Food Data and Food Vocabularies

PNH 2020 deliverable D5.3

dr. J.L. (Jan) Top, drs. M.J. (Marielle) Timmer, dr. G. (Görkem) Simsek-Senel

WAGENINGEN
UNIVERSITY & RESEARCH

Food Data and Food Vocabularies

PNH 2020 deliverable D5.3

Authors: dr. J.L. (Jan) Top, drs. M.J. (Marielle) Timmer, dr. G. (Görkem) Simsek-Senel

This research project has been carried out by Wageningen Food & Biobased Research commissioned and funded by TKI Agri & Food (TKI-AF-15262). The work was conducted in an objective way by researchers who act impartially with respect to the client(s) and sponsor(s).

Wageningen Food & Biobased Research
Wageningen, December 2020

Public

Report 2115
DOI https://doi.org/10.18174/538580

WFBR Project number: 6234108700
Version: Final
Reviewer: dr. H. (Hajo) Rijgersberg
Approved by: dr.ir. H. (Henk) Wensink
Client and Sponsor: TKI Agri & Food
Confidentiality of the report: Public

The research that is documented in this report was conducted in an objective way by researchers who act impartial with respect to the client(s) and sponsor(s). This report can be downloaded for free at https://doi.org/10.18174/538580/ or at www.wur.eu/wfbr (under publications).

PO box 17, 6700 AA Wageningen, The Netherlands, T + 31 (0)317 48 00 84, E info.wfbr@wur.nl, www.wur.eu/wfbr.

Contents

Summary

The project Personalised Nutrition and Health (PNH) aims to develop methods and knowledge needed to make personalised food and health advice. Smart applications can support consumers in making healthy, but also safe and sustainable choices. Such applications require the availability of high quality data about food products and ingredients. This includes attributes such as nutritional values, ingredients, way of application, associated CO_2 impact, etc. This information needs to be accessible, timely, accurate, reliable, comprehensive and reusable.
Currently, many data sources about food are available. Although very rich in content, it is very difficult to extract and combine the data from these different sources, and even then many values remain missing or unreliable. An approach based on the principles of Linked Data can make the data available for automated processing, in particular if additional attributes are specified. Another issue related to the availability of product data is that it should ideally be openly available, directly from the producer of a product.

In this document we focus on food data for health, in particular in the context of personalised dietary advice. We present an overview of existing data sources.

It appears that for the Netherlands, nutritional values at the level of generic products can still best be found the NEVO table, but they are also increasingly available at the level of commercial product from the GS1 data pool. Many other attributes related to consumer preferences, application context, sustainable responsible production are scattered, if available at all.

Several other data sources are available for retrieving food data relevant for applications in personalised nutrition. These data sources are still difficult to combine. Therefore the development of semantic standards is crucial. The most promising approach would be to combine the efforts by GS1, representing food supply chains worldwide and FoodOn as a research effort to introduce Linked Data standards in the world of food.

1 Introduction

The objective of the project Personalised Nutrition and Health (PNH) is to develop and demonstrate new technologies for personalised food and health advice. Smart applications can support consumers in making healthy, but also safe and sustainable choices. Such applications require the availability of high quality data about food products and ingredients.
In the case of advice on healthy food products, today’s consumer typically relies on data provided in the food product label, including

- Product name
- Product category
- Ingredients
- Nutritional values
- Allergens
- Certificates

In 2014 the EU regulation No 1169/2011 on the provision of food information to consumers[1] entered into application, followed by the obligation to provide nutrition information in 2016. It is quite an effort for companies to provide this information and keep it updated[2]. Product Information Management[3] is considered as the next step in enterprise content management and product data are considered to add distinct value to products.

However, for providing targeted and actionable advice, we also need to know other product attributes. These can make dietary swaps acceptable for consumers. Some examples of such attributes are:

- Meal moment
- Taste
- Texture
- Processing suggestion, recipe
- Consistency
- Serving size
- Health claim
- CO_2 footprint

Such properties are not yet readily available for most food products, not even at the level of food types or categories.

Another issue related to the availability of product data is that it should ideally be openly available, directly from the producer of a product. So, rather than being passed on through the supply chain, each producer should have its own ‘data access point’ that provides all information for produced items. Moreover, this information should be provided as *linked* data, such that the data can be retrieved and processed automatically. Linked data[4] uses data standards that define unique identifiers (URIs, i.e. locations on the web) for each item, attributes of these items, and relations between them. A so-called Knowledge Graph defines the underlying conceptual model. Semantic technologies can process linked data and make software applications ‘smart’.

