

## VTT Technical Research Centre of Finland

### Data harmonisation as a key to enable digitalisation of the food sector

Zeb, Akhtar; Soininen, Juha-Pekka; Sözer, Nesli

*Published in:*  
Food and Bioproducts Processing

*DOI:*  
[10.1016/j.fbp.2021.02.005](https://doi.org/10.1016/j.fbp.2021.02.005)

Published: 01/01/2021

*Document Version*  
Publisher's final version

*License*  
CC BY

[Link to publication](#)

*Please cite the original version:*

Zeb, A., Soininen, J-P., & Sözer, 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>



VTT  
<http://www.vtt.fi>  
P.O. box 1000FI-02044 VTT  
Finland

By using VTT's Research Information Portal you are bound by the following Terms & Conditions.

I have read and I understand the following statement:

This document is protected by copyright and other intellectual property rights, and duplication or sale of all or part of any of this document is not permitted, except duplication for research use or educational purposes in electronic or print form. You must obtain permission for any other use. Electronic or print copies may not be offered for sale.



Contents lists available at ScienceDirect

## Food and Bioproducts Processing

journal homepage: [www.elsevier.com/locate/fbp](http://www.elsevier.com/locate/fbp)

IChemE ADVANCING CHEMICAL ENGINEERING WORLDWIDE



# Data harmonisation as a key to enable digitalisation of the food sector: A review

Akhtar Zeb\*, Juha-Pekka Soininen, Nesli Sozer

VTT Technical Research Centre of Finland Ltd., P.O. Box 1000, FI-02044 VTT, Finland

## ARTICLE INFO

## Article history:

Received 29 October 2020

Received in revised form 19 January 2021

Accepted 9 February 2021

Available online 17 February 2021

## Keywords:

Food sector

Data harmonisation

Food ontology

Harmonisation challenges

## ABSTRACT

The food sector is driven by a large number of actors, including primary producers, manufacturers, logistics providers, retailers, and consumers. At each phase of the food value chain, a significant amount of data is generated that provides important information to the agents involved in processing and flow of food products from farm to fork. Proper handling of food data has a crucial role in providing safe, quality and affordable products to the increasing world population. The independent production of food data, without following any specific guidelines and procedures, often results in inconsistent and incomparable datasets that cannot be directly utilised by multiple users. Data harmonisation means reconciling various types, levels and sources of data in formats that are compatible and comparable, and thus useful for better decision making. In the food sector, one way of performing data harmonisation is to represent food data according to reliable classification and description systems. Another approach towards harmonisation is to match various food concepts to the existing and widely used ontologies. Furthermore, harmonisation is facilitated by following specific guidelines and procedures during data collection processes. This study explores some of the most important tools, frameworks and methodologies for data harmonisation in the food sector.

© 2021 The Author(s). Published by Elsevier B.V. on behalf of Institution of Chemical Engineers. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The food sector is under active digitalisation due to the implementation of advanced technologies (e.g. sensors, robots, drones) for efficient utilisation of natural resources and providing safe, nutritious, affordable and tasty food to the increasing world population. Food sector is complex and includes various actors in the value chain, for example; primary producers, breeding companies, ingredient and food producers, packaging companies, retailers, food services and catering industry, logistics providers, and consumers (Fig. 1). At each phase of the food value chain, a significant amount of data is generated that provides important information to the actors involved in processing and flow of food products from farm to fork. The primary producers plan their produc-

tion activities by analysing the demand, market value and weather data from different sources. They also create data for their own needs as well as for other partners by collecting information on, e.g., harvesting time and conditions, application of machinery, nature and type of products, feed, use of fertilisers and pesticides, packaging, storage, delivery time and method. Food manufacturers utilise raw material composition and origin data not only as an input for product development but also during marketing. Food manufacturing operations generate various levels of data e.g. storage conditions of raw materials and ingredients, preliminary operations (e.g. washing, cutting, mixing), transformation operations (e.g. extraction, heating, drying), packaging, food composition, nutrient content and shelf-life. Data collected from various sensors measuring temperature, humidity, chemical and bio-

\* Corresponding author.

E-mail addresses: [akhtar.zeb@vtt.fi](mailto:akhtar.zeb@vtt.fi) (A. Zeb), [Juha-Pekka.Soininen@vtt.fi](mailto:Juha-Pekka.Soininen@vtt.fi) (J.-P. Soininen), [nesli.sozer@vtt.fi](mailto:nesli.sozer@vtt.fi) (N. Sozer).  
<https://doi.org/10.1016/j.fbp.2021.02.005>

0960-3085/© 2021 The Author(s). Published by Elsevier B.V. on behalf of Institution of Chemical Engineers. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

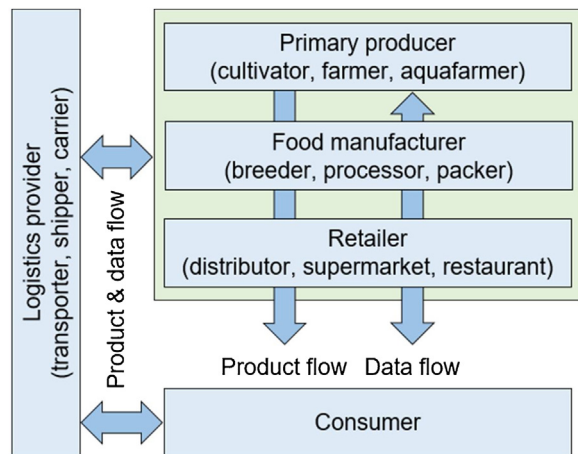


Fig. 1 – Food value chain.

chemical components are used to predict actual shelf-life, microbiological stability of food products or ripeness of fruits and vegetables during logistics and distribution. Logistic and supply delivery chain data contains also information about the mean and mode of transportation and delivery time. Different sensor scans, radio frequency identification (RFID), smart tags (near field communication (NFC), QR codes) and barcode technologies are used to link the products to services that enhance consumers experience and provide solutions to minimise food consumption waste. Moreover, consumers demand product information related to nutritional composition, allergens, origin and authenticity while they also generate valuable information for the retailers and food companies through their consumption patterns.

Besides, national food composition, consumption and export data are essential to assess diet quality of populations, need or success of food-based dietary guidelines, providing a useful tool for public health and nutrition. It is used by many actors, including dieticians, epidemiologists, researchers, health educationalists, decision-makers, and the food industry. It is essential to assess food-oriented risks (exposure to harmful substances and pathogenic microorganisms), and nutrient intake for early recognition of diet-related diseases (Pakkala et al., 2014), (Szűcs et al., 2013). In addition, food consumption datasets when combined with other datasets, such as food composition data and health records, can enable thorough investigations.

### 1.1. Data driven food manufacturing system

The modern food industry requires data from multiple sources for product development and formulation, labelling and to comply with national and international standards and regulations (Finglas et al., 2014), (Williamson, 2005). In order to stay competitive in the food market, manufacturers and retailers need to modify their products based on consumer needs and preferences. When developing new products, the food manufacturers require both market and consumer data, in addition to the ingredient and recipe data. The composition data is also needed for reformulating existing products either to improve their, e.g. taste, or to meet certain nutritional criteria, e.g. requirements for fat, sugar and salt content (Black et al., 2011). Food data is also required for understanding the effects of new food processing technologies on the nutritional quality of food products. Besides, information on the nutritional composition of foods in the form of nutrition and health

claims (e.g. 'low in salt', 'reduced fat', 'no added sugar', 'high in fibre') are exploited by the food manufacturers and retailers for marketing and promotion purposes in accordance with consumer preferences and needs.

