain IC constitutes the second essential output of *Phase 1* of our workflow. It provides machine-actionable data dictionaries that specify harmonized data structures (file formats) for different FNS data types, as well as harmonized terminology and reporting formats (data value characteristics) for relevant domain-specific and cross-domain FNS-related variables. IC thus underlie the platform's semantic data model and enable (semi-)automated integration of diverse FNS data into an FNS platform.

Phase 1 is understood as an iterative process. If new FNS data domains and data types emerge, new research questions become relevant or new data standards and ontologies are developed, action steps in *Phase 1* are repeated to support platform adaptability.

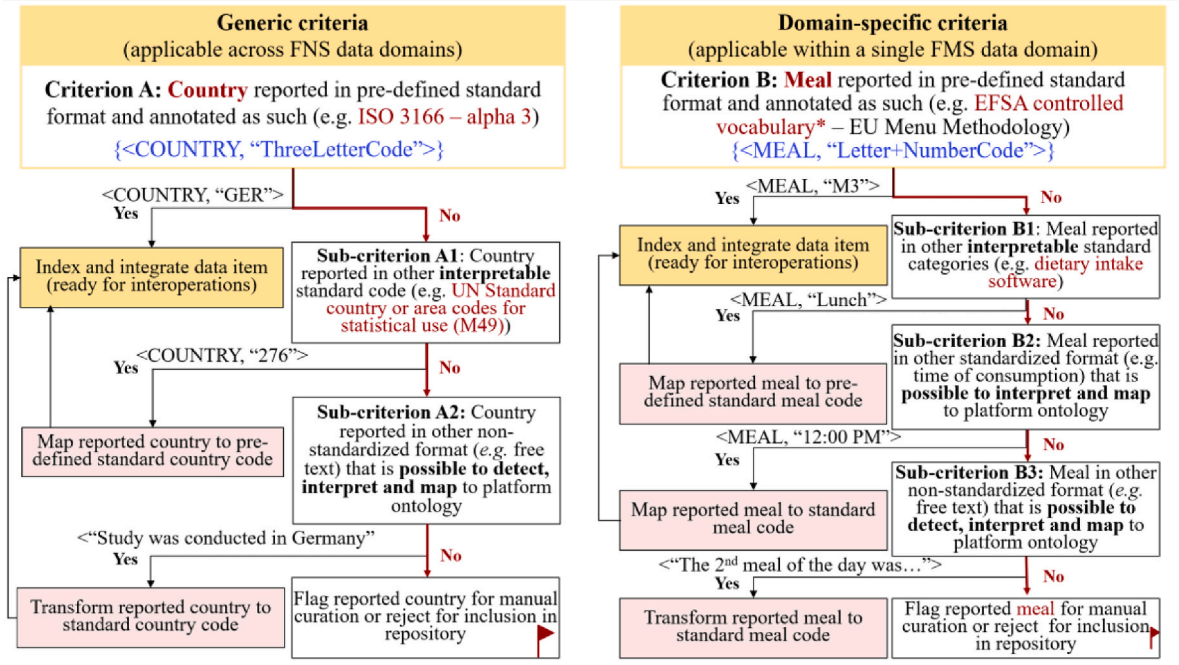

Fig. 2. Examples of generic and domain-specific interoperability criteria (IC), capturing the selected standard reporting format for selected variables. Incoming data are checked against the platform's IC ('background processes') and, if possible, are extracted and converted to target formats, or else flagged for manual curation.

3.2. Implementation of criteria and data processing

In *Phase 2* of the workflow, we move on to repeated platform operations including data pre-processing, extraction and implementation of interoperability criteria.

Phase 2 starts with data anonymization, which is an imperative first step when dealing with a number of sensitive and protected FNS data such as personal patient data, socioeconomic data or proprietary industry data (e.g. sales data) (Pravst, Lavriša, Kušar, Miklavec, & Žmitek, 2017; Sinaci et al., 2020). Under the European General Data Protection Regulation (GDPR), anonymized data are not recognized as 'personal data' and can be used without having to comply with data protection requirements (Regulation (EU) 2016/679), which greatly simplifies subsequent data processing, repurposing and analysis. Naturally, not all FNS data types will require data anonymization. However, which data are sensitive and which data items need anonymization is to be defined by the dataset owner prior to data upload and integration into any FNS platform ('input from platform user' in Fig. 1).

Anonymized data are further processed via matching of data to the respective data type, FNS data domain and file formats to help identify any domain-/data-type specific IC that will apply to the incoming data.

The next step involves the extraction of key variables from incoming data such as food images or scientific articles and conversion of the data on key variables into the target standard machine-readable (i.e. normalized) file format, as well as into the standard terminologies and reporting formats set forward by the pre-defined IC. The implementation of IC, thus, enables the normalization and harmonization of data structures, variable names data values, which allows linking of information from different datasets on semantically-equivalent key variables.

A variety of studies have used different automation approaches to solve a number of problems and challenges in the FNS field such as food recognition, food matching, calories estimation, leftover estimation quality detection and food contamination (e.g. Chin et al., 2019; Eftimov et al., 2017a; Koroušić Seljak et al., 2018; Lamarine, Hager, Saris, Astrup, & Valsesia, 2018). Information technology-based approaches focused specifically on automating data (pre-)processing, normalization and the implementation of IC (i.e. the different steps of *Phase 2* in our proposed workflow) are increasingly being developed with a rise in their adaptation for applications within the FNS field. For example, unstructured text data describing FNS can be processed using text mining and NLP, whilst Deep Learning (DL) methods are typically used to extract information about products from image libraries. Extracted information can then be stored as structured data, matching the defined standard file format, and be used for further application of terminology- and data value-related IC.

Table 1 presents some examples of the automation approaches applied so far within the FNS domain. The rule-based engine 'FoodIE', for instance, which relies on computational linguistics and semantic information (Popovski, Seljak, & Eftimov, 2019a), has been developed to extract and annotate key food- and nutrition-related variables from free-form text, thereby converting unstructured textual data to structured data. Also with structured data, there are problems that may still need solving using computer techniques. For example, to convert inconsistent names of key variables across datasets, lexical and semantic similarity approaches can be applied (Ispirova, Eftimov, Seljak, & Korošec, 2017).