In this report we want to find out whether this type of data is available and which knowledge graphs (ontologies, controlled vocabularies) are available to support a linked data approach. We also discuss what is needed to proceed in this direction. Note that we frequently rely on text provided by data suppliers on their website. We include references to these websites, but do not always show explicitly where literal quotes are used.

[1] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A32011R1169
[2] https://www.gs1.nl/nieuws/nieuws/2019/albert-heijn-alleen-artikelen-betrouwbare-data-komen-ons-assortiment
[3] https://en.wikipedia.org/wiki/Product_information_management
[4] https://en.wikipedia.org/wiki/Linked_data

2 Data sources

This chapter lists several digital sources for food product data. Important aspects to be addressed are

- Which (types of) products are covered?
- Which attributes are covered?
- What can be said about the quality of the data?
- Is the data accessible?
- Who is responsible for providing and maintaining the data?

It appears that not all questions can be answered for all sources, but at least we can give an impression of what is known.

2.1 Food product suppliers

The most obvious place to look for product data would be repositories provided by the producers of these products, such as farmers, food and beverage industry (premium labels), but also retailers (private labels). For example, Unilever provides data on the ingredients of personal care products and their function (e.g., 'Andrélon shampoo for men iedere dag' contains 'sodium benzoate' as a 'preservative')[5]. The Food Information Law (European Regulation 1169/2011) sets requirements for what must be included on the label of foods. There is no law yet for the data being digitally available, unless they are provided through a web shop. Food product data can indeed be found on web outlets, however, as far as we know, none of these companies have openly accessible portals (APIs) for food product data. In the Personalised Nutrition and Health project we encountered practical difficulties in retrieving and using data from retail organisations. For example, it was not possible to get product data from one of the retailers as bulk data, even though it was indirectly available on their website. For the Dutch market, most data is channelled through GS1 and PS in foodservice, which we will discuss in later sections.

2.2 GS1 Data Source

The mission of GS1 is to issue unique barcodes for products (GTIN, previously known as EAN).
In addition, *GS1 Data Source* provides a central data pool for exchanging product information[6].
The focus is on data for trading and logistics. The producers who provide product data can decide who can have access to the data. There are annual costs involved in accessing this data. About 1400 companies (food health and beauty) provide data, whereas about 100 companies and organisations use data. Data providers include AHold, Superunie, Jumbo, but not Lidl and Aldi.[7] Organisations using data are for example hospitals, service providers, retailers, traders.

The *Global Data Synchronisation Network GDSN*[8] is the world's largest product data network. With GDSN, high quality product content is uploaded, maintained and shared automatically, ensuring trading partners have immediate access to the most current and complete information needed to exchange products on both local and global markets.

[5] https://pioti.unilever.com/pioti/en/p1.asp
[6] https://www.gs1.nl/gs1-datapools-delen-artikelgegevens
[7] https://www.gs1.nl/wie-maken-gebruik-gs1-data-source
[8] https://www.gs1.org/services/gdsn

For consumer decision support a more relevant service is the GS1 Datalink Service. This service is typically used by organisations such as Voedingscentrum, Questionmark and RIVM. It only includes data for premium (not private label) products.
Product attributes available from the GS1 Datalink service relevant for PNH are: name, category, ingredients, nutritional values, allergens, but also 'health claims', general labels and certificates (such as *Bewuste keuze, halal, suitability for a 880-1220 kcal diet, low on sugar, vegetarian, Demeter, Fair Trade*, etc), serving suggestion, preparation instructions, serving size, etc.[9]

GS1 runs a program on data quality called DatakwaliTijd 2.0. In this program they check consistency between printed labels on physical product and data in the data source[10]. However, this does not imply that the data is completely correct in all cases.

2.3 PS in foodservice

PS in foodservice[11] supports producers, farmers and growers to make their food data available for their clients. The food data concerns the legally required specifications (ingredients, allergens and nutritional values) and logistics data size, weight, packaging), but also commercial information such as the story behind a product, the quality remarks, logos, product photos, certificates, recipes and videos of preparation and presentation methods. Data providers get their individual website for data entry and can have their own digital catalogue, giving an overview of their assortment. Currently 926 producers are using this platform.