Similarly, access to up-to-date information on food consumption and demand is important for the producers to understand the market needs and adjust their production activities accordingly. The food demand and consumption are associated with price, household income, as well as socio-economic and geographic factors. Information about food consumption and food demand, both at the individual and population levels (e.g. regional, national, or international levels), are important for ensuring food security, as well as for making better policies (e.g. income policies, price policies) in the food sector (Hoang, 2009).

Furthermore, the food traceability systems and food ontologies, e.g. ISO-FOOD ontology (Eftimov et al., 2019), have been introduced in order to assure efficient exchange of information among the food actors from initial production to consumption (supply chain forward) or consumer to production source (supply chain backward) (Sabrina et al., 2018). A good traceability system helps in ensuring the safety and quality of food, and prevents the distribution of foodborne outbreaks, thus avoiding negative effects on consumer health.

### 1.2. Data driven food choices and ordering

Food choices have a huge impact on the economy, health and welfare of the society. Although healthy food options are available to the consumers, making choices reclining towards healthy direction is challenging. Consumers generally choose low cost, convenient and indulgent foods instead of picking healthy diets (Barreiro-Hurlé et al., 2010). Some of the reasons behind the false choices are; inferior taste of the healthy foods compared to their unhealthy versions, psychophysical, lifestyle and environmental factors. The coronavirus pandemic, known as Covid-19, has greatly influenced food consumption patterns. A recent consumer survey carried out in ten EU countries focusing on the impact of Covid-19 pandemic on consumers' food and buying behaviour revealed that they are shopping differently and consuming more food than before, have increased planning to what food products that purchase, included more home cooking and rising demand for healthy, sustainable and local foods (EIT Food, 2020). Furthermore, (Eftimov et al., 2020) applied an artificial intelligence based methodology that showed the highest increase in consumption of pulse ingredients, where this kind of analyses could be utilised to prepare for acute crises situations and take necessary actions in the food supply and distribution chain.

Unhealthy eating and food choices leads to obesity and associated diseases which puts economic burden to healthcare and productivity (Capacci, 2012). In this regard, policy-makers utilise food consumption, price and nutrient data to improve eating habits of the population. The nutritional status of the population is assessed through consumption data of specific food items (e.g. fresh foods, meat, dairy products, cereal foods) and their nutrition content.

However, often times consumer knowledge on healthy and sustainable foods is rather limited and therefore there is a need to assist them during the decision making process by using various pools of food related data starting from agriculture production to food product integrating safety, envi-

ronmental, source, nutrient, allergen, price and availability information. The availability of information in a convenient and easy to understand format will facilitate consumer's selection of healthy foods according to their needs, beliefs and preferences.

As briefly summarized above, the food sector consists of scattered data pools that can play a significant role in building a safe, sustainable and healthy food system. However, it is necessary to generate such data in formats that are easy to handle and compatible with other data sources. The aim of this article is to provide an overview of the data-driven food system and introduce the tools, frameworks and methodologies that can be used in harmonising heterogeneous food data sources.

## 2. Data harmonisation

The generation, compilation and exchange of food data involve several actors. For instance, the major sources of food composition data are the national food composition tables and datasets that are produced and published by governmental bodies. However, such data may also be generated by independent research institutes and other non-governmental agencies (Kapsokafalou et al., 2019), for example, using information extraction or semantic annotation from natural language texts (Popovski et al., 2019a,b,c), (Eftimov et al., 2017a,b). Similarly, food consumption surveys are conducted at many different levels (i.e. national level, household level or individual level) using a variety of methods, for example, paper questionnaires, interviews, 24-hour recall, food record or diary, computers and Internet (Szűcs et al., 2013). These datasets are developed according to the local or regional requirements, following national food description and classification systems (Egan et al., 2007). The use of terms, concepts and semantic relationships in inconsistent manners make the data sharing very challenging. Furthermore, different applications require different types of data, and it is not possible to fulfil the needs of multiple users in a specific way. As an example, data heterogeneity in the food sector may occur due to differences in experimental settings in the fields and food labs (Aubin et al., 2019).

The independent data generation by various parties (e.g. laboratories, research institutes, industries) may vary significantly depending on the experimental protocol as well as the food analysed. This creates inconsistent and incomparable data sources, requiring a large amount of time and other resources to identify profitable opportunities or avoid unknown risks.

Integration of heterogeneous data sources is necessary for finding answers to new questions. For example, researchers require access to multiple datasets that are created from the farm to flour production to investigate the effects of different agricultural treatments on protein and starch quality of cereals. However, before combining such data, it is necessary to understand the similarities and differences among the various data collection methods and technologies, the description of data, and the quality of data. This is to ensure that the data can be made compatible without any significant loss of information. It is for this reason the concept of data harmonisation has been introduced, which is a process of combining data from heterogeneous sources (e.g. national datasets) into integrated, consistent and unambiguous information products (e.g. European datasets). Once the data is considered comparable and

Data element name	Data field
Name	
Report type <ul style="list-style-type: none"> <li>Initial, revised or final report</li> <li>Annual or final performance report</li> </ul>	
Date	<ul style="list-style-type: none"> <li>Calendar drop down</li> <li>Text field allows data format</li> <li>Text field allows free-text</li> </ul>
Address <ul style="list-style-type: none"> <li>Street 1</li> <li>Address line 1</li> <li>Street address line 1</li> </ul>	

**Fig. 2 – An example of a data form and its main components.**

compatible, only then the analyses and interpretations can be considered accurate. The production of data in a consistent manner provides access to a wide range of datasets, enables comparison of food data at broader levels, facilitates benchmarking, and enhances cross-border trade (Egan et al., 2007).

### 2.1. Definition

Personal or business data is traditionally provided in a specific data form. A basic data form has two main components: *data element name* or *field heading* expressed in a free-form text or code describing the nature of the data in a data field; and *data field* which is an area designed for a specified data entry (UNESCAP, 2012). However, many forms and online data collection systems are built independently, without knowing what other systems are like. As shown in Fig. 2, information about address can be collected using “street 1”, “address line 1” or “street address line 1”. The same information is asked but the questions appear differently. Also, different data forms may use the same question to obtain different information, confusing users. For example, one system may be asking whether a report is the “initial”, “revised” or “final” report, while another system may be asking whether the report is an “annual performance” or a “final performance” report. Without any clear instructions, data under the same category, i.e. data element, provided by different users might appear differently. For example, in the case of a data element that allows date in a free-text format, information about a product to be delivered on 05 October 2020 can be expressed as “05-10-2020” by user A and “2020-10-05” by user B. Furthermore, different vocabularies and semantic models are needed to describe the meaning of a data element. For example, we human understand that *surname*, *last name*, and *family name* of a person mean the same, however, computer applications fail to recognize such similarities. This creates problems for everyone: users filling out the forms are confused; programs receive inaccurate information; and data cannot be easily compared or analysed across systems.

