Toward improved semantic annotation of food and nutrition data

Lidija Jovanovska
Jožef Stefan International Postgraduate School &
Jožef Stefan Institute
Ljubljana, Slovenia
lidija.jovanovska@ijs.si

Panče Panov
Jožef Stefan Institute &
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia
pance.panov@ijs.si

ABSTRACT

This paper aims to provide a critical overview of the state-of-the-art vocabularies used for semantic annotation of databases and datasets in the domain of food and nutrition. These vocabularies are commonly used as a backbone for creating metadata that is usually used in search. Furthermore, the paper aims to provide a summary of ICT technologies used for storing food and nutrition datasets and searching digital repositories of such datasets. Finally, the results of the paper will provide a roadmap for moving towards FAIR (findable, accessible, interoperable, and reusable) food and nutrition datasets, which can then be used in various AI tasks.

KEYWORDS

ontologies, semantic technologies, data mining, food and nutrition

1 INTRODUCTION

Today more than ever before in history, we live in an age of information-driven science. Vast amounts of information are being produced daily as a result of new types of high-throughput technology in all walks of life. Consequently, the quantity of available scientific information is becoming overwhelming and without its proper organization, we would not be able to maximize the knowledge we harvest from it. Namely, research groups carry out their research in different ways, with specific and possibly incompatible terminologies, formats, and computer technologies. To tackle these issues, researchers have developed high-level knowledge organization systems (KOS), such as ontologies, which constitute the core of the semantic web stack. Throughout the years, an abundance of ontologies has been developed and released, slowly expanding from the biomedical sciences to the fields of information science, machine learning, as well as the domain of food and nutrition science.

There is an old, yet simple saying which goes: "You are what you eat". As the world becomes more globalized and food production grows massively, it is becoming increasingly difficult to track the farm-to-fork food path. In the last few decades, digital technology has been profoundly affecting many health and economic aspects of food production, distribution, and consumption. Issues regarding food safety, security, authenticity as well as conflicts arising from biocultural trademark protection are issues that were further enhanced by the lack of a centralized food data repository without which there is a great difficulty in achieving cross-cultural and expert consensus. [1]

In this paper, we will briefly go through the fundamental components of the Semantic Web technologies, as well as the standards for the development of high-level KOS (Section 2). Next, we provide a critical overview of the most significant semantic resources in the domain of food and nutrition (Section 3). Finally, we present a proposal for the design and implementation of a broad ontology that would allow us to harmonize and integrate reference vocabularies and ontologies from different sub-areas of food and nutrition (Section 4).

2 BACKGROUND

The goal of the Semantic Web is to make Internet data machine-readable by enhancing web pages with semantic annotations. Linked data is built upon standard web technologies, also including semantic web technologies in its technology stack [11]. **Resource Description Framework (RDF)** allows the representation of relationships between entities using a simple subject-predicate-object format known as a triple. The triples form an RDF database — called a triplestore — which can be populated with RDF facts about some domain of interest. **RDF Schema (RDFS)** was developed immediately after the appearance of RDF as a set of mechanisms for describing groups of related resources and the relationships between them. **Simple Protocol and RDF Query Language (SPARQL)** is the query language for querying RDF triples stored in RDF triplestores.

The Web Ontology Language (OWL) is based on Description Logics, a family of logics that are expressively weaker than First Order Logic, but enjoy certain computational properties advantageous for purposes such as ontology-based reasoning and data validation. Most of the ontologies used today are represented in the OWL format.

All the semantic technologies operate on top of various KOS. A KOS is intended to encompass all types of schemes for organizing information and promoting knowledge management [7]. One example of a KOS is a thesaurus as a structured, normalized, and dynamic vocabulary designed to cover the terminology of a field of specific knowledge. It is most commonly used for indexing and retrieving information in a natural language in a system of controlled terms. When looking at the expressiveness of a KOS, a thesaurus is on the lower side of the scale. On the other side, ontologies enjoy greater expressiveness than thesauri due to the inclusion of description logics. Arp, Smith, and Spear define the term ontology as "A representation artifact, comprising a taxonomy as proper part, whose representations are intended to designate some combination of universals, defined classes, and certain relations between them" [1].