With producers, farmers and growers making their product information available PS in foodservice supports wholesalers to get the commercial, product-related and logistics information from all their suppliers. In this way the wholesalers can comply with their legal obligations to provide information to their customers. Presently 98 wholesalers, which is 99% of the wholesalers, are using the application. Also hospitality professionals such as healthcare, caterers, hospitality and fresh-retail can have a digital insight into their product range, so they can provide up-to-date information to their guests.

PS Food in foodservice offers FoodBook[12], a free online overview of 250.000 public products in the PS in foodservice database. Foodbook shows photos, videos, allergens, nutritional values, quality marks, the story behind and much more of products.

2.4 NEVO and NELADA

The Dutch Food Composition Database (NEVO)[13] contains data on the composition of foods consumed frequently by a large part of the Dutch population. These foods contribute significantly to the intake of energy and nutrients. Foods of importance for specific groups of the Dutch population are also included. NEVO is owned by the Dutch Ministry of Health, Welfare and Sports, and maintained at the Netherlands Institute for Public Health and the Environment (RIVM). RIVM collaborates with the Netherlands Nutrition Centre in collecting nutritional data. Data in NEVO originate from chemical food analyses, food manufacturers, international food composition tables or (recipe) calculations. Data

[9] https://www.gs1.org/sites/default/files/docs/gs1-source/GS1_TSD_Product_Data_Modules_Standard_v1%202%201_r_2016-09-20.pdf

[10] https://www.gs1.nl/sectorafspraken-over-standaarden/levensmiddelen-en-drogisterij/gs1-data-source-levensmiddelen-en-8

[11] http://info.psinfoodservice.nl/

[12] http://foodbook.psinfoodservice.nl/prod

[13] https://www.rivm.nl/en/dutch-food-composition-database/about-nevo-data

published in NEVO online are freely accessible and can be used to engage in scientific research (nutrition research in particular), the food industry, dietetics and nutrition counselling and/or public health education.

Data for NEVO online[14] are derived from NEVO. NEVO online 2019 contains data of 2152 food items. In NEVO online the composition of foods can be found without having direct access to the complete dataset. It is possible to search on NEVO food code, name of the food item (both Dutch and English) or food group. Synonyms are added to the food names (in Dutch only), to improve search results. Results can be retained to compare them with new search results. Results are shown per component group (e.g., macronutrients, minerals or fat-soluble vitamins). Information on how to use NEVO online is available on the NEVO website[15].

NEVO is part of the *Levensmiddelendatabase (NELEDA)* or Food Database[16]. This database is an independent database with extensive information on more than 90,000 food products sold in the Netherlands.
The Netherlands Nutrition Centre (Voedingscentrum) and RIVM manage the data of various products in the Food Database:

- General, unbranded foods, such as apples and potatoes.
- (Private) brand items: foods from a certain manufacturer or supermarket.
- Nutritional supplements such as vitamins and minerals

The Food Database is only accessible with permission from the Netherlands Nutrition Centre.

2.5 USDA FoodData Central

USDA offers *FoodData Central*[17] as a service for researchers, policy makers, nutrition and health professionals, product developers and others. FoodData Central contains five distinct databases that provide information on food and nutrient profiles: Foundation Foods, Food and Nutrient Database for Dietary Studies 2015-2016 (FNDDS 2015-2016), National Nutrient Database for Standard Reference Legacy Release (SR Legacy), USDA Global Branded Food Products Database (Branded Foods), and Experimental Foods. It provides a broad snapshot of the nutrients and other components found in a wide variety of foods and food products.

Two of the data types—Foundation Foods and Experimental Foods—will be the primary focus of efforts as FoodData Central is expanded and developed in coming years. Foundation Foods includes values for nutrients and other food components on a diverse range of foods and ingredients as well as extensive metadata. These metadata include the number of samples, sampling location, date of collection, analytical approaches used, and if appropriate, agricultural information such as genotype and production practices. Experimental Foods contains foods produced, acquired or studied under specific conditions, such as alternative management systems, experimental genotypes, or research/analytical protocols. The foods in this data type may not be commercially available to the general public or the data may expand information about the specific food. Experimental Foods are for research purposes and may not be appropriate as a reference for the consumer or for diet planning.