In order to deal with different food description and classification systems, StandFood was developed as an NLP- and ML-based method for assigning FoodEx2 terms to food items using textual data (i.e. names) (Eftimov, Korošec, & Koroušić Seljak, 2017). Another approach based on stochastic optimisation was developed for matching FoodEx2 terms to LanguaL and vice versa (Koroušić Seljak et al., 2018). The method could be upgraded to enable matching of FoodEx2 to e.g. GS1 Global Product Classification and other food description and categorization systems (e.g. Dunford et al., 2012).

Another important resource that can serve in the implementation of IC is the first annotated corpus of FNS entities, FoodBase (Popovski, Seljak, & Eftimov, 2019b). Similarly, FoodOntoMap has been developed as the first method that provides normalization of food concepts to different semantic resources, additionally providing a link between them (Popovski et al., 2019c). FoodOntoMap is designed in a flexible way, enabling its integration with other ontologies such as ONS (Vitali et al., 2018) and ONE (Yang et al., 2019) as well as other systems like FairSpace, which provides a FAIR Virtual Research Environment (the-hyve, 2019).

Another important issue often faced by platform developers and

Table 1
Examples of automation approaches and of their applications mapped against workflow steps, including a short description of the approaches' advantages and technological readiness.

Workflow step	Example automation approaches	Available example tools or applications	Advantages and technological readiness
Match incoming data to domains, data types & file formats	Natural language processing (NLP), computation linguistics	FoodIE (Popovski, Kochev, Seljak, & Eftimov, 2019), FoodOntoMap (Popovski et al., 2019a)	• NER based on a small number of computational linguistics' rules and semantic information about food entities • Linking food concepts across different food ontologies
Detect and extract key variables in incoming data and save in standard file format	Named Entity Recognition (NER)	(Eftimov, Koroušić Seljak, & Korošec, 2017), FoodNER (Stojanov, Popovski, Cenikj, Seljak, & Eftimov, 2021)	• Speed of data processing •Higher accuracy achieved with advanced ML approaches
	Support vector machines (SVMs)	(Anthimopoulos, Gianola, Scarnato, Diem, & Mougiakakou, 2014; Yuan, Holtz, Smith, & Luo, 2016)	•Training datasets required •Time-consuming process of model training •Require as complete semantic resources as possible
	Conditional random fields (CRFs)	Lafferty, McCallum, and Pereira (2001)	•In image recognition: Volume estimation is still an unresolved problem
	Deep neural networks	NutriNet (Mezgec & Koroušić Seljak, 2017), (Kagaya, Aizawa, & Ogawa, 2014)	
Convert key variables to standard terminology & standard reporting formats	NLP, ontologies	StandFood (Eftimov, Korošec, & Koroušić Seljak, 2017), Lamarine et al. (2018)	• Speed of data processing •High possible accuracy •Ability to quantify mapping uncertainty •Require as complete semantic resources as possible
Link standardized key variables from different datasets	Named Entity Linking (NEL)	SAFFRON (Cenikj, International, Eftimov, Koroušić, & Seljak, 2021)	• Allows detecting relations between food and disease entities
Structure/filter results based on user criteria		FoodViz (Stojanov et al., 2020)	• Consistent and comprehensive data structure schemas • Requires as complete semantic resources as possible

ultimately by data users is that some relevant data would be missing in a given dataset. For example, data for specific components or selected foods are often missing from food composition databases or food labels. Imputation methods based on statistics (Ispirova, Eftimov, Korošec, & KoroušićSeljak, 2019) as well as food matching techniques (Eržen, Rayner, & Pravst, 2015; Lamarine et al., 2018) have so far proven successful for filling these data gaps. Similarly, branded/packaged food data often include only data on the macronutrient content of food, while data on micronutrients are generally missing, as they are not a mandatory part of food declaration labelling. However, if some information about the branded food ingredients is available, linear programming can be used to address this issue (Koroušić Seljak et al., 2013; van Dooren, 2018).

By applying different computational automation methods for data (pre-)processing, extraction of key variables and implementation of predefined IC to harmonize data structures, terminologies and reporting formats, *Phase 2* of our workflow prepares various incoming FNS data for data integration/linking, reuse and joint analysis. While automation approaches will continue to play an important role in such FNS data integration, our semantic resources today are far from complete to trust machines learning algorithms to do the work for us (Fantke et al., 2021). In the majority of cases of e.g. food matching, human intervention will still be required (in the foreseeable future) to guarantee accurate data interpretation and integration.

3.3. Response to user queries

The final phase of the workflow is to host the standardized, annotated and linked data on the interoperable data platform, while providing multiple services for data analysis and visualization. Useful conceptual connections between user queries and harmonized data available on an FNS platform will rely on accurate interpretation of user information needs (Humphreys & Lindberg, 1993). An advanced natural-language user interface can help identify key concepts in user queries based on the platform-internal knowledge system and map them to possible research questions, data domains and data types (*Phase 3* in Fig. 1). Automated approaches to achieve such mapping are still at their infancy (see for instance OpenAI, 2020) but are likely to become an effective solution in the future. Key concepts in user queries can additionally be mapped to metadata information (e.g. study design, geographical region/country or study population) to filter for the most conceptually suitable platform data.

FAIR and linked FNS data available on an FNS data platform may still be protected by data licensing agreements with data owners ('input from data provider' in Fig. 1), and may thus be only accessed under specified conditions (e.g. attribution to author) or only upon request. Ethical considerations defined at the time of data collection and consent, along with regulatory aspects of the GDPR, will generally govern data availability and access options. Consequently, a vital step in operating a FAIR FNS data platform, while managing ethical and legal requirements related to the (re-)use of FNS data, pertains to 'licensing and access control'. With each published dataset, clear descriptions of re-use rights and restrictions must be provided (e.g. can the data user copy and redistribute the data, can the data be used commercially), along with security protocols to follow by data user, and where applicable, clear processes to request reuse permission (Cessda Training Team, 2017; Sinaci et al., 2020). Depending on the specific licensing agreement, access control – i.e. the process by which users are authenticated and authorized to use data for which they are granted access rights – is an indispensable component of data security on integrated data platforms. On an FNS data platform, access control may implemented by asking data users to create password-protected accounts and provide access authorization for a given dataset based on multiple user attributes (e.g. user's organizational group or defined purpose for data usage). Additionally, the intended design of an FNS data platform (e.g. pooled vs. federated) will influence the way data are structured and determine whether access to raw vs. aggregated data is provided (Doiron et al., 2013).