Information Society 2020, 5–9 October, 2020, Ljubljana, Slovenia

[1] https://www.nature.com/scitable/knowledge/library/food-safety-and-food-security-68168348/, accessed 22/04/2020

The Open Biomedical Ontologies (OBO) Foundry applies the key principles that ontologies should be open, orthogonal, instantiated in a well-specified syntax, and designed to share a common space of identifiers. Open means that the ontologies should be available for use without any constraint or license and also receptive to modifications proposed by the community. Orthogonal means that they ensure the additivity of annotations and compliance with modular development. The proper and well-specified syntax is expected to support algorithmic processing and the common system of identifiers enables backward compatibility with legacy annotations as the ontologies evolve [17].

The FAIR guiding principles for scientific data management and stewardship were conceived to serve as guidelines for those who wish to enhance the reusability and invaluableness of their data holdings [19]. The power of these principles lies in the fact that they are simple and minimalistic in design and as such can be adapted to various application scenarios. *Findability* ensures that a globally unique and persistent identifier is assigned to the data and the metadata which describes the data. *Accessibility* ensures that the data and the metadata can be retrieved by their identifier using a standardized communications protocol. *Interoperability* ensures that data, as well as metadata, use a formal, accessible, and shared language for knowledge representation. *Reusability* ensures that data and metadata are accurately described, released with a clear and accessible license, have detailed provenance, and meet domain-relevant community standards.

3 CRITICAL OVERVIEW OF FOOD AND NUTRITION SEMANTIC RESOURCES

In this section, we provide a critical overview of the most relevant KOS in the field of food and nutrition. We start by describing LanguaL [8], a thesaurus that serves as a foundation for most of the ontologies in this domain. We are more focused on analyzing ontologies which belong to different sub-spheres of the food and nutrition domain. Namely, FoodOn [4], as a more general food description ontology, ONS [18], relevant in the field of nutritional studies and ISO-Food [6], relevant in the field of annotating isotopic data acquired from food samples.

LanguaL [8] is a thesaurus used for describing, capturing, and retrieving data about food. Since 1996, it has been used to index numerous European Union (EU) and US agency databases, among which, the US Department of Agriculture (USDA) Nutrient Database for Standard Reference and 30 European Food Information Resource (EuroFIR) databases. Food ingredients are represented with indexing terms, preferably in the form of a noun or a phrase. The thesaurus also includes precombined terms which are food product names to which facet terms have been assigned. There are 4 main facets in LanguaL: A (Product Type), B (Food Source), C (Part of Plant or Animal), and E (Physical State, Shape, or Form). Other food product description facets include chemical additive, preservation or cooking process, packaging, and standard national and international upper-level product type schemes.

The LanguaL thesaurus complies with the FAIR guidelines. The completeness of LanguaL's indexing is to a large extent assured by the Langual Food Product Indexing (FPI) software, which verifies that all facets have been indexed for each food in the list [8]. It is available online[2] and can be queried using a food descriptor or synonym. Its interoperability and reusability are eminent as it represents a cornerstone in the development of more sophisticated ontologies, such as FoodOn. Even though the OBO Foundry principles apply only to ontologies, we can use the more general ones as evaluation criteria for the LanguaL thesaurus. For instance, as previously mentioned, the thesaurus is open, made available in an accepted concrete syntax, versioning is ensured, textual definitions are available for all the terms and a sufficient amount of documentation is provided.

FoodOn [4] is an open-source, comprehensive ontology composed of term hierarchy facets that cover basic raw food source ingredients, process terms for packaging, cooking, and preservation, and different product type schemes under which food products can be categorized. FoodOn is applicable in several use-cases, such as personalized foods and health, foodborne pathogen surveillance and investigations, food traceability and food webs, and sustainability. FoodOn echoes most of LanguaL's plant and animal part descriptors —– both anatomical (arm, organ, meat, seed) and fluid (blood, milk) —– but reuses existing Uberon [12] and Plant Ontology [10] term identifiers for them. Multiple component foods are more challenging because LanguaL provides no facility for giving identifiers to such products.

Building on top of this, FoodOn allows food product terms like lasagna noodle to be defined directly in the ontology, and allows them to reference component products through various relations which do not exist in LanguaL, such as: "has ingredient", "has part", "composed primarily of". As a suggestion, these relations can all be represented with a single relation "has ingredient" and the quantity can be expressed explicitly when annotating the objects. All of the ontology terms have unique identifiers and the ontology is accessible and can be searched via The European Bioinformatics Institute (EMBL-EBI) and its Ontology Lookup Service (OLS).[3] The ontology itself is open-source and is a member of the OBO Foundry. It also includes the upper-level Basic Formal Ontology (BFO) [1]. The adherence to BFO proves useful in the case of aligning ontologies covering different domains because they share the same top-level.