“Data harmonisation is an act of reconciling the definition and representation formats of data elements in a domain of interest” (UNESCAP, 2012). It requires consistent use of data



Prospective harmonisation	Retrospective harmonisation
<ul style="list-style-type: none"> <li>• Harmonisation before the data collection</li> <li>• Both stringent and flexible approaches</li> <li>• Agreement on identical data collection tools and procedures</li> <li>• Results in standardised and compatible data</li> <li>• Requires huge amounts of time &amp; other resources</li> </ul>	<ul style="list-style-type: none"> <li>• Harmonisation after the data collection</li> <li>• Flexible approach, mostly</li> <li>• Studies choose data collection procedures</li> <li>• Utilisation of available data sources</li> <li>• Quantity of valid data to be harmonised is limited</li> <li>• Relatively modest time and other costs</li> </ul>

**Fig. 3 – Prospective and retrospective harmonisation.**

elements in terms of their meaning and representation, thus providing the basis for standardisation of data requirements, which formalises the definition and representation format of data elements, reducing the risk of errors, costs and delays. For example, the report type data element in Fig. 2, can be split into “report type” and “performance report type” data elements. However, data harmonisation differs from standardisation as it does not impose a single methodology or norm, instead involves finding ways of integrating or making an agreeable effect from information gathered through disparate techniques (FAO, 2020). In the above example, using “performance report” as a single data element name during the data collection process may reduce confusion.

Harmonisation is an iterative process of capturing, defining, analysing and reconciling information, and it can be achieved in several different ways. The ultimate goal of data harmonisation is to enable the exchange, integration and interoperability of data among the individual information systems for making better decisions. Novel natural language processing (NLP) techniques are an interesting alternative to create semantic content allowing easier data harmonisation approaches (Popovski et al., 2019a).

## 2.2. Prospective and retrospective harmonisation

Data harmonisation is carried out either prospectively or retrospectively following a *stringent* or *flexible* approach (Fig. 3). In case of prospective harmonisation (i.e. harmonisation before data collection), a group of studies agree on the use of identical data collection tools and procedures prior to the actual collection of data. This approach is known as *stringent-prospective harmonisation*, as the participating investigators use specific measures and standard operating procedures, resulting in standardised and compatible data sources. However, the approach does not provide any flexibility for the individual studies to apply specific cultural or scientific context tools and procedures for data collection, making the implementation of similar measures and procedures across a large number of studies very challenging. On the other hand, *flexible-prospective harmonisation* requires agreement on common variables but also provides flexibility for the individual studies to adapt data collection procedures that best suit their requirements (Fortier et al., 2011). The resultant datasets are integrated after analysing the similarities and removing heterogeneities from the data sources. The approach requires a high level of coordination among data collectors. Both *stringent-* and

*flexible-prospective harmonisation* make use of identical procedures for data collection and require huge amounts of time and other resources (Doiron et al., 2012; Granda and Blasczyk, 2016).

The retrospective harmonisation (i.e. harmonisation after data collection) involves the integration of existing data sources produced in separate studies. As very few studies use similar methods for data collection, retrospective harmonisation by its nature has to be *flexible*. This approach requires a deep understanding of the level of compatibility among the data sources. Also, it requires defining a set of target variables to be pooled together, and determining the potential of individual study to produce each target variable. Hence, retrospective harmonisation requires time, access to appropriate expertise, and adequate methodologies. Compared with prospective harmonisation, the quantity of valid data to be harmonized in retrospective harmonisation is limited and depends on the heterogeneity of studies and data collection tools. However, as retrospective harmonisation makes use of the previously collected data, it can be achieved with relatively modest time and costs (Doiron et al., 2012). As there is no unique solution for data harmonisation, both *stringent* and *flexible* approaches should be considered for the integration and interoperability of existing and future food data sources (Fortier et al., 2011).

## 2.3. Data integration – ETL process

Data harmonisation requires a target data model for integrating source data of varying file formats, naming conventions, and/or columns into a cohesive dataset. The data integration (Lenzerini, 2002) involves the Extract, Transform, Load (ETL) process for moving data into a target model (Bansal, 2014). The ETL process comprises of three database operations for transferring data into a common database. In the extraction operation, data from the source files are extracted. The extracted data is then transformed into a format suitable for querying and analysis, during the transformation operation. The data is then loaded into the target database in the final loading operation. The ETL process should be carried out as accurately as possible because an error at one stage of the process will result in inaccurate or missing data in the subsequent stages. Harmonisation adds new dimensions to the target data model that may not be initially planned with the original model, thus providing new insights for better decision making (Munzberg et al., 2018).

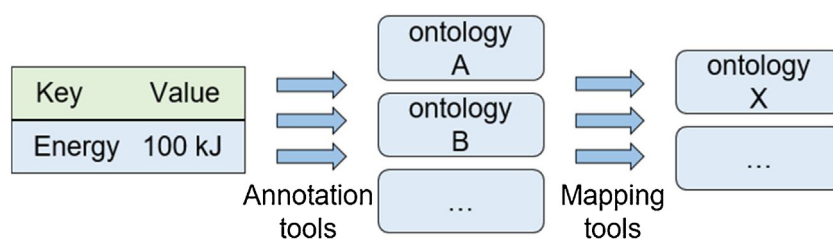


Fig. 4 – Ontology annotation and mapping.

#### 2.4. Ontology and ontology mapping

The variety of classification systems (nomenclature, hierarchies) is usually one of the main reasons behind data heterogeneity. For this purpose, domain-specific ontologies are designed that provide standardised or common vocabularies for the representation of shared knowledge. Basically, an ontology is a set of well-defined hierarchical vocabulary connected with logical relationships. It is used to model a domain of interest by defining classes (common concepts or object types), relations (object properties) and data attributes (data properties) (Gruber, 2009). The concepts and relations within an ontology are identified by globally unique identifiers (e.g. URI, IRI). The ontology itself is defined in a machine-readable syntax (e.g. XML, JSON-LD), using a generic data model for the content. The data model comprises of Resource Description Framework (RDF) triples (subject, predicate, object) that associate a subject to an object via a predicate (W3C, 2014). The triples are stored in an RDF triple store and can be queried using, e.g., the SPARQL (a semantic query language for databases) query language (W3C, 2013).

As shown in Fig. 4, the data (especially the metadata) typically appear in the key-value pairs or strings of text. The ontology annotation tools are used to map the oncoming string of text to the term within an ontology (e.g. ontology A or B) that uniquely identifies the concept conveyed by that particular string of text. In order to get additional information or explore various relationships, the ontology mapping tools are used to identify other relevant ontologies (e.g. ontology X). The RDF triple store creates a data space, and the LinkedData extends this data space to the Internet so that it is possible to create queries that collect data from various cloud services (W3C, 2016).

Ontologies facilitate data interoperability by creating and/or reusing standardised vocabularies for indexing data sources with those vocabularies. Also, ontologies offer automated reasoning and can be used to develop advanced information systems for managing heterogeneous data sources. Furthermore, they are used to design platforms for more collaborative scientific data analysis and decision support (Aubin et al., 2019).

The ontologies developed by various communities have limited applications and are customised according to the user's needs. The distributed and heterogeneous ontologies are combined using ontology mapping (Choi et al., 2006). Ontology mapping (or matching) is the process of finding connections among the concepts of ontologies. The ontology mapping tools analyse the ontologies and indicate the similarities between each concept by using lexical information (indicating the similarity of class names and properties) and structural information (indicating the similarity of relations to other classes or properties) (Euzenat and Shvaiko, 2007). In the food domain, the heterogeneous datasets can be linked

by using various mapping tools, e.g. FoodOntoMap (Popovski et al., 2019c), and annotation tools, e.g. FoodVIZ (Stojanov et al., 2020).