In cases where submitted FNS data are not accessible to the public, aggregation of data to provide e.g. study statistics along with publishable metadata becomes a crucial, yet challenging, step. Similar to mapping rules that help transform one data item in a given representation into a platform's target format, aggregation rules for linked FNS data will have to be defined and integrated into a platform's internal knowledge network. For example, food composition and food consumption data may not be publicly accessible at the level of specific foods, but can be aggregated for dissemination at the food group level. A related challenge is to guarantee capturing all relevant information (acknowledgements, collaborations and accessibility) as data are

extracted, integrated, aggregated and retrieved for user reuse and analysis. If data owners alternatively choose to only link to their data on other repositories, our workflow along with a set of defined interoperability criteria for different FNS data types ('criteria catalogue') can still provide guidance to help data owners achieve interoperability of their data with other FNS data deposited on the platform.

After access control and data aggregation, where necessary, available harmonized and machine-readable data relevant for a given user query are retrieved and structured (e.g. shown in rank order according to a specific user criterion). As there are numerous ways of presenting linked data, an interoperable data platform should be designed in an open way, enabling its integration with different outside tools that may further structure, analyse or visualize integrated data and knowledge with respect to specific needs of its end FNS data users.

4. Case study on branded food data

4.1. Example research question and selected datasets

As a novel and central aspect of the proposed workflow towards building interoperable, cross-domain FNS data platforms, we present an illustrative case study to specifically demonstrate the process of defining and implementing IC that help normalize, standardize and link heterogeneous FNS data. In this case study, we harmonize three heterogeneous *branded food datasets* based on predefined IC for selected key variables in order to answer an example research question. The following research question was defined as input for e.g. assessing health impacts associated with food consumption patterns: 'How does the median total sugar content (i.e. sum of naturally occurring and added sugars) of marketed soft drinks and breakfast cereals differ among European countries?'

The two food product categories 'breakfast cereals' and 'soft drinks' were selected as examples, as these constitute common food product groups consumed across Europe and are important contributors to sugar consumption in European diets, linked to multiple health complications (Amoutzopoulos, Steer, Roberts, Collins, & Page, 2020; Ruiz et al., 2017; Zupanič et al., 2020).

Similar to a user query, our research question can be broken down to identify relevant concepts ('total sugar', 'soft drinks' 'breakfast cereals' and 'Europe*') and accordingly find data matched to these aspects (*Phase 3* of our workflow, Fig. 1).

We selected three branded food datasets from Slovenia (SI), the Netherlands (NL) and Switzerland (CH) as illustrative examples. Slovenian data were extracted from the 'Composition and Labelling Information System' (CLAS), created and managed by the Nutrition Institute in Ljubljana (Pravst et al., 2017). Data collection in CLAS started in 2015 from photographs of food labelling and included digital recognition of EAN codes. Currently, the CLAS database contains 10,674 food items with data on e.g. the product name, company, brand, list of ingredients and nutritional values. Dutch data originated from the 'Dutch Branded Food Database' (LEDA), where food label information for approximately 100,000 branded foods (covering ~75% of the Dutch retail food market) have been collected between 2007 and 2017 (Westenbrink, van der Vossen-Wijmenga, Toxopeus, Milder, & Ocké, 2021). Mandatory food label information (e.g. product name, brand, GTIN/EAN code ingredients and macronutrients) is sufficiently covered in the LEDA, whereas non-mandatory information (e.g. micronutrients, dietary claims or portion sizes) is not always available. Finally, Swiss data were derived from the publicly available 'Swiss Food Composition Database' (Version 5.3), which contains data on approximately 10,5000 foods, a subset of which are branded food data (Bieler, 2020).

4.2. Definition of interoperability criteria for selected key variables

To arrive at three harmonized branded food datasets and answer our research question, we followed *Phase 1* and *Phase 2* steps of our proposed workflow which would precede data harmonization and implementation of IC. The selected datasets belong to the data type 'branded food data' as part of the 'Agri-food' domain. We initially researched and mapped branded food data to multiple data standards, reporting guidelines and ontologies (including thesauri) that govern this type of data (*Phase 1a* in Fig. 1). These included for instance the 'Codex Alimentarius General Standard for the Labelling of Prepackaged Food, CXS 1–1985' (Joint FAO/WHO Codex Alimentarius Commission, 2018), the European Food Safety Authority's (EFSA) food classification and description system (FoodEx2) (EFSA, 2015) and the EuroFIR thesauri (Macháčková, Møller, & Ireland, 2017).

To specifically answer the research question, we required data on the total sugar content for all products classified as 'breakfast cereal' or 'soft drink' in the different datasets, standardized and harmonized in terms of used terminology for variable names as well as reporting formats, including unit and matrix unit. This renders the component 'sugar, total' and the 'food product category' our two key variables to define IC that will enable data integration and linking. Additionally, to be able to interpret and subsequently harmonize these key variables based on the defined IC, we required metadata pertaining to (1) the country of origin of the dataset, (2) the food product and product group classification system used, and (3) the units of reporting of the component (sugar, total). With the exception of the Swiss data, which contained information on the units of reporting, the required metadata were either given directly through personal communication with the data owners or searched for in literature published on the respective data sources (i.e. LEDA, CLAS or Swiss Food Composition Database).

We defined the IC to capture the standard terminology and standard reporting format (i.e. target semantic objects and semantic context) for each key variable based on relevant FNS data standards and ontologies (see Table 2). We further defined a criterion relating to the harmonization of the file format (data structure) and chose Office Open XML (OOXML) as the common format to save the data. To take an example from Table 2, the IC for harmonization of the food product category would be:

- The variable name for food product category is "FOOD_CATEGORY",
- The food product category is reported following FoodEx2 classification (EFSA, 2015).

4.3. Implementation of interoperability criteria and case study results

To arrive at harmonized and linked key variables across the three

Table 2
Data standards and ontologies (including thesauri) used to define standard terminology and standard reporting formats for selected key variables in branded food data.

Variable (harmonized property)	Data standard or ontology	Standard terminology or standard reporting format (code, if applicable)
Food product category (variable name)	–	FOOD_CATEGORY
Food product category (data value)	FoodEx2 (EFSA, 2015)	String (e.g. A03DZ for soft drinks)
Component – sugar, total (variable name)	EuroFIR Component Thesaurus (Macháčková et al., 2017)	SUGAR
Unit (variable name)	–	UNIT
Unit component – sugar, total (data value)	EuroFIR unit thesaurus (Macháčková et al., 2017)	Gram (g)
Matrix unit component – sugar, total (data value)	EuroFIR matrix unit thesaurus for unit denominators (Macháčková et al., 2017)	per 100 g total food (T) or per 100 ml food volume (V)

datasets (SI, NL and CH), we implemented the defined IC (i.e. converted the heterogeneous data to the target semantic objects and context). This required a variety of manual case study-specific data wrangling, mapping and conversion (data augmentation) steps, which are presented in Fig. 3, along with IT tools that can be applied at each step to automate the process. The steps presented in Fig. 3 thus provide an example of how the general steps of *Phase 2* of our proposed workflow ('Implementation of IC and data processing' in Fig. 1) can look like in practice, given a set of heterogeneous dataset and a guiding research question.