The FoodData Central API provides REST access to FoodData Central (FDC). It is intended primarily to assist application developers wishing to incorporate nutrient data into their applications or website. Documentation is available via links on Data Type Documentation. This documentation provides the

[14] https://www.rivm.nl/documenten/specifications-of-references-in-nevo-online-2019

[15] http://www.rivm.nl/en/Topics/Topics/D/Dutch_Food_Composition_Database/Access_NEVO_data/NEVO_online/Search_and_compare

[16] https://www.rivm.nl/nederlands-voedingsstoffenbestand/organisatie/levensmiddelendatabank

[17] https://fdc.nal.usda.gov/

detailed definitions and descriptions needed to understand the data elements referenced in the API documentation.

2.6 VCP

The Food Consumption Survey (VCP) provides insight into the Dutch consumption of foodstuffs, the intake of macronutrients and micronutrients, and the intake of potentially harmful chemical substances. The data is used for information purposes and scientific research. The most recent food consumption survey was conducted by RIVM (National Institute for Public Health and the Environment) from 2012 to 2016.

Data on the national diet can be relevant for personalised nutrition in the sense that they can provide default values for a population if individual data on food intake is missing. Secondly that can assist in suggesting which product and ingredients would be acceptable in Dutch culture in a more healthy or sustainable diet.

2.7 EFSA

The EFSA Comprehensive European Food Consumption Database[18] contains data on food consumption habits and patterns across the EU. It provides detailed information for a number of European countries in refined food categories and specific population groups. Statistics from the database enable quick screening for chronic and acute exposure to substances that may be found in the food chain.
For certain specific food categories, particularly those related to regulated products, EFSA makes summary statistics of food consumption data.
For each country, food consumption data based on dietary surveys are presented. They are shown per age class (Infants, Toddlers, Other children, Adolescents, Adults, Elderly and Very elderly), for the total population and for consumers only. Food consumption statistics are reported both in grams/day and in grams/kg body weight per day. Dutch data are based on de VCP surveys by RIVM.

2.8 SELF NutritionData

SELF Nutrition Data[19] aims to provide accurate and comprehensive nutrition data, and to make it accessible and understandable to all. The information in Nutrition Data's database comes from the USDA's National Nutrient Database for Standard Reference and is supplemented by listings provided by restaurants and food manufacturers. In addition to food composition data, Nutrition Data also provides a variety of proprietary tools to analyse and interpret that data. These interpretations represent Nutrition Data's opinions and therefor subjective. They are based on calculations derived from Daily Reference Values (DRVs), Reference Daily Intakes (RDIs), published research, and recommendations of the FDA.

2.9 Portie-online

Portie-online[20] contains information about weights and contents of food available in the Netherlands and is published by the National Institute for Public Health and the Environment (RIVM National Institute for Public Health and the Environment). The information comes from a database that is compiled and maintained by RIVM in collaboration with the Nutrition Centre (Voedingscentrum) and Wageningen University (WUR Wageningen University & Research). Portie-online contains data about

[18] https://www.efsa.europa.eu/en/food-consumption/comprehensive-database
[19] https://nutritiondata.self.com/
[20] https://portie-online.rivm.nl/

generic foods in a number of product groups. One can search by food group, (part of) the product name, synonyms or by NEVO code.
In addition to the data that can be looked up in Portie-online for each food, a number of standard household sizes can be distinguished. This is, for example, the content of a cup, glass, bowl or spoon.

2.10 Open Food Facts

Open Food Facts[21] is a free, open and collaborative database of food products from the entire world, developed by thousands of volunteers from all around the world. It is in particular filled with products that are on the market in France, secondly US. The Netherlands is currently marginally present (5200 products).
The database contains about 940.000 products which are findable by barcode. The product information provides the Nutriscore of the product and a comparison with other products. The website uses a Wikipedia–approach with multiple authors and mechanisms for corrections[22].
A mobile app is available[23] that supports barcode scanning, provides Nutri-Score values and lists the NOVA group of ultra-processed products. If a product does not exist yet in the database, it can be added in less than a minute. The app allows you to create personal lists and add products. This can be a shopping list, or lists to manage your stocked products (freezer, cellar, pantry etc.) and avoid waste.