### 3. Data harmonisation in the food sector

In the food sector, one way of performing data harmonisation is to collect food data according to specific guidelines and procedures. The European Food Information Resource Network (EuroFIR), the International Network of Food Data Systems (INFOODS), and the European Food Safety Authority (EFSA) have documented several guidelines and procedures for harmonising food data (see section 3.1). Another approach of data harmonisation is to represent food data according to certain classification and description systems, such as Lingual thesaurus (Lingual™, 2020) and FoodEx2 (European Food Safety Authority, 2011). There are several other tools, frameworks and methodologies that have been proposed for enhancing food data harmonisation (see section 3.2). Furthermore, harmonisation can be facilitated by matching various food concepts to the existing and widely used ontologies. A number of ontologies have been designed for the representation of food-related data, such as FoodOn (FoodOn, 2020) and MESCO (Pizzuti et al., 2017). Several other ontologies, e.g. Food Products Ontology, Open Food Facts, FoodWiki, FOODS–Diabetes Edition, and AGROVOC have been reviewed by (Boulos et al., 2015). Also, the scope of various food ontologies has been discussed in (Popovski et al., 2020b). These ontologies contain concepts and relationships to which different types of data can be linked, resulting in consistent and compatible data sources (see section 3.3). The recently published European data strategy (European Commission, 2020) proposes a concept of data spaces for creation of trustworthy data infrastructures. It aims at enabling data and service sharing and digitalisation of industry domains including the AgriFood sector. This will increase the need for efficient and seamless data harmonisation in the food sector.

#### 3.1. Harmonised food databases

The food composition data is compiled at the national level following the country-specific food description and classification systems. For instance, “energy calculation methods vary between countries; carbohydrate data may or may not include dietary fibre; protein values may have been calculated from nitrogen or other standard factors; vitamins and minerals may be determined using a variety of analytic methods and may be reported with various units and modes of expression” (Black et al., 2011). Differences in the collection and documentation of food composition data generated by individual studies prevent them to be pooled together for thorough analyses (Ioannidou et al., 2020). On the other hand, the globalisation of the food market and international research studies on

food consumption, nutrient intake, nutritional labelling and many other areas require data from more than one national food composition database. Also, each food from each country cannot be simply analysed, requiring the exchange of data among partners. A high level of harmonisation at the food, component and value level is a prerequisite while using food composition data from multiple countries (Westenbrink et al., 2019). Many EU funded initiatives (e.g. EuroFOODS, Cost Action 99, the IARC European Nutrient Data Bank project, INFOODS, EuroFIR) have been carried out for the harmonisation of food composition data to enable the exchange and comparison of data from multiple countries (Kapsokafalou et al., 2019).

INFOODS sponsored by the Food and Agricultural Organization of the United Nations (FAO) is a worldwide network that promotes “international participation, cooperation and harmonisation in the generation, compilation and dissemination of adequate and reliable data on the composition of foods, beverages, and their ingredients in forms appropriate to meet the needs of the various users, including: government agencies, nutrition scientists and educators, health and agriculture professionals, policy makers and planners, food producers, processors, retailers, and consumers” (INFOODS, 2017). INFOODS and FAO facilitate the harmonisation of food composition data by providing guidelines, standards, compilation tools, databases, capacity development tools, policy advice, advocacy tools, technical assistance at country level (FAO, 2017b), (Murphy et al., 2016). Also, INFOODS has played a major role in the development of several food composition databases and has co-published many food composition tables, such as ASEANFOODS, LATINFOODS, Pacific Islands, Lesotho, Brazil, Armenia, West Africa, and Bangladesh. The application of data collection tools and guidelines and procedures provided by the INFOODS and FAO produces harmonised food composition datasets, which can be utilised for multiple applications.

EuroFIR project (2005–2010) was conducted to develop the first European single online platform with up-to-date food composition data across Europe (EuroFIR, 2020b). All steps from data production and data entry, through data aggregation, compilation and control, to data dissemination were thoroughly documented. Detailed description related to foods, components, values, recipes, and references, each with a varying number of properties were documented. Food and component identifier, unit, matrix unit (or mode of expression), value and acquisition type, value reference, method type and method indicator are mandatory properties for value description. In addition, details on sampling, analytical methods and quality indicators can be documented. European compiler organizations agreed on a standardised recipe calculation procedure and food description was harmonised using the LanguaL food description system (LanguaL™, 2020), which is a multi-lingual, multifaceted thesaurus for describing characteristics of a food, including food group, food origin, physical attributes, processing, packaging, dietary uses and geographic origin (Black et al., 2011). The EuroFIR guidelines were agreed upon by all compiler organisations and used in their national food composition databases (Westenbrink et al., 2019).

The harmonised food composition data from 26 European countries is available from the FoodExplorer, which is a web-based tool developed by EuroFIR for searching and comparing food composition data (EuroFIR, 2020a). FoodExplorer also contains data from several Non-European food composition

databases, such as USA, Canada, New Zealand and Japan, however, these databases are developed and maintained by each country according to their own guidelines and compilation processes. Although some of the current datasets in FoodExplorer do not give all the details needed due to incomplete documentation, nevertheless, most of the food composition databases in Europe are now documented in a harmonised manner following the EuroFIR guidelines (Westenbrink et al., 2019).

As updating food composition databases never stops due to the growing and changing food market, value documentation should also never stop, thus new or updated and fully documented versions of the national food composition datasets should be made available to the users as soon as possible (Westenbrink et al., 2019). The work of EuroFIR has been continued by the establishment of EuroFIR AISBL (Association Internationale Sans But Lucratif), which is an international, member-based, non-profit association of food composition data compilers, expert users and stakeholders. It supports the development, management, publication and exploitation of food composition, and promotes international cooperation and harmonisation of standards to improve data quality, storage and access (EuroFIR, 2020b).

Similarly, harmonised food consumption data is necessary for nutrition surveillance, diet and health-related studies, as well as for food exposure assessments. However, such data is collected in several different ways, using different methodologies and food description and classification systems, resulting in incomparable data. Several initiatives have been taken for harmonising food consumption data, for example, the first version of “EFSA Comprehensive European Food Consumption Database” included food and beverage consumption information at the most disaggregated level possible, extracted from the existing national dietary information of the member states according to the FoodEx2 (European Food Safety Authority, 2011) classification system. However, the food consumption data provided by the member states were collected using different national dietary survey methodologies, different national clustering of age groups, and different national food description and classification systems, and thus could not be used for EU-wide or country-to-country comparisons. In 2011, EFSA started a project named as “What’s on the Menu in Europe? - EU Menu” to conduct national dietary surveys using harmonised methods by following the EFSA guidelines, which is estimated to be completed by 2023 (Ioannidou et al., 2020).

In order to harmonise information about food entities, the FoodBase corpus utilising recipes data has been developed by (Popovski et al., 2019b). The annotated corpus consists of 12844 food entity annotations that describe 2105 unique food entities. The resultant dataset can be useful for various machine learning tasks, such as multi-class classification, multi-label classification, and hierarchical multi-label classification. Furthermore, the FoodBase corpus can be helpful in detecting semantic similarities or differences between food concepts, as well as other predictive studies.