We started with mapping the selected datasets to FNS data domains (e.g. 'Agri-food' domain) and data types (i.e. branded food data) and file types. This step is crucial in determining which data standards and ontologies apply to this domain and feed into the definition of IC. In the case of the selected datasets, an extra step for translating product names, product categories and ingredients lists (required later for mapping between national food classification systems and FoodEx2) to English had to be added, as both the SI and NL data were available in the respective national languages. Following translation, our key variables

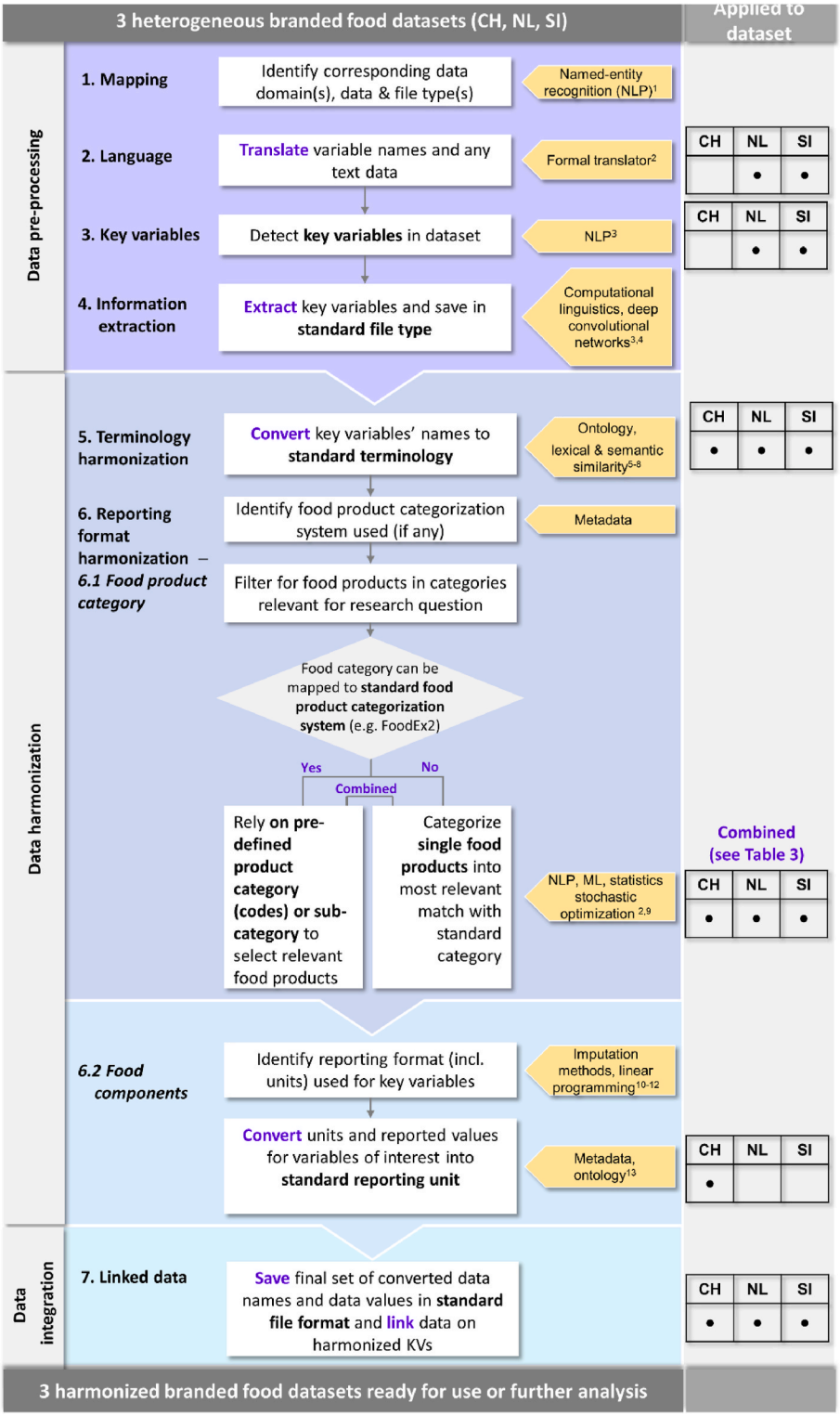

Fig. 3. Data wrangling, mapping and conversion (data augmentation) steps required to get from heterogeneous branded food data from Slovenia (SI), Switzerland (CH) and the Netherlands (NL) to harmonized and linked data on key variables necessary to answer an initially defined research question. [1]Eftimov et al., 2017b, [2]Eftimov et al., 2017a, [3]Popovski et al., 2019a, [4]Mezgec et al., 2019, [5]Vitali et al., 2018, [6]Ispirova et al., 2017, [7-8]Popovski et al., 2019b,c, [9]Koroušić Seljak et al., 2018, [10]Isprova et al., 2019, [11]Koroušić Seljak et al., 2013, p. [12]van van Dooren, 2018, p. [13]Gkoutos, Schofield, & Hoehndorf, 2012.

were detected and data extracted to be saved in the common file format. Given that the data were already provided as tabulated spreadsheets, no conversion into the standard file format was necessary.