2.11 CheBI

Chemical Entities of Biological Interest (ChEBI)[24] is a freely available dictionary of molecular entities, focusing on 'small' chemical compounds. Although these are not food products as such, information about these ingredients is often needed to determine properties of consumer products.
The term 'molecular entity' refers to any constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity. The molecular entities in question are either products of nature or synthetic products used to intervene in the processes of living organisms. ChEBI incorporates an ontological classification, whereby the relationships between molecular entities or classes of entities and their parents and/or children are specified.
The data is available in OWL and OBO and contains more than 50.000 compounds.

2.12 Cook's thesaurus

The *Cook's Thesaurus*[25] is a cooking encyclopaedia that covers thousands of ingredients and kitchen tools. The entries include pictures, descriptions, synonyms, pronunciations, and suggested substitutions. This encyclopaedia is not maintained anymore.

2.13 FooDB

FooDB[26] is a comprehensive resource on food constituents, covering fruits, spices, vegetables, herbs, etc., but no processed foods. Presently it contains 797 items. FooDB is offered to the public as a freely available resource.

[21] http://openfoodfacts.org/
[22] http://en.wiki.openfoodfacts.org/Main_Page
[23] https://world.openfoodfacts.org/open-food-facts-mobile-app
[24] https://www.ebi.ac.uk/chebi/
[25] http://www.foodsubs.com/
[26] http://foodb.ca/

It provides information on macronutrients and micronutrients. Each entry in the FooDB contains more than 100 separate data fields covering detailed compositional, biochemical and physiological information (obtained from the literature). This includes data on the compound's nomenclature, its description, information on its structure, chemical class, its physico-chemical data, its food source(s), colour, aroma, taste, physiological effect, presumptive health effects (from published studies), and concentrations in various foods. Users are able to browse or search FooDB by food source, name, descriptors, function or concentrations. Depending on individual preferences users are able to view the content of FooDB from the Food Browse (listing foods by their chemical composition) or the Compound Browse (listing chemicals by their food sources). FoodB can be accesed though a publicly available API[27].

2.14 Innova Market Insights

The *Innova Database*[28] is a commercial online food and beverage product database, designed for leading food and beverage manufacturers. It aims to be a source of new product tracking, trends and innovations available anywhere.

[27] https://foodb.ca/api_doc

[28] https://www.innovamarketinsights.com/services/new-product-insights/

3 Standards and vocabularies

Smart software applications can assist decision making by consumers, food designers, retailers and other actors. This is even more true in today's 'Internet of Food' [1]. In order to make data about food products actionable in such applications, it needs to be made understandable for machines, i.e. automated processing. Traditional, text-based standards such as the *GPC Global Product Classification* have been a first step towards harmonizing data from different sources, such that they can be used and combined in different applications. The next step is to move towards standards based on URIs as abstract, unique and dereferenceable identifiers that (automatically) lead to persistent locations containing relevant information. These identifiers can point to objects, concepts, people, etc. but also to relations between them. This allows software to automatically recognize these concepts and reason about them.

When selecting a semantic standard for a given purpose, it is important to decide which type of modelling is needed. We distinguish three types[29]:

- A *controlled vocabulary* describes the terminology used in a selected domain. It provides terms and their synonyms, broader, narrower and related terms. It is a light-weight model, loosely structured and typically expressed in SKOS. It is typically used in search functions on unstructured data sources (such as documents).
- An *ontology* is an abstract model describing classes, relations between them and restrictions on those relations. Ontologies are typically expressed in RDFS/OWL. They support some level of automated reasoning using generic reasoners. They act as layer on top of a dataset, describing the meaning of the variables in that dataset, their provenance and other metadata.
- A *knowledge base* or *knowledge graph* is a repository containing an ontology and instances of the concepts defined in that ontology. It is stored in a triple store such as RDF4J, Virtuoso, GraphDB, etc. It is an alternative to a relational database, adding the power of flexibility and logics.

An important aspect of data modelling is the construction of hierarchical class structures or taxonomies. First, a controlled vocabulary allows weakly defined broader and narrower relations, which may mean 'is kind of', but also 'is part of', or even 'falls in the category of'. Although these readings are possible, for certain applications it is advisable to stick to the 'kind of' interpretation as much as possible. This is also the interpretation used in the more formal ontological interpretation.