ONS [18] is the first systematic effort to provide a solid and extensible ontology framework for nutritional studies. ONS was built to fill the gap between the description of nutrition-based prevention of disease and the understanding of the complex impact nutrition has on health. Its structure consists of 3334 terms imported from already existing ontologies and 100 newly defined terms. The usability of ONS was tested in two scenarios: an observational study, which aims at developing novel and affordable nutritious foods to optimize the diet and reduce the risk of diet-related diseases among groups at risk of poverty, and an intervention study represented by the impact of increasing doses of flavonoid-rich and flavonoid-poor fruit and vegetables on cardiovascular risk factors in an "at risk" group study.

The development of ONS followed FAIR principles and as a result, it has been published in the FAIR-sharing database.[4] Before defining new terms, the developers of ONS have ensured that they are not yet defined, with the use of the ONTOBEE web service. Terms that were already defined were imported using the ontology reuse service — ONTOFOX [20]. In compliance with the OBO Foundry principles, the ONS has been developed to be interoperable with other ontologies, as it has been formalized

[2]https://www.langual.org, accessed 22/04/2020

[3]https://www.ebi.ac.uk/ols/ontologies/FoodOn, accessed 22/04/2020

[4]https://fairsharing.org/bsg-s001068/, accessed 22/04/2020

using the latest OWL 2 Web Ontology Language and RDF specifications and edited using Protégé [13] and the Hermit reasoner for consistency checking. It is also accessible, under the Creative Commons license (CC BY 4.0), published on GitHub and at NCBO BioPortal. Moreover, this ensured the adoption of a well-defined and widely adopted structure for the top and mid-level classes and principally the adherence to BFO as upper-level ontology.

ISO-Food is an ontology that was conceived to aid with the organization, harmonization, and knowledge extraction of datasets containing information about isotopes, that represent variants of a particular chemical element which differ in neutron number. To develop this ontology a mixed approach was used, a combination of both expert knowledge-driven (bottom-up) and data-driven (top-down) methods. Its main classes include Isotope, Sample, Location, Measurement, Article. The main class Isotope is connected to the rest of the classes with respective relations. The Food and Nutrient classes are linked to the RICHFIELDS ontology [5]. The ontology was further applied in a study for describing isotopic data, to annotate a data sample that consists of isotopic measurements of milk and potato samples.

The ISO-Food ontology can be accessed online via the BioPortal repository of biomedical ontologies.[5] It reuses terms from several ontologies, such as the concept Unit from the Units of Measurements Ontology (UO), the classes Food and Component from the RICHFIELDS ontology [5], the class Document from the Bibliographic Ontology (BIBO) [3].

4 PROPOSAL

Ontologies for data mining. To provide a suitable formalized representation of the outcomes of the research in the food and nutrition domain, as well as to suggest new ways to extract knowledge from the ever-abundant data produced in this field, we turn to ontologies that are used to formally represent the data analysis process. More specifically, we focus on the **OntoDM** ontology, which provides a unified framework for representing data mining entities. It consists of three modular ontologies: **OntoDM-core** [15] which represents core data mining entities, such as datasets, data mining tasks, algorithms, models and patterns, **OntoDT** [16] — a generic ontology of datatypes, and **OntoDM-KDD** [14] which describes the process of knowledge discovery.

The ontology defines top-level concepts in data mining and machine learning, such as data mining task, algorithm, and their generalizations, which denote the outputs of applying an implementation of an algorithm on a particular dataset. Starting with these general concepts, OntoDM also defines the components of the algorithms, such as distance and kernel functions, and other features they may contain. From the input and output data perspective, in this ontology, there is a hierarchical representation of data, from general concepts such as dataset to more specific concepts regarding its structure, such as the number of features, their role in a given task, concluding with the datatype of each attribute. These properties of OntoDM provide a complete formal representation of the data mining process from beginning to end.

Combining orthogonal domain ontologies. Our goal is to align the selected ontologies in the domain of food and nutrition with the OntoDM ontology of data mining to improve the semantic annotation of the food and nutrition domain datasets, as well as to formally represent data analysis tasks performed in the domain of food and nutrition (see Figure 1). In this way, we can also use the benefits of cross-domain reasoning. Since FoodOn, ONS, and OntoDM all use BFO as a main top-level ontology, they speak the same general language and are consequently, easier to align.