In the first step of data harmonization of extracted key variables, we converted variable names to match our standard target terminology. For instance, in CLAS the food category was defined as 'Food product group' and thus converted to 'FOOD_CATEGORY'. Harmonization of the reporting format for the 'FOOD_CATEGORY' was more challenging, as distinct national food classification systems are used in each dataset. This led to a mixed mapping strategy to determine which products matched the two selected food categories 'breakfast cereals' and 'soft drinks' as defined in our standard classification system FoodEx2. Both CH and NL data used database-specific food product categorization systems with no further information currently available on the defined food (product) groups and what they include/exclude (Dutch Nutrition Center, 2020; Federal Food Safety and Veterinary Office, 2020). SI data were coded using the food product classification system proposed by Dunford et al. (2012) for categorization of branded foods. To map reported food product categories to FoodEx2 as our predefined standard food classification and description system, and select only products matching the FoodEx2 definition of 'breakfast cereals' and 'soft drinks',[1] we used the following two approaches (see Fig. 3):

A. Where a direct match between a reported food category and the selected FoodEx2 food categories was possible, this allocation was adopted.
B. Where a direct match between reported food category and the selected FoodEx2 food categories was not possible, manual mapping of single products to FoodEx2 food categories was performed. For this kind of manual mapping, we relied on translations of food product names and of food descriptions included in the datasets, as well as on web searches (if available, using EAN or GTIN code), on brand name and on analysing the ingredients list.

Table 3 shows examples of applying approach (A) or approach (B) to reported food items in the different datasets. In the case of CH and NL data, the main food groups incorporated several categories (e.g. 'soft-drinks, fruitjuice and lemonade' in NL data and 'bread, flakes and breakfast cereals' in CH data). In some cases, the reported food sub-categories (e.g. 'softdrinks, fruitjuice and lemonade – soft drink' or 'softdrinks, fruitjuice and lemonade – smoothie' in NL data) provided a direct match with FoodEX2 categories and made the inclusion (– soft drink) or exclusion (– smoothie) of whole food sub-categories possible by following approach (A).

The final step of data harmonization was to apply the defined IC related to converting units and reported sugar content. Although this information was missing in the datasets, the sugar content in both SI and NL data was given in the desired standard unit (conclusion based on literature/communication with data owners). Therefore, no further conversion of the reported sugar content values was necessary. In the CH data, the unit for breakfast cereals was gram per 100 g *edible* product. Assuming that there is no difference between the food product and its edible fraction in breakfast cereals, we considered the units in the CH data to be semantically equivalent to our standard reporting unit 'g/100 g T'. We added the variable 'UNIT' to the otherwise harmonized datasets and indicated it to be in 'g/100 g T' or 'g/100 ml V'. This concluded our standardization and harmonization efforts for key variables. Linking of data on semantically equivalent food product groups (following FoodEx2 classification) and in harmonized units is now possible and enables us to calculate average sugar content per product group in each of the considered European countries or across all three datasets.

Branded food data used in our case study presents a good example of data privacy issues within the FNS domain. For example, due to the license agreements with data providers, the Dutch LEDA database, which

Table 3
Examples of manual mapping of single food products in the three selected branded food datasets (Switzerland, The Netherlands, Slovenia) to the desired standard food categories 'breakfast cereals' or 'soft drinks' based on FoodEx2 definitions.

Data-set	Reported product name	Reported food category/ food sub-category	Applied approach for mapping (A. or B.)	Include/ Exclude	Decision based on
CH	Actilife Crunchy Mix Fibre (Migros)	Bread, flakes and breakfast cereals/ Muesli mixes and breakfast cereals	A.	Include	Reported food sub-category
CH	Biotta Getränk Bio-Energy (Biotta AG)	Non-alcoholic beverages/ Soft drinks	B.	Exclude	Web search → energy drink
NL	Siroop grenadine	Softdrinks, fruitjuice and lemonade/ syrup, rose water, lemonade, …	B.	Exclude	Product name in Dutch ('siroop') → Syrups/ Cordials
NL	Aloe vera natural	softdrinks, fruitjuice and lemonade/ fresh fruit drink and fruit water	B.	Exclude	Ingredients list → fruit content >25%
SI	Brezalkoholna negazirana pijača brez energijske vrednosti z okusom pomaranče in L-Karnitinom, s sladilom	Soft drinks/ n.a.	B.	Exclude	Product name in English ('Non-alcoholic energy-free drink with orange and L-Carnitine flavour, with sweetening matter') → Isotonic drink/sports drink
SI	Brezalkoholna negazirana pijača iz limoninega soka, z limonino pulpo	Soft drinks/ n.a.	B.	Include	Product name in English ('Non-alcoholic lemon juice drink with lemon pulp') and ingredients list (fruit content = 15%)

[1] According to FoodEx2 'breakfast cereals' include all cereal-based derivatives or products intended to be consumed mostly at breakfast. This can be rolled grains and porridge (to be diluted or ready to eat), cereal bars, muesli and mixed breakfast cereals. 'Soft drinks' include any type of soft drink with minor amounts of fruit (below the minimum for nectars, i.e. 25%) or flavours and sweeting ingredients. These drinks are mostly carbonated. Functional drinks (e.g. isotonic/sports drinks, energy drinks) and cordials are not considered part of the 'soft drinks' category.

is hosted by the Netherlands Food Information Resource (NethFIR)/ Institute for Public Health and the Environment (RIVM), can only be used by these institutions for nutritional research and consumer information (Westenbrink et al., 2021). The provision of data for purposes of our case study was granted under the condition of not presenting any product-/brand-specific data.

Fig. 4 shows the outcome of the mapping between reported food categories and selected FoodEx2 categories (i.e. number of finally included records) following approaches (A) and (B), as well as case study results that were calculated based on the three harmonized datasets. As some extracted data records had missing values (e.g. 'NA') in the associated 'Sugar' column (13 records in SI data, 3 records in NL data, 2 records in CH data), a post-processing step was necessary to remove such records, leading to an additional reduction of the selected records for analysis. There are different reasons for the missing information in the datasets. For example, the SI dataset originates from a food labelling monitoring study that was conducted in 2017 right after mandatory labelling of nutrition declaration was introduced in the EU in December 2016. At the time, marketing of old products without labelled nutrition declaration was allowed until their stocks are exhausted. Our case study reveals that median total sugar content in soft drinks lies at 7.8, 4.6 and 7 g/100 ml food volume, while for breakfast cereals is 18.3, 13.5 and 19 g/100 g total food in SI, NL and CH, respectively. In both the SI and NL data, the maximum sugar content was an outlier, amounting to 34 g/100 ml for a 'Strawberry flavoured spray drink' in the SI dataset, and 25 g/100 ml for an orange-flavoured soft drink in the NL data.

The analysis of the case study results has shown that some post-processing steps from the user may be necessary to make proper use of the data, for example, understanding the distribution of any missing data across categories and datasets or determining outliers (as identified in this analysis) and considering their inclusion/exclusion in the data analysis. Our workflow thus does not replace close data analysis and a data quality check, nor does it consider the correct interpretation and use of analysis results. It is nonetheless suitable for building interoperable data platforms that integrate and link heterogeneous FNS data for joint use and co-analysis.