Classification is in particular a delicate issue in modelling food products and ingredients. These are produced from natural products. In principle every single product on the shelf is unique and has its own identity, in particular if they are fresh produce. In practice easily many competing classifications are proposed, which can cause confusion. For example, *wholegrain bread* can be classified either as *bread* or as *wholegrain* product. In many cases it is advised to replace classifications by specialised properties; in this case we could say *wholegrain bread - is produced from – wholegrain cereals*. Then *wholegrain bread – is a kind of – bread* can still be maintained as a static classification. Note that in machine learning the term classification is used to refer to ordering of concepts based on any property.

It is good practice to reuse existing vocabularies and ontologies as much as possible, as this facilitates interoperability. In practice always some tweaking is needed to meet the requirements of a particular application. In this section we list a number of publicly available controlled vocabularies and ontologies in the food domain. These models can be a starting point for data providers and data users to share semantically enriched data.

[29] There is no clear consensus about precise definitions of these types, but for most practical cases this categorisation works.

3.1 GS1 standards

The *GPC Global Product Classification*[30] contains a classification of thousands of products from food, beverage and tobacco. It is a text-based controlled vocabulary available in XML, TX and XSLX. Next to a hierarchy at four levels (segment, family, class and brick) it also provides a number of attributes and comments. New versions are published each half year, from which the latest version is only available for registered members. Except for the classification, the data is not very relevant for personalised nutrition applications.

For representing product data in the GS1 data pools, the GDSN Standard[31] has been developed. Properties in GDSN relevant for personalised nutrition are for example: grade, genus, size, species, cooking type, ingredients, inner flesh colour.

The *GS1 Web vocabulary*[32] or GS1 Smart Search standard is an external extension of schema.org[33]. The extension vocabulary builds upon an extensive set of pre-existing B2B standards. While this means that in some places there is some divergence between the GS1 terminology and Schema.org's, they build upon the core vocabulary of schema.org and upon underlying foundational standards from W3C such as JSON-LD. The combination of schema.org and GS1's vocabularies is expected to provide for significantly richer online product descriptions for use in Web search, combining the descriptive depth of GS1 terminology with the broad coverage of Schema.org's.

3.2 LanguaL

LanguaL[34] is a trademarked thesaurus (controlled vocabulary) for describing data about food, specifically for classifying food products for information retrieval [2]. It was used in the USA and Europe for numeric data banks on food composition (nutrients and contaminants), food consumption and legislation. The work on LanguaL was started in the late 1970's in the US. Since 1996, the European LanguaL Technical Committee has administered the thesaurus.

In LanguaL, each food is described by a set of controlled terms chosen from properties characteristic of the nutritional and/or hygienic quality of a food, as for example the biological origin, the methods of cooking and conservation, and technological treatments. In order to make LanguaL language independent it was organized using abstract identifiers based on unique codes, even long before RDF/SKOS was created. Each unique code points to equivalent terms in different languages (e.g. Czech, Danish, English, French, German, Hungarian, Italian, Portuguese and Spanish).

In total, more than 40000 European, North American foods and foods from other countries are now LanguaL indexed. LanguaL 2017 contains a total of 12605 descriptors.

The most current version of LanguaL™ is version 2017. It has been updated extensively - especially in the facets *Product Type* and *Food Source*.

- *Product Type* has been updated according to the GS1 GPC Standard for *Food* and *Beverage*. Similarly, the EFSA FOODEX2 (described later) has been included in LanguaL 2017, following FoodEx2 version Matrix 9.0, and comprises the FoodEx2 Exposure Hierarchy.
- *Food Source* has been updated thoroughly with new descriptors, and existing descriptors have been updated with new information.

[30] https://www.gs1.org/standards/gpc/jun-2019
[31] https://www.gs1.org/standards/gdsn/current-standard
[32] https://www.gs1.org/voc/
[33] http://blog.schema.org/2016/02/gs1-milestone-first-schemaorg-external.html
[34] https://www.langual.org/default.asp

Since LanguaL does not provide specific product attributes (other than the general taxonomic relations), its direct applicability for personalised nutritional advice is limited. However, each entry contains a section ‘AI - additional information about the terms’, which can be useful for access by automated information retrieval.

3.3 Agrovoc

The Agrovoc Multilingual Thesaurus[35,36] is a large controlled vocabulary covering all areas of interest of the Food and Agriculture Organization (FAO) of the United Nations. These areas include food, nutrition, agriculture, fisheries, forestry, environment etc. The vocabulary is published by FAO and edited by a community of experts. It is widely used in specialized libraries as well as digital libraries and repositories to index content and for the purpose of text mining.