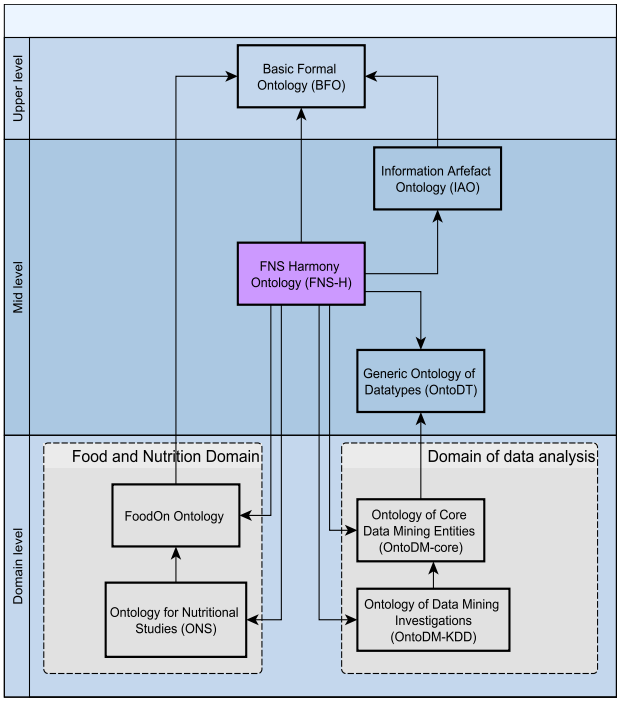

Figure 1: Diagram representing the alignment of the proposed ontology with the identified relevant upper-level and domain ontologies.

Towards the FNS Harmony ontology. In the context of the H2020 project FNS Cloud[6] (food, nutrition, security) the goal is to develop an infrastructure and services to exploit food, nutrition and security data (data, knowledge, tools – resources) for a range of purposes. To support the different functionalities required by the cloud platform, we started with the development of the FNS-Harmony (FNS-H). The application ontology would allow us to harmonize and integrate the different reference vocabularies and ontologies from different sub-areas of food and nutrition, as well as ontologies representing the domain of data analysis.

Initial ontology development. The development of FNS-H, which is intended to bridge the gap between the field of data analysis and food and nutrition will be guided by common best practice principles for ontology development. The aim is to maximize the reuse of available ontology resources and simultaneously follow the Minimum Information to Reference an External Ontology Term (MIREOT) principles [2]. In the first phase, we will integrate the FoodOn ontology and the ONS ontology with the OntoDM suite of ontologies. With this integration, we will be able to (1) define domain-specific data types for the domain of food and nutrition by extending OntoDT generic data types; (2) define food and nutrition analysis pipelines for the domain of food and nutrition by extending OntoDM-core, and (3) define

[5]http://bioportal.bioontology.org/ontologies/ISO-FOOD, accessed 22/04/2020

[6]https://www.fns-cloud.eu/

food and nutrition knowledge discovery scenarios by extending OntoDM-KDD ontology.

The development of the ontology already started in a top-down fashion, it is expressed in OWL2 and being developed using the Protégé ontology development tool. Aspiring to maximize accessibility, the ontology will be available for access on a GitHub repository, [7] as well as via BioPortal. In the current stage of development, an initial set of higher-level domain terms, data types, data formats, data provenance metadata, lists of external ontologies and vocabularies were extracted from the literature and FNS-Cloud project documents.

In the next steps, we will first align the extracted terms with the BFO ontology and then integrate them with domain terms from the domain ontologies based on BFO, such asFoodOn, and ONS, at the first instance, as well as with the OntoDM set of ontologies. Other potentially relevant ontologies include the Ontology for Biomedical Investigations (OBI), Ontology of Biological and Clinical Statistics (OBSC), Ontology of Chemical Entities of Biological Interest (ChEBI), Ontology of Statistical Methods (STATO), and others. To achieve integration of different ontological resources, we will use the ROBOT tool [9] that supports the automation of a large number of ontology development tasks and helps developers to efficiently produce high-quality ontologies.