The challenges faced during the case study further reinforce the importance of semantic annotation of incoming data (enrichment with metadata) to allow (semi-)automated data extraction, interpretation and harmonization. They further emphasize the need for clear mapping rules and semantic conversion functions to be included into a platform's internal semantic data model in order to allow integration of data represented in different formats and terminologies. Platform developers will therefore have to strive for as complete a semantic model as possible, which is likely to take years to keep up with the growing FNS field.

5. Conclusions

We propose a comprehensive and generalizable workflow that serves as a step-by-step guide ('blue-print') for designing interoperable, cross-domain FNS data platforms and clouds that integrate various FNS data types in a human- and machine-readable format to make them findable, accessible, interoperable and reusable. The workflow can be applied to any setting, where data managers are seeking to integrate and link diverse FNS datasets into a centralized data repository or cloud infrastructure to increase data reuse and joint analysis. While inspired by previous work on the FAIRification of data, guidance on the definition and implementation of generic and cross-domain interoperability criteria that formalize the conditions under which FNS data are interoperable constitutes the main novelty of our workflow. IC address the harmonization of data structures, terminologies and data reporting formats (semantic and syntactic interoperability) for predefined lists of FNS-relevant key variables that extend beyond metadata. Our workflow further includes unique platform design steps (*Phase 1*) that involve e.g. mapping the whole FNS field in preparation for defining IC, as well as platform operation processes (*Phase 3*), which pertain to processing user queries (e.g. via NLP technologies) and returning harmonized, consistently annotated FNS data. Flexibly adaptable to the conditions of any given platform, IC and standard workflow steps will not only guide data managers during data harmonization and integration, but, if published as e.g. 'criteria catalogue', they can also be used by data providers to generate new data that are 'FAIR-by-design' (Jacobsen et al., 2020).

Emerging information technology tools were mapped to workflow steps to demonstrate the rising role of e.g. deep learning and NLP technologies in data processing, annotation and linking within the FNS field, while highlighting the focal points at which new IT-based solutions are still needed. Our case study demonstrates that there are substantial challenges when implementing IC, further emphasizing the need for as complete a semantic model as possible for the FNS field to enable cross-domain data integration and analysis.

Our workflow tackles a key challenge for advancing the FNS field by

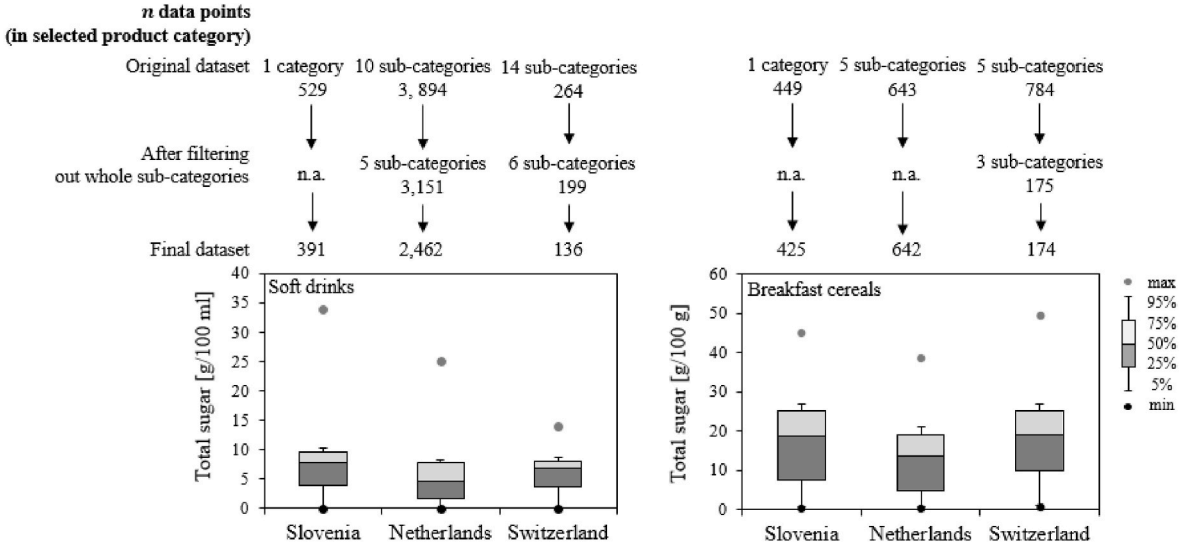

Fig. 4. Final outcome of mapping reported food categories or single food products to the FoodEx2 categories 'soft drinks' (left) and 'breakfast cereals' (right), and results of the case study for the distribution of the total sugar content in selected products in Slovenia, the Netherlands and Switzerland following dataset harmonization and integration according to the proposed workflow.

guiding data managers and researchers towards building cross-domain, interoperable FNS data platforms that overcome the current fragmentation of FNS interoperability resources and FNS data to enable joint data analysis and interpretation. This, in turn, will help address more complex research questions around food and nutrition security and contributes to a more open data and open science environment.

Acknowledgements

We thank Susanne Westenbrink (RIVM, The Netherlands), Karl Presser & Agnieszka Matuszczak (Premotec GmbH, Switzerland), and Anita Kušar & Katja Žmitek (Nutrition Institute, Slovenia) for providing access to datasets used in the case study. This work was undertaken within the FNS-Cloud ('Food Nutrition Security Cloud') project (www.fns-cloud.eu), which 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.

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

Anthimopoulos, M. M., Gianola, L., Scarnato, L., Diem, P., & Mougiakakou, S. G. (2014). A food recognition system for diabetic patients based on an optimized bag-of-features model. *IEEE Journal of Biomedical and Health Informatics, 18*(4), 1261–1271. https://doi.org/10.1109/JBHI.2014.2308928

Becker, W., Unwin, I., Ireland, J., & Møller, A. (2008). Proposal for structure and detail of a EuroFIR standard on food composition data II. Technical Annex. In *EuroFIR technical report. The EuroFIR consortium*.

Beyan, O., Emam, I., Rocca-Serra, P., Sansone, S.-A., Juty, N., Alharbi, E., et al. (2021). *FAIRification of IMI and EFPIA data WP2 – standards definition and process development. D2 . 5 FAIRplus FAIR data maturity framework*.