AGROVOC consists of 37,000+ concepts and 750,000+ terms in up to 37 languages. Currently, the vocabulary is a SKOS-XL concept scheme, and a Linked Open Data (LOD) set edited by VocBench. AGROVOC is aligned with 18 other multilingual knowledge organization systems. It can be accessed through a SPARQL endpoint [37] and an API[38]. AGROVOC is available as a linked data set and is aligned (linked) with 18 vocabularies related to agriculture. The linked data version of the vocabulary is exposed as RDF and HTML, through a content-negotiation mechanism.

Agrovoc has limited use in applications in personalised nutrition. The classification of food products may be helpful to check whether all relevant products are covered by an application, but the level of detail in the taxonomy is limited. For example ‘whole grain foods’, ‘street foods’ and ‘beef’ are end-nodes, having no narrower terms.

3.4 FoodEx2

EFSA provides a standardised food classification and description system called FoodEx2[39]. The system consists of descriptions of a large number of food items organised in a broader-narrower hierarchy. Central to the system is a core list of food item descriptions that represent the minimum level of detail needed for intake or exposure assessments. More detailed terms can be found on the ‘extended list’. A parent-child relationship exists between a core list food item and its related extended list food items. The current version has seven hierarchies: five domain-specific and a general purpose one available for the users, and a service hierarchy for the management of the terminology. Facets are used to add further detail by describing properties and aspects of foods from various perspectives.

FoodEx2 may be useful for identifying food products and ingredients in personalised nutrition as is has the right level of detail (‘Gouda cheese’, ‘Rye Bread’). However, the hierarchy and labelling of products are not immediately suitable as input for nutritional advice.

3.5 EUroFIR Thesauri

EUroFIR offers a number of controlled vocabularies for food composition data. The specifications are based on a relational database model containing four main and other additional entities. The four main entities (Food, Component, Value, Reference) are mandatory for data documentation and they build the core set that is necessary for proper description of food consumption data. The description of the food is regulated through the LanguaL Thesaurus.

[35] http://aims.fao.org/vest-registry/vocabularies/agrovoc
[36] https://en.wikipedia.org/wiki/AGROVOC
[37] http://agrovoc.uniroma2.it/sparql
[38] http://agrovoc.uniroma2.it:8080/SKOSWS/services/SKOSWS?wsdl
[39] https://www.efsa.europa.eu/en/data/data-standardisation

3.6 FoodOn

FoodOn[40] is an ambitious, worldwide effort to develop public food ontologies [3]. FoodOn aims to create content to cover gaps in the representation of food-related products and processes. This ontology is being applied to research and clinical datasets in academia and government. They also welcome industry uptake since agricultural and consumer devices connected to the Internet of Things will require a standard food vocabulary that has a global, multilingual reach.

FoodOn is compatible with the Basic Formal Ontology (BFO)[41], which means that all the classes provided by FoodOn are organized under BFO classes, and reasoning over the ontology does not lead to BFO-related contradictions. Figure 1 shows some basic concepts and relations defined in FoodOn.