Besides, many other EU projects have been directed towards the efficient utilisation of food data. For example, the RICHFIELDS project focused to “design a consumer-data platform to collect and connect, compare and share information about our food behaviours, to revolutionise research on everyday choices made across Europe” (RICHFIELDS, 2015), whereas the Food Nutrition Security (FNS-Cloud) project is dedicated to “develop an infrastructure and services to exploit food, nutri-



tion and security data for a range of purposes” (FNS-Cloud, 2020).

### 3.2. Harmonised data formats, frameworks and tools

The application of specific tools, frameworks, and data exchange formats also generate comparable data sources. Many studies have been conducted to overcome food data integration and interoperability issues. The food exposure assessments require combining food consumption data with concentration data of adverse chemicals in foods. The food consumption data can be formatted at different levels, for instance, food as consumed (e.g. bread), ingredient (e.g. flour), and RAC level (Raw Agricultural Commodities, e.g. wheat, tomato, apple). The RAC components of foods are comparable at the country level. Thus, for the harmonisation of risk assessment procedures at the European level, (Boon et al., 2009) suggested that the food consumption data need to be formatted at the edible-RAC level (RAC without non-edible part, e.g. banana without peel). The authors described an approach of formatting national consumption data at the RAC level and generating the RAC conversion databases, which can further be used for edible-RAC conversion. The RAC conversion databases can be used at the European level in risk assessments of environmental contaminants, heavy metals, pesticides, glycoalkaloids and other chemicals.

Due to multiple non-comparable formats originating from different methods or software tools, data from heterogeneous sources cannot be easily combined with the food consumption data. (Pakkala et al., 2014) proposed an Extensible Markup Language (XML) based data interchange format that can be used as a common interface for linking different software tools. The XML is a general-purpose markup language for encoding information in a format that is both human- and machine-readable, and it is used for sharing data especially on the Internet. As different requirements demand different software tools, the authors concluded that harmonisation is a never-ending issue, and new requirements and methods will emerge by time.

Global Individual Food consumption data Tool (FAO/WHO GIFT) has been developed to harmonise the individual quantitative food consumption (IQFC) data and evaluate the various eating habits of different population groups based on geographical location, especially in the low- and middle-income countries. The platform is a global growing repository of IQFC microdata with all the datasets are harmonised according to the EFSA's food classification and description system called FoodEx2 (modified for global use) (European Food Safety Authority, 2015), hence the end-users can aggregate the available data. FoodEx2 tool provides ready-to-use food-based indicators for the analyses of key data according to population segments and food groups, thus facilitating policymakers in the development of evidence-based policies for better nutrient and food safety in low- and middle-income countries (Leclercq et al., 2019). However, the difficulty with FoodEx2 is the need for manual classification and description of food, thus requiring thorough understanding of the system as well as the food (i.e. composition, processing, marketing, etc.). In (Eftimov et al., 2017b), a semi-automatic system called StandFood has been introduced to classify and describe foods following the FoodEx2 system. It uses machine learning, NLP and post-processing, and having an overall accuracy of 79 percent.

The food traceability systems contain information about the foods throughout the food supply chain, i.e. from farm to fork. Such information can be utilised for complying with the legal and quality standards as well as achieving consumer trust. Based on the XML technology, (Folinas et al., 2006) proposed a very simple framework for food traceability. The suggested framework is easy to use (especially at the base of the supply chain, i.e. farmers, fisherman, cattle breeders, etc.), and exchange information through commonly accessible means, such as cell phones, email and Internet. Also, (Pizzuti and Mirabelli, 2015) proposed a general framework of web-based traceability system (track and trace system) that can be queried for food-related information at every stage of the food supply chain. Similarly, (Holmberg and Åquist, 2018) studied the application of Blockchain technology for food traceability and developed a framework for the milk supply chain. Blockchain is an open, distributed and decentralised verification system for digital transactions where data about the transaction is stored in the so-called blocks which form chains in the network. All transactions, and each of the block in the chain, can be identified as an encrypted piece of information. Anyone connected to the network can add information in the Blockchain once the transaction is verified by everyone in the network, however, no one can change or delete information without authorisation. The authors concluded that even though Blockchain technology promises secure and transparent access to information, the technology is still immature to be applied in the food supply chain, and there is a need for collaboration, information sharing and standardisation in the food traceability.

The diversity of data models utilised by the food data providers hinders the integration of information into a central database. Although there exist obligations for providing ingredient information on the packaging of food products in certain countries, however, there are no legal requirements for digitally providing information on ingredients, nutrient values and allergens to consumers in a standardised format. (Munzberg et al., 2018) investigated the ETL process for integration of data from heterogeneous sources into a central database. As a case study, the authors demonstrated the integration of data coming from five different sources into a central database. From the central database, the information is then transmitted to mobile health applications (used for health monitoring purposes) using the JSON (JavaScript Object Notation) data format, which is an open-standard data interchange file format used for storing and transmitting data.

(Accorsi et al., 2018) designed a decision-support platform that runs on desktop software. The platform can help in decision making by retrieving and managing information related to the food supply chain, such as transportation costs, food demand and country-specific rules.

### 3.3. Ontology based approaches

(Mertins et al., 2012) proposed an ontology-based framework to facilitate effective coordination among the small- and medium-sized enterprises (SMEs) in the food supply chain networks. The service side of the framework provides toolkits for SMEs to develop local customised ontologies, which are then mapped to a common repository of ontologies (generic ontology). At the back-end, domain-experts maintain the generic ontology structures, reference ontologies and the associated knowledge base by using service side toolkits. The consistency among ontologies is analysed and information is provided to



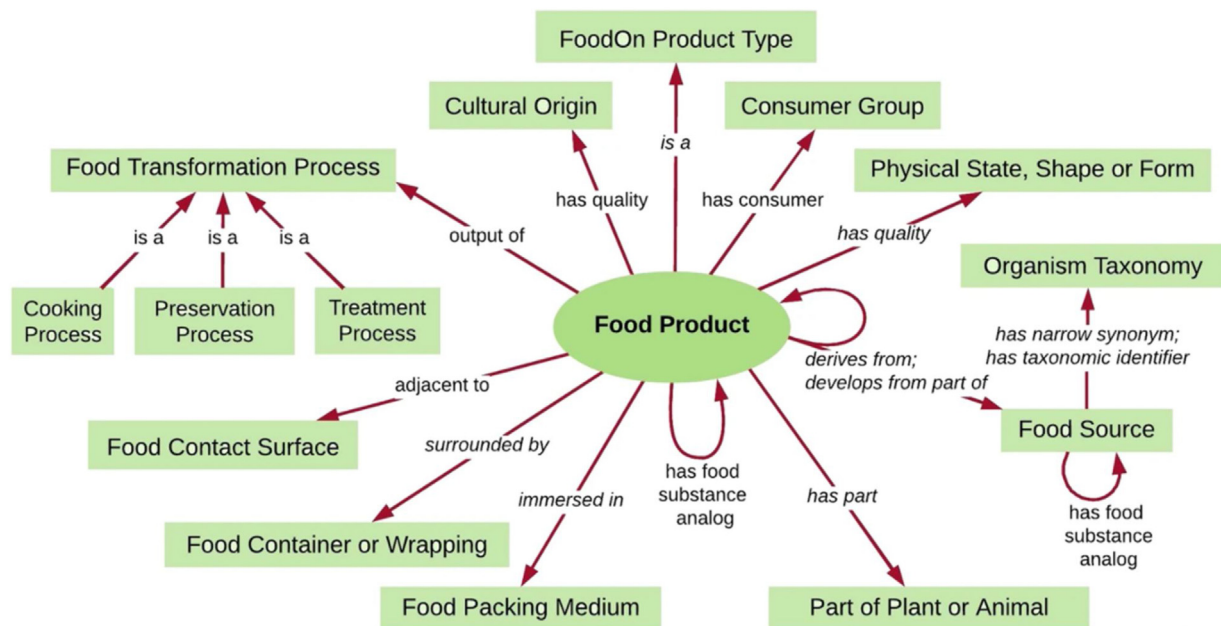


Fig. 5 – FoodOn ontology for the integration and interoperability of food data (Dooley et al., 2018).