Bieler, S. (2020). *Schweizer Nährwerttabelle (3. Ausgabe). Bundesamt für Lebensmittelsicherheit und Veterinärwesen BLV*.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research, 32*, 267–270. https://doi.org/10.1093/nar/gkh061. DATABASE ISS.).

Bogaardt, M. J., Geelen, A., Zimmermann, K., Finglas, P., Raats, M. M., Mikkelsen, B. E., et al. (2018). Designing a research infrastructure on dietary intake and its determinants. *Nutrition Bulletin, 43*(3), 301–309. https://doi.org/10.1111/nbu.12342

Boon, P. E., Ruprich, J., Petersen, A., Moussavian, S., Debegnach, F., & van Klaveren, J. D. (2009). Harmonisation of food consumption data format for dietary exposure assessments of chemicals analysed in raw agricultural commodities. *Food and Chemical Toxicology, 47*, 2883–2889. https://doi.org/10.1016/J.FCT.2009.08.003

Bornhövd, C. (2000). Semantic metadata for the integration of web-based data for electronic commerce. In *Proceedings of international workshop on advance issues of E-commerce and web-based information systems*. Cat. *No.PR00334)*.

Bukhari, A. C., Klein, A., & Baker, C. J. O. (2013). Towards interoperable BioNLP semantic web services using the SADI framework. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 7970 LNBI* (pp. 69–80). https://doi.org/10.1007/978-3-642-39437-9_6

Cenikj, G., International, J. S., Eftimov, T., Koroušić, B., & Seljak, K. (2021). *Saffron: transfer learning for food-disease relation extraction* (pp. 30–40). https://doi.org/10.18653/V1/2021.BIONLP-1.4

Cessda Training Team. (2017). *CESSDA data management expert guide*. https://www.cessda.eu/DMGuide.

Chin, E. L., Simmons, G., Bouzid, Y. Y., Kan, A., Burnett, D. J., Tagkopoulos, I., et al. (2019). Nutrient estimation from 24-hour food recalls using machine learning and database mapping: A case study with lactose. *Nutrients, 11*(12), 3045. https://doi.org/10.3390/NU11123045, 11.

Crosswell, L. C., & Thornton, J. M. (2012). Elixir: A distributed infrastructure for European biological data. *Trends in Biotechnology, 30*(5), 241–242. https://doi.org/10.1016/j.tibtech.2012.02.002. Trends Biotechnol.

Doiron, D., Burton, P., Marcon, Y., Gaye, A., Wolffenbuttel, B. H. R., Perola, M., et al. (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., et al. (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

van Dooren, C. (2018). A review of the use of linear programming to optimize diets, nutritiously, economically and environmentally. *Frontiers in Nutrition, 5*. https://doi.org/10.3389/fnut.2018.00048

Dunford, E., Webster, J., Metzler, A. B., Czernichow, S., Mhurchu, C. N., Wolmarans, P., et al. (2012). International collaborative project to compare and monitor the nutritional composition of processed foods. *European Journal of Preventive Cardiology, 19*(6), 1326–1332. https://doi.org/10.1177/1741826711425777

Dutch Nutrition Center. (2020). *Levensmiddelendatabank*. https://www.voedingscentrum.nl/levensmiddelendatabank.

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, 2015*Wiley. https://doi.org/10.2903/sp.efsa.2015.en-804. EN-80).

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

Eftimov, T., Koroušić Seljak, B., & Korošec, P. (2017). A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One, 12*(6), Article e0179488. https://doi.org/10.1371/journal.pone.0179488

Emara, Y., & Fantke, P. (2020). *Food nutrition security cloud. Deliverable 3.1 data requirements and applicability criteria* (Issue Project Number: 863059).

Eržen, N., Rayner, M., & Pravst, I. (2015). A comparative evaluation of the use of a food composition database and nutrition declarations for nutrient profiling. *Journal of Food and Nutrition Research, 54*(2), 93–100.

Fantke, P., Cinquemani, C., Yaseneva, P., De Mello, J., Schwabe, H., Ebeling, B., et al. (2021). Transition to sustainable chemistry through digitalisation. *Inside Cosmetics, 7*, 2866–2882. https://doi.org/10.1016/j.chempr.2021.09.012

Fantke, P., Friedrich, R., & Jolliet, O. (2012). Health impact and damage cost assessment of pesticides in Europe. *Environment International, 49*, 9–17. https://doi.org/10.1016/j.envint.2012.08.001

Fantke, P., & Jolliet, O. (2016). Life cycle human health impacts of 875 pesticides. *International Journal of Life Cycle Assessment, 21*, 722–733. https://doi.org/10.1007/s11367-015-0910-y

FAO/INFOODS. (2012). *Guidelines for converting units, denominators and expressions. version 1.0*.

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

Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology, 26*, 541–547. https://doi.org/10.1038/nbt1360

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

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.

Ispirova, G., Eftimov, T., Korošec, P., & Koroušić Seljak, B. (2019). Might: Statistical methodology for missing-data imputation in food composition databases. *Applied Sciences, 9*(19), 4111. https://doi.org/10.3390/app9194111

Ispirova, G., Eftimov, T., Seljak, B. K., & Korošec, P. (2017). Mapping food composition data from various data sources to a domain-specific ontology. *IC3K 2017 - Proceedings of the 9th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, 2*, 203–210. https://doi.org/10.5220/0006504302030210

Jacobsen, A., Kaliyaperumal, R., Santos, L. O. B. da S., Mons, B., Schultes, E., Roos, M., et al. (2020). A generic workflow for the data FAIRification process. *Data Intelligence, 2*, 56–65. https://doi.org/10.1162/DINT_A_00028

Kagaya, H., Aizawa, K., & Ogawa, M. (2014). *Food detection and recognition using convolutional neural network*. https://doi.org/10.1145/2647868.2654970. *MM 2014 - Proceedings of the 2014 ACM Conference on Multimedia*, 1085–1088.

Koroušić Seljak, B., Korošec, P., Eftimov, T., Ocke, M., van der Laan, J., Roe, M., et al. (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

Koroušić Seljak, B., Stibilj, V., Pograjc, L., Mis, N. F., & Benedik, E. (2013). Food composition databases for effective quality nutritional care. *Food Chemistry, 140*(3), 553–561. https://doi.org/10.1016/j.foodchem.2013.02.061

Lafferty, J., McCallum, A., & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th international conference on machine learning 2001. ICML 2001)*, 282–289 http://portal.acm.org/citation.cfm?id=655813.