5 CONCLUSION

In this paper, we provided an overview of the most relevant knowledge organization systems in the domain of food and nutrition. We started with the LanguaL food thesaurus that served as a foundation for the development of the more sophisticated ontologies — FoodOn, used for a multi-faceted description of various foods; ONS, used for observational and interventional nutrition studies; ISO-Food for the studies of isotopic data in foods. Next, we assessed the selected vocabularies with respect to the FAIR principles and OBO Foundry guidelines for scientific data management. All of the selected vocabularies showed compliance with these accomplishment criteria, with only minor suggestions for improvement provided from our side. Finally, in our proposal, we lay down the foundations of a new ontology which would connect data mining concepts in the domain of food and nutrition using domain ontologies (FoodOn, ONS) with ontologies for datatypes, data mining, and knowledge discovery in databases (OntoDT, OntoDM-core, OntoDM-KDD). By doing so, we can provide richer semantic annotation and discover new scenarios of harvesting knowledge from the food and nutrition data.

ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency through the grant J2-9230, as well as the European Union's Horizon 2020 research and innovation programme through grant 863059 (FNS-Cloud, Food Nutrition Security).

REFERENCES

[1] Robert Arp, Barry Smith, and Andrew D Spear. 2015. *Building ontologies with basic formal ontology*. Mit Press.

[2] Mélanie Courtot, Frank Gibson, and Allyson L Lister et al. 2011. Mireot: the minimum information to reference an external ontology term. *Applied Ontology*, 6, 1, 23–33.

[3] Bojana Dimić Surla, Milan Segedinac, and Dragan Ivanović. 2012. A bibo ontology extension for evaluation of scientific research results. In *Proceedings of the Fifth Balkan Conference in Informatics*, 275–278.

[4] Damion M Dooley, Emma J Griffiths, and Gurinder S Gosal et al. 2018. Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration. *npj Science of Food*, 2, 1, 1–10.

[5] Tome Eftimov, Gordana Ispirova, and Peter Korosec et al. 2018. The richfields framework for semantic interoperability of food information across heterogenous information systems. In *KDIR*, 313–320.

[6] Tome Eftimov, Gordana Ispirova, and Doris Potočnik. 2019. Iso-food ontology: a formal representation of the knowledge within the domain of isotopes for food science. *Food chemistry*, 277, 382–390.

[7] Heather Hedden. 2016. *The accidental taxonomist*. Information Today, Inc.

[8] Jayne D Ireland and A Møller. 2010. Langual food description: a learning process. *European journal of clinical nutrition*, 64, 3, S44–S48.

[9] Rebecca C Jackson, James P Balhoff, and Eric Douglass. 2019. Robot: a tool for automating ontology workflows. *BMC bioinformatics*, 20, 1, 407.

[10] Pankaj Jaiswal, Shulamit Avraham, and Katica Ilic et al. 2005. Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comparative and functional genomics*, 6, 7-8, 388–397.

[11] Brian Matthews. 2005. Semantic web technologies. *E-learning*, 6, 6, 8.

[12] Christopher J Mungall, Carlo Torniai, and Georgios V Gkoutos et al. 2012. Uberon, an integrative multi-species anatomy ontology. *Genome biology*, 13, 1, R5.

[13] Mark A Musen. 2015. The protégé project: a look back and a look forward. *AI matters*, 1, 4, 4–12.

[14] Panče Panov, Larisa Soldatova, and Sašo Džeroski. 2013. Ontodm-kdd: ontology for representing the knowledge discovery process. In *International Conference on Discovery Science*. Springer, 126–140.

[15] Panče Panov, Larisa Soldatova, and Sašo Džeroski. 2014. Ontology of core data mining entities. *Data Mining and Knowledge Discovery*, 28, 5-6, 1222–1265.

[16] Panče Panov, Larisa N Soldatova, and Sašo Džeroski. 2016. Generic ontology of datatypes. *Information Sciences*, 329, 900–920.

[17] Barry Smith, Michael Ashburner, and Cornelius Rosse et al. 2007. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25, 11, 1251–1255.

[18] Francesco Vitali, Rosario Lombardo, and Damariz Rivero et al. 2018. Ons: an ontology for a standardized description of interventions and observational studies in nutrition. *Genes & nutrition*, 13, 1, 12.

[19] Mark D Wilkinson, Michel Dumontier, and IJsbrand Jan Aalbersberg et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3.

[20] Zuoshuang Xiang, Mélanie Courtot, and Ryan R Brinkman et al. 2010. Ontofox: web-based support for ontology reuse. *BMC research notes*, 3, 1, 175.

[7] https://github.com/panovp/FNS-Harmony