SMEs for choosing the right ontology modules. The knowledge base provides knowledge about, e.g., the structure of generic ontology. The implementation of such a framework promotes interoperability by collecting, storing, and sharing information in a consistent manner.

The manufacturing of food products with high nutritional and sensory values and low environmental impact requires multi-disciplinary knowledge. The integration of data and knowledge from different domains in food science (e.g. nutrition, food digestion, sensory and perception, eco-design, microbiology, biochemistry, process engineering) with data and knowledge in environmental analysis is difficult due to the heterogeneity of data sources in terms of formats and vocabularies used by domain experts (Ibanescu et al., 2018; Ibanescu et al., 2016). Process and Observation Ontology (PO2) has been designed to model the food transformation processes and their observations. Based on the core PO2 ontology, a specialized domain ontology called PO2DG (Process and Observation Ontology for Dairy Gels) was built and applied for the production process of French hard cheeses (Aubin et al., 2019). Both PO2 and PO2DG are available from the AgroPortal repository, which is a common platform for agronomy related vocabularies and ontologies (Jonquet et al., 2018).

The accessibility to the required information at every stage of the food supply chain makes it easier to withdraw food products in case of foodborne disease outbreaks. Also, consumer confidence is gained by making food information available in a convenient way. In case of the meat supply chain, (Pizzuti et al., 2017) proposed MESCO (MEat Supply Chain Ontology) ontology that can be used for the traceability of meat as it moves from the primary producer, processor, logistics and distribution channel to the final consumer. The MESCO ontology was validated and checked for consistency using the Pellet reasoner (an open-source reasoner for checking consistencies in the ontology) and it can be searched for information and knowledge related to the meat traceability using the Description Logic (a family of formal knowledge representation languages) queries.

The integration of different data sources also helps in better understanding of the food market and in the decision-making

of when, what type, and how much of food products should be produced. However, due to lack of trust, stakeholders do not make their data available to others. In order to access different data sources, (Verhoosel et al., 2018) developed a platform called HortiCube based on semantic technologies. The contents of data sources on HortiCube are described by developing different ontologies. The HortiCube platform will provide a secure interface to the application developers for checking the availability of data sources, the type of data in the data sources, and the data values in the data sources. For finding answers to different questions, the SPARQL language can be used.

FoodOn is an effort towards data harmonisation in the complex food system involving farms, processing, distribution chains, and consumers (FoodOn, 2020). The aim of FoodOn project is to build an OWL (Web Ontology Language) web ontology that uses well-defined vocabulary and relationships for different food product terminologies. Such a global ontology will promote system interoperability and facilitate food traceability by providing information about food sources, food types, preservation processes, packaging and others, as shown in Fig. 5. At the moment, FoodOn is based on the vocabulary taken mostly from LanguaL, which is a mature and popular food indexing thesaurus (Dooley et al., 2018).

#### 4. Barriers and challenges in food data harmonisation

The food market is continuously changing and growing due to the varying consumption patterns and the launching of new food products. Both food manufacturers and retailers must respond to the food market dynamics to keep their competitive edge. The food manufacturers need to develop new products, alter the composition of processed foods (e.g. reducing sugar, salt, fats) to protect or increase the market share and profits, comply with the governmental policies, and achieve consumers' satisfaction. Similarly, retailers need food data for the marketing of branded foods. Usually, third-party organisations are involved who receive data from the food producers, compile the data into specific datasets and then supply this

data to the retailers. Sharing all the data with third-party organisations may cause data privacy issues. Also, food manufacturers and retailers manipulate the names of food products for marketing purposes, which results in misinterpretation of the data (Popovski et al., 2020a). Furthermore, two commercial food products with the same brand-name can have a different composition (FAO, 2017a). Therefore, following the food market needs and updating and harmonising composition data simultaneously is a challenging task and may require developing new tools and software to assist data compilation, networks to disseminate data and promote training, and data infrastructures to enable linking to other networks, platforms and research organisations (Kapsokefalou et al., 2019).

The food composition data should provide information on national food habits and consumption patterns. Such data should be generated according to international guidelines which result in comparable and reliable data. This data should also be the representative of a large number of highly-consumed foods and their main components. However, the majority of food composition data is based on fresh foods, whereas, information on processed and fortified foods are missing. Also, most of the national food composition databases are incomplete, outdated and do not provide accurate data. Besides, many developing and also some developed countries are still lacking their own national datasets and borrow data from the neighbouring countries. Furthermore, some countries do not consider the importance of food composition data and have no policies for supporting the institutions working voluntarily on food data collection (FAO, 2017a). All these factors are negatively affecting the quality of generated data and making the data harmonisation process difficult.

The food consumption data are usually collected through surveys by the traditional 24-hour recall and dietary record methods. These surveys take longer times and the quality of collected data is affected by the consumers' behaviour. For example, the consumers do not want to participate or mention about the consumption of certain foods or diets. This can be a barrier in the collection and dissemination of data. The process of data collection can be improved by using advanced technologies in national dietary surveys. Also, the development of web-based online platforms, mobile or tablet applications will encourage consumers to provide accurate and detailed information about their consumption (Ioannidou et al., 2020).

## 5. Conclusions

This work emphasised on the importance of data in the food sector, introduced the necessity for food data harmonisation, highlighted the barriers and challenges in harmonising food data, and provided examples of studies focusing on food data integration and interoperability issues. Data harmonisation is the process of combining data from heterogeneous data sources into integrated, consistent and unambiguous information products. It can be achieved either prospectively or retrospectively, depending on the objective and type of investigation. The resultant datasets and data spaces help to identify profitable opportunities or avoid unknown risks. The development of various food databases has resulted in the harmonisation of certain food data types, including food composition data and food consumption data. For other types of data, such as food processing and traceability data, researchers have proposed several tools, frameworks, and methodologies. Ontologies can play a significant role in har-

monising food data by providing standardised vocabularies. As different studies require different tools and procedures, data harmonisation is an iterative and continuous process.

## Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

- Accorsi, R., Cholette, S., Manzini, R., Tufano, A., 2018. A hierarchical data architecture for sustainable food supply chain management and planning. *J. Clean. Prod.* 203, 1039–1054, <http://dx.doi.org/10.1016/j.jclepro.2018.08.275>.
- Aubin, S., Bisquert, P., Buche, P., Dibie, J., Ibanescu, L., Jonquet, C., Roussey, C., 2019. Recent progresses in data and knowledge integration for decision support in agri-food chains. *Les Actes Des 30èmes Journées Francophones d'ingénierie Des Connaissances (IC PFIA 2019)*, 43–59, <https://hal.inrae.fr/hal-02609466>.
- Bansal, S.K., 2014. Towards a semantic extract-transform-load (ETL) framework for big data integration. 2014 IEEE International Congress on Big Data, 522–529, <http://dx.doi.org/10.1109/BigData.Congress.2014.82>.
- Barreiro-Hurlé, J., Gracia, A., de-Magistris, T., 2010. Does nutrition information on food products lead to healthier food choices? *Food Policy* 35 (3), 221–229, <http://dx.doi.org/10.1016/j.foodpol.2009.12.006>.
- Black, L.J., Ireland, J., Møller, A., Roe, M., Walton, J., Flynn, A., Finglas, P.M., Kiely, M., 2011. Development of an on-line Irish food composition database for nutrients. *J. Food Compos. Anal.* 24 (7), 1017–1023, <http://dx.doi.org/10.1016/j.jfca.2011.01.015>.
- 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 Chem. Toxicol.* 47 (12), 2883–2889, <http://dx.doi.org/10.1016/j.fct.2009.08.003>.
- Boulos, M., Yassine, A., Shirmohammadi, S., Namahoot, C., Brückner, M., 2015. Towards an “internet of food”: food ontologies for the internet of things. *Fut. Internet* 7 (4), 372–392, <http://dx.doi.org/10.3390/fi7040372>.
- Capacci, S., 2012. Harmonizing Data From the UK Expenditure and Food Survey and the National Food Survey: An Application to Fruit and Vegetable Demand. <http://amsacta.unibo.it/3465/>.
- Choi, N., Song, I.-Y., Han, H., 2006. A Survey on Ontology Mapping. *ACM Digital Library*, <http://dx.doi.org/10.1145/1168092.1168097>.
- Doiron, D., Raina, P., Raina, P., L'Heureux, F., Fortier, I., 2012. Facilitating collaborative research: implementing a platform supporting data harmonization and pooling. *Norsk Epidemiologi* 21 (2), 221–224, <http://dx.doi.org/10.5324/nje.v21i2.1497>.
- 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 Sci. Food* 2 (23), <http://dx.doi.org/10.1038/s41538-018-0032-6>.
- 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 Chem.* 277, 382–390, <http://dx.doi.org/10.1016/j.foodchem.2018.10.118>.
- Eftimov, T., Korošec, P., Koroušić Seljak, B., 2017a. StandFood: standardization of foods using a semi-automatic system for classifying and describing foods according to FoodEx2. *Nutrients* 9 (6), 542, <http://dx.doi.org/10.3390/nu9060542>.