Lamarine, M., Hager, J., Saris, W. H. M., Astrup, A., & Valsesia, A. (2018). Fast and accurate approaches for large-scale, automated mapping of food diaries on food composition tables. *Frontiers in Nutrition, 5*, 38. https://doi.org/10.3389/fnut.2018.00038

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.

Mezgec, S., & Koroušić Seljak, B. (2017). NutriNet: A deep learning food and drink image recognition system for dietary assessment. *Nutrients, 9*(7), 657. https://doi.org/10.3390/nu9070657

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

OpenAI. (2020). *OpenAI technology, just an*. HTTPS call away https://beta.openai.com/?demo=1.

Pinart, M., Nimptsch, K., Bouwman, J., Dragsted, L. O., Yang, C., De Cock, N., et al. (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

Popovski, G., Kochev, S., Seljak, B., & Eftimov, T. (2019). FoodIE: A rule-based named-entity recognition method for food information extraction. *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods*, 915–922. https://doi.org/10.5220/0007686309150922

Popovski, G., Seljak, B., & Eftimov, T. (2019a). FoodOntoMap: Linking food concepts across different food ontologies. In *Proceedings of the 11th international joint conference on knowledge discovery, knowledge engineering and knowledge management* (pp. 195–202). https://doi.org/10.5220/0008353201950202

Popovski, G., Seljak, B. K., & Eftimov, T. (2019b). FoodBase corpus: A new resource of annotated food entities. *Database*, (1), 1–13. https://doi.org/10.1093/database/baz121, 2019.

Poppe, K. J. (2019). *Towards a European food , nutrition and health research infrastructure – organisational aspects*.

Pravst, I., Lavriša, Ž., Kušar, A., Miklavec, K., & Žmitek, K. (2017). Changes in average sodium content of prepacked foods in Slovenia during 2011–2015. *Nutrients, 9*(9), 952. https://doi.org/10.3390/nu9090952

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). of the European Parliament and of the Council of27 April 2016 on the protection ofnatural persons with regard to the processing of personal data and on the free move- ment of such data, and repealing Directive 95/46/EC (General Da, Pub. L. No. 2016/679 OF (2016)*. https://doi.org/10.1308/rcsfdj.2018.54

Rocca-Serra, P., Giessmann, R., Splendiani, A., Boiten, J. W., Courtot, M., Burdett, T., et al. (2020). *D2.4 FAIR Cookbook - PublicVersion*. https://doi.org/10.5281/zenodo.3924596. Zenodo Version 1.0. .

Ruiz, E., Rodriguez, P., Valero, T., Ávila, J., Aranceta-Bartrina, J., Gil, Á., et al. (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

Rychlik, M., Zappa, G., Añorga, L., Belc, N., Castanheira, I., Donard, O. F. X., et al. (2018). Ensuring food integrity by metrology and FAIR data principles. *Frontiers of Chemistry, 6*, 1–7. https://doi.org/10.3389/fchem.2018.00049. MAY.

Sansone, S. A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A. L., et al. (2019). FAIRsharing as a community approach to standards, repositories and policies. *Nature Biotechnology, 37*(4), 358–367. https://doi.org/10.1038/s41587-019-0080-8, 2019 37:4.

Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., et al. (2012). Toward interoperable bioscience data. In *, Vol. 44. nature genetics* (pp. 121–126). Nature Publishing Group. https://doi.org/10.1038/ng.1054.

Sinaci, A. A., Núñez-Benjumea, F. J., Gencturk, M., Jauer, M. L., Deserno, T., Chronaki, C., et al. (2020). From raw data to FAIR data: The FAIRification workflow for health research. *Methods of Information in Medicine, 59*(S 01), E21–E32. https://doi.org/10.1055/S-0040-1713684

Snoek, H. M., Eijssen, L. M. T., Geurts, M., Vors, C., Brown, K. A., Bogaardt, M. J., et al. (2018). Advancing food, nutrition, and health research in Europe by connecting and building research infrastructures in a DISH-RI: Results of the EuroDISH project. In *, Vol. 73. Trends in food science and technology* (pp. 58–66). Elsevier Ltd. https://doi.org/10.1016/j.tifs.2017.12.015.

Stojanov, R., Popovski, G., Cenikj, G., Seljak, B. K., & Eftimov, T. (2021). A fine-tuned bidirectional encoder representations from transformers model for food named-entity recognition: Algorithm development and validation. *Journal of Medical Internet Research, 23*(8). https://doi.org/10.2196/28229

Stojanov, R., Popovski, G., Jofce, N., Trajanov, D., Seljak, B. K., & Eftimov, T. (2020). FoodViz: Visualization of food entities linked across different standards. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 12566 LNCS* (pp. 28–38). https://doi.org/10.1007/978-3-030-64580-9_4

Taylor, C. F., Paton, N. W., Lilley, K. S., et al. (2019). The minimum information about a proteomics experiment (MIAPE). *Nature Biotechnology, 25*(8), 887–893. https://doi.org/10.1038/nbt1329 thehyve, 2007.

thehyve. (2019). FAIR Virtual Research Environment. *A use case of Fairspace for Institut Curie*. https://thehyve.nl/cases/fair-vre-institut-curie/.

Vitali, F., Lombardo, R., Rivero, D., Mattivi, F., Franceschi, P., Bordoni, A., et al. (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., van der Vossen-Wijmenga, W., Toxopeus, I., Milder, I., & Ocké, M. (2021). LEDA, the branded food database in The Netherlands: Data challenges and opportunities. *Journal of Food Composition and Analysis, 102*, Article 104044. https://doi.org/10.1016/J.JFCA.2021.104044

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (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., et al. (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

Yuan, J., Holtz, C., Smith, T., & Luo, J. (2016). Autism spectrum disorder detection from semi-structured and unstructured medical data. *EURASIP Journal on Bioinformatics and Systems Biology*, (1), 1–9. https://doi.org/10.1186/S13637-017-0057-1/TABLES/4, 2017.

Zeb, A., Soininen, J.-P., & Sozer, N. (2021). Data harmonisation as a key to enable digitalisation of the food sector: A review. *Food and Bioproducts Processing, 127*, 360–370. https://doi.org/10.1016/J.FBP.2021.02.005

Zupanič, N., Hristov, H., Gregorič, M., Blaznik, U., Delfar, N., Koroušić Seljak, B., et al. (2020). Total and free sugars consumption in a slovenian population representative sample. *Nutrients, 12*(6), 1729. https://doi.org/10.3390/nu12061729