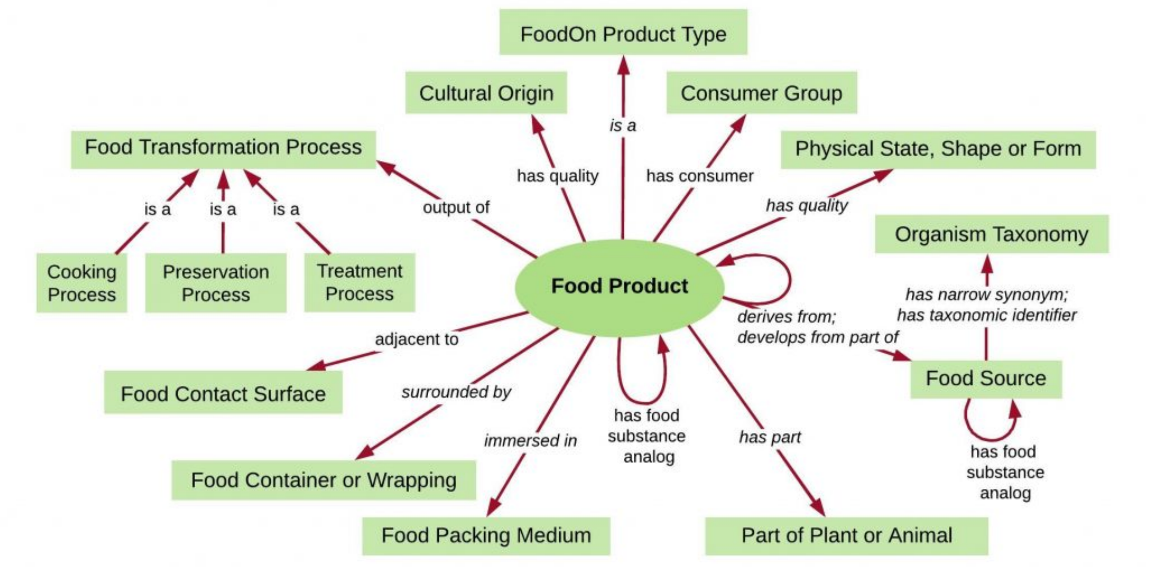

Figure 1 Overview of basic FoodOn concepts

3.7 OM for quantities and units

OM (Ontology of units of Measure)[42, 43] is a general ontology to express quantities, measure and units [4]. Although this ontology is not specific for the food domain, it is mentioned here given the role it can play in personalised nutrition systems. It provides a semantic basis for expressing amounts in food intake and food advice, nutritional values and other food properties, recipes, etc. OM is continuously being maintained and updated as an open source effort. An independent review has concluded that OM is the most comprehensive and detailed ontology for expressing quantities and units [5].
GS1 provides a basic set of units in their UOM_ByCat definition, most of which is already incorporated in OM.

[40] https://foodon.org/
[41] https://basic-formal-ontology.org/
[42] https://github.com/HajoRijgersberg/OM
[43] http://www.foodvoc.org/page/om-2

4 Conclusion

For the Netherlands, nutritional values at the level of generic products can still best be found the NEVO table, but they are also increasingly available at the level of commercial products from the GS1 data source. Many other attributes related to consumer preferences, application context, and sustainable production are scattered, if available at all.

Several other data sources are available for retrieving food data relevant for applications in personalised nutrition. These data sources are still difficult to combine. Therefore the development of semantic standards is crucial. The most promising approach would be to combine the efforts by GS1, representing food supply chains worldwide and FoodOn as a research effort to introduce Linked Data standards in the world of food.

References

[1] M.N.K. Boulos, A. Yassine, S. Shirmohammadi, C.S. Namahoot, and M. Bruckner, "Towards an "Internet of Food": Food Ontologies for the Internet of Things," (in English), *Future Internet,* vol. 7, no. 4, pp. 372-392, Dec 2015, doi: 10.3390/fi7040372.

[2] A. Moller and J. Ireland, "LanguaL™ 2017 – The LanguaL™ Thesaurus," Danish Food Informatics, 2018.

[3] D.M. Dooley *et al.*, "FoodOn: a harmonized food ontology to increase global food traceability, quality control and data integration," *NPJ Sci Food,* vol. 2, p. 23, 2018, doi: 10.1038/s41538-018-0032-6.

[4] H. Rijgersberg, M. van Assem, and J.L. Top, "Ontology of Units of Measure and Related Concepts," *Semantic Web,* vol. 4, pp. 3-13, 2013, doi: 10.3233/SW-2012-0069.

[5] J.M. Keil and S. Schindler, "Comparison and evaluation of ontologies for units of measurement," (in English), *Semantic Web,* vol. 10, no. 1, pp. 33-51, 2019, doi: 10.3233/Sw-180310.

To explore the potential of nature to improve the quality of life

Wageningen Food & Biobased Research
Bornse Weilanden 9
6708 WG Wageningen
The Netherlands
www.wur.eu/wfbr
E info.wfbr@wur.nl

Report 2115

The mission of Wageningen University and Research is “To explore the potential of nature to improve the quality of life”. Under the banner Wageningen University & Research, Wageningen University and the specialised research institutes of the Wageningen Research Foundation have joined forces in contributing to finding solutions to important questions in the domain of healthy food and living environment. With its roughly 30 branches, 6,500 employees (5,500 fte) and 12,500 students, Wageningen University & Research is one of the leading organisations in its domain. The unique Wageningen approach lies in its integrated approach to issues and the collaboration between different disciplines.