- Eftimov, T., Koroušić Seljak, B., Korošec, P., 2017b. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PLoS One* 12 (6), e0179488, <http://dx.doi.org/10.1371/journal.pone.0179488>.
- Eftimov, T., Popovski, G., Petković, M., Koroušić Seljak, B., Kocev, D., 2020. COVID-19 pandemic changes the food consumption patterns. *Trends in Food Science and Technology*, vol. 104. Elsevier Ltd., pp. 268–272, <http://dx.doi.org/10.1016/j.tifs.2020.08.017>.
- Egan, M.B., Fragodt, A., Raats, M.M., Hodgkins, C., Lumbers, M., 2007. The importance of harmonizing food composition data across Europe. *Eur. J. Clin. Nutr.* 61, 813–821, <http://dx.doi.org/10.1038/sj.ejcn.1602823>.
- EIT Food, 2020. COVID-19 Impact on Consumer Food Behaviours in Europe. <https://www.eitfood.eu/media/news-pdf/COVID-19-Study.-European-Food-Behaviours.-Report.pdf>.
- EuroFIR, 2020a. Discover Food Data Through Our Webtools – EuroFIR Webtools. <http://www.eurofir.org/our-tools/>.
- EuroFIR, 2020b. European Food Information Resource Network – EuroFIR. <https://www.eurofir.org/about/eurofir/>.
- European Commission, 2020. Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. A European Strategy for Data. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0066>.
- European Food Safety Authority, 2011. The Food Classification and Description System FoodEx 2 (draft-revision 1). EFSA Supporting Publications 2011:215. doi:10.2903/sp.efsa.2011.EN-215.
- European Food Safety Authority, 2015. The Food Classification and Description System FoodEx 2 (revision 2) – European Food Safety Authority (EFSA). EFSA Supporting Publications 2015:EN-804. doi:10.2903/sp.efsa.2015.en-804.
- Euzenat, J., Shvaiko, P., 2007. *Ontology Matching*. Springer Berlin Heidelberg, <http://dx.doi.org/10.1007/978-3-540-49612-0>.
- FAO, 2017a. INFOODS: Food Composition Challenges. <http://www.fao.org/infoods/infoods/food-composition-challenges/en/>.
- FAO, 2017b. International Network of Food Data Systems – INFOODS: About INFOODS. <http://www.fao.org/infoods/infoods/en/>.
- FAO, 2020. 9. Data Harmonisation. <http://www.fao.org/3/X0587E/x0587e09.htm>.
- Finglas, P.M., Berry, R., Astley, S., 2014. Assessing and improving the quality of food composition databases for nutrition and health applications in Europe: the contribution of EuroFIR. *Adv. Nutr.* 5 (5), 608S–614S, <http://dx.doi.org/10.3945/an.113.005470>.
- FNS-Cloud, 2020. Overview – FNS-Cloud. <https://www.fns-cloud.eu/overview/>.
- Folinas, D., Manikas, I., Manos, B., 2006. Traceability data management for food chains. *Br. Food J.* 108 (8), 622–633, <http://dx.doi.org/10.1108/00070700610682319>, Emerald Group Publishing Limited.
- FoodOn, 2020. FoodOn: A Field to Fork Ontology. <https://foodon.org/>.
- Fortier, I., Doiron, D., Burton, P., Raina, P., 2011. Invited commentary: consolidating data harmonization – how to obtain quality and applicability? *Am. J. Epidemiol.* 174 (3), 261–264, <http://dx.doi.org/10.1093/aje/kwr194>.
- Granda, P., Blasczyk, E., 2016. Data Harmonisation. Guidelines for Best Practice in Cross-Cultural Surveys, <http://www.ccsgr.isr.umich.edu/>.
- Gruber, T., 2009. Ontology – Definition in the Encyclopedia of Database Systems, <https://tomgruber.org/writing/ontology-definition-2007.htm>.
- Hoang, L.V., 2009. Estimation of Food Demand from Household Survey Data in Vietnam, <http://www.depocenwp.org/modules/publications/watch.php?id=64>.
- Holmberg, A., Åquist, R., 2018. *Blockchain Technology in Food Supply Chains: A Case Study of the Possibilities and Challenges with an Implementation of a Blockchain Technology Supported Framework for Traceability*, <https://www.semanticscholar.org/paper/Blockchain-technology-in-food-supply-chains-%3A-A-of-Holmberg-Åquist/9fec814ef3c09683ab5949e8bdf2d79c811dc871?p2df>.
- Ibanescu, L., Allard, T., Dervaux, S., Dibie, J., Guichard, E., Pénicaud, C., Raad, J., 2018. A use case of data integration in food production. In: 21st International Conference on Knowledge Engineering and Knowledge Management, EKAW, Nancy, France <https://hal.archives-ouvertes.fr/hal-01924171/>.
- Ibanescu, L., Dibie, J., Dervaux, S., Guichard, E., Raad, J., 2016. PO2 – a process and observation ontology in food science. Application to dairy gels. *Commun. Comput. Inf. Sci.* 672, 155–165, [http://dx.doi.org/10.1007/978-3-319-49157-8\\_13](http://dx.doi.org/10.1007/978-3-319-49157-8_13).
- INFOODS, 2017. International Network of Food Data Systems (INFOODS). <http://www.fao.org/infoods/infoods/en/>.
- Ioannidou, S., Horváth, Z., Arcella, D., 2020. Harmonised Collection of National Food Consumption data in Europe. Food Policy, <http://dx.doi.org/10.1016/j.foodpol.2020.101908>.
- Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M.A., Pesce, V., Larmande, P., 2018. AgroPortal: a vocabulary and ontology repository for agronomy. *Comput. Electron. Agric.* 144, 126–143, <http://dx.doi.org/10.1016/j.compag.2017.10.012>.
- Kapsokafalou, M., Roe, M., Turrini, A., Costa, H.S., Martínez-Victoria, E., Marletta, L., Berry, R., Finglas, P., 2019. Food composition at present: new challenges. *Nutrients* 11 (8), 1714, <http://dx.doi.org/10.3390/nu11081714>.
- LanguaL™, 2020. LanguaL™ – The International Framework for Food Description. <https://www.langual.org/default.asp>.
- Leclercq, C., Allemand, P., Balcerzak, A., Branca, F., Sousa, R.F., Lartey, A., Lipp, M., Quadros, V.P., Verger, P., 2019. FAO/WHO GIFT (Global Individual Food consumption data Tool): a global repository for harmonised individual quantitative food consumption studies. *Proc. Nutr. Soc.* 78 (4), 484–495, <http://dx.doi.org/10.1017/S0029665119000491>.
- Lenzerini, M., 2002. Data integration: a theoretical perspective. Proceedings of the Twenty-First ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems – PODS '02, <http://dx.doi.org/10.1145/543613.543644>.
- Mertins, K., Jaekel, F.-W., Deng, Q., 2012. Towards information customization and interoperability in food chains. *Lecture Notes Bus. Inf. Process.* 122, 92–103, [http://dx.doi.org/10.1007/978-3-642-33068-1\\_10](http://dx.doi.org/10.1007/978-3-642-33068-1_10).
- Munzberg, A., Sauer, J., Hein, A., Rosch, N., 2018. The use of ETL and data profiling to integrate data and improve quality in food databases. 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Limassol 2018, 231–238, <http://dx.doi.org/10.1109/WiMOB.2018.8589081>.
- Murphy, S.P., Charrondiere, U.R., Burlingame, B., 2016. Thirty years of progress in harmonizing and compiling food data as a result of the establishment of INFOODS. *Food Chem.* 193, 2–5, <http://dx.doi.org/10.1016/j.foodchem.2014.11.097>.
- Pakkala, H., Christensen, T., Presser, K., Victoria, I.M.de, 2014. Towards harmonized data interchange in food consumption data. *Comput. Standards Interfaces* 36 (3), 592–597, <http://dx.doi.org/10.1016/j.csi.2013.10.004>.
- Pizzuti, T., Mirabelli, G., 2015. The Global Track&Trace System for food: general framework and functioning principles. *J. Food Eng.* 159, 16–35, <http://dx.doi.org/10.1016/j.jfoodeng.2015.03.001>.
- Pizzuti, T., Mirabelli, G., Grasso, G., Paldino, G., 2017. MESCO (MEat Supply Chain Ontology): an ontology for supporting traceability in the meat supply chain. *Food Control* 72, 123–133, <http://dx.doi.org/10.1016/j.foodcont.2016.07.038>.
- Popovski, G., Ispirova, G., Hadzi-Kotarova, N., Valenčič, E., Eftimov, T., Seljak, B., 2020a. Food data integration by using heuristics based on lexical and semantic similarities. 13th International Joint Conference on Biomedical Engineering Systems and Technologies – Volume 5: HEALTHINF, 208–216, <http://dx.doi.org/10.5220/0008990602080216>.
- Popovski, G., Kocev, S., Seljak, B., Eftimov, T., 2019a. 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, <http://dx.doi.org/10.5220/0007686309150922>.
- Popovski, G., Korousic Seljak, B., Eftimov, T., 2020b. A survey of named-entity recognition methods for food information extraction. In: IEEE Access, Vol. 8, Institute of Electrical and Electronics Engineers Inc, pp. 31586–31594, <http://dx.doi.org/10.1109/ACCESS.2020.2973502>.
- Popovski, G., Koroušić Seljak, B., Eftimov, T., 2019b. FoodBase corpus: a new resource of annotated food entities. Database 2019 (1), <http://dx.doi.org/10.1093/database/baz121>.
- Popovski, G., Seljak, B., Eftimov, T., 2019c. FoodOntoMap: Linking Food Concepts across Different Food Ontologies. Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, vol. 2, 195–202, <http://dx.doi.org/10.5220/0008353201950202>.
- RICHFIELDS, 2015. About – RICHFIELDS. <https://www.richfields.eu/about/>.
- Sabrina, C.I., Andrés, B., MME, A., 2018. Traceability in the Food Supply Chain: review of the literature from a technological perspective. Dirección y Organización 64, 50–55 <https://riunet.upv.es/handle/10251/121089>.
- Stojanov, R., Popovski, G., Jofce, N., Trajanov, D., Koroušić Seljak, B., Eftimov, T., 2020. FoodViz: Visualization of Food Entities Linked Across Different Standards. Springer, Cham, pp. 28–38, [http://dx.doi.org/10.1007/978-3-030-64580-9\\_4](http://dx.doi.org/10.1007/978-3-030-64580-9_4).
- Szűcs, V., Szabó, E., Bánáti, D., 2013. Short overview of food consumption databases. Czech J. Food Sci. 31, 541–546, <http://dx.doi.org/10.17221/160/2013-CJFS>.
- UNESCAP, 2012. Data Harmonization and Modelling Guide for Single Window Environment. United Nations. <https://www.unescap.org/resources/data-harmonization-and-modelling-guide-single-windows-environment>.
- Verhoosel, J., Bekkum, M. van, Verwaart, T., 2018. Semantic interoperability for data analysis in the food supply chain. Int. J. Food Syst. Dyn. 9 (1), 101–111, <http://dx.doi.org/10.18461/ijfsd.v9i1.917>.
- W3C, 2013. SPARQL Query Language for RDF. <https://www.w3.org/TR/rdf-sparql-query/>.
- W3C, 2014. Resource Description Framework (RDF): Concepts and Abstract Syntax. <https://www.w3.org/TR/rdf-concepts/>.
- W3C, 2016. LinkedData. <https://www.w3.org/wiki/LinkedData>.
- Westenbrink, S., Kadvan, A., Roe, M., Seljak, B.K., Mantur-Vierendeel, A., Finglas, P., 2019. 12th IFDC 2017 Special Issue – evaluation of harmonized EuroFIR documentation for macronutrient values in 26 European food composition databases. J. Food Compos. Anal. 80, 40–50, <http://dx.doi.org/10.1016/j.jfca.2019.03.006>.
- Williamson, C., 2005. Synthesis Report No 2: The Different Uses of Food Composition Databases. [https://www.bezpecnostpotravin.cz/UserFiles/File/Kvasnickova2/EuroFIR\\_2.pdf](https://www.bezpecnostpotravin.cz/UserFiles/File/Kvasnickova2/EuroFIR_2.pdf).