Category: Ontology ecosystem to support research and operational adoption

Ontology for Nutritional Epidemiology (ONE): a standards-based framework to FAIR share nutritional epidemiologic data

Prepared by Chen Yang[1] & Carl Lachat[1]

1 Department of Food Technology, Safety and Health, Ghent University, Ghent, Belgium

Background

Nutritional epidemiology assesses nutritional causes of health and disease in humans and typically relies on observational evidence combined with findings from human intervention studies. Joint analysis of secondary data in nutritional epidemiologic research can increase power to detect often small effects of nutrition or diet on human health. To date, however, there is lack of tools to index data/study characteristics for data retrieval and data reuse. Data standards and manuscript reporting guidelines summarize essential data and study characteristics. Meanwhile, an ontology, as an information representation frame in computer science, enables machine processing of semantic information. For nutritional epidemiology, a standards-based ontology can help to track data/study characteristics for identification and integration of data from multiple nutritional epidemiologic studies. An ontology that deals with study level information complement existing work at the data level.

To facilitate joint data analysis in nutritional epidemiology, we developed the Ontology for Nutritional Epidemiology (ONE) using three authoritative standards: 1) minimal information requirements of nutritional epidemiologic data; 2) quality descriptors of nutritional epidemiologic data and 3) the STROBE-nut (strengthening the reporting of observational studies in epidemiology—an extension for nutritional epidemiology) reporting guidelines.

The ONE is not a novel controlled vocabulary, but an information representation framework. The ONE, hence, uses as many existing ontology terms as possible while limiting the introduction of new terms. The FAIR principles (Findable, Accessible, Interoperable, and Reusable) provide an important framework to establish the European Open Science Cloud. The ONE can contribute to make research data and manuscripts FAIR in such open science platform, and enable graphical knowledge representations, logical reasoning, natural language processing, etc. in nutritional epidemiology.

Methods

a) In order to code the ONE, we first reviewed existing ontologies in epidemiology, as well as ontologies in the relevant disciplines such as food science, nutrition science, etc. All ontologies in the three main medical ontology libraries "OBO Foundry", "BioPortal", and "Ontology Lookup Service" were reviewed in 2018, and an update of the review was carried out in 2019. Existing ontologies were selected and classified if they met part of the controlled vocabulary requirement of nutritional epidemiology.

b) Second, we developed the Ontology for Nutritional Epidemiology (ONE) according to the three authoritative guidelines in nutritional epidemiology. In case a required term was found in more than one existing ontology, the existing term with the most appropriate description was selected. When no exact terms were found in the selected ontologies, a synonym term was selected or a new term was added by the involved domain experts.

c) Third, we conceptualized a graph database for nutritional epidemiologic research based on the DIKW (Data, Information, Knowledge, and Wisdom) pyramid, an established model in information science. The ONE as well as other authoritative ontologies provided the nodes and edges of the graph database. The data findability was showcased using the SPARQL and SQWRL query languages.

d) Finally, we developed a use case of the ONE in Python Virtual Environment. A Web Crawler was developed to extract information from nutritional epidemiologic research articles (n=15) through the Springer Nature application programming interface (API) portal. The extracted information was stored in a graph database, and the statistics were computed and visualized.

Result

a) In total, 237 ontologies were selected and classified from the 1146 reviewed ontologies. Of the 237 ontologies, 158 ontologies were for data annotation (33 for food/dietary agricultural products, 4 for nutrients/chemical compounds, 100 for disease and specific population, and 21 for data management), and 35 were for research terminology. Besides, 44 ontologies were selected to describe supplementary (meta) data such as ethical issues, demographics, etc. However, among the 1146 reviewed ontologies, no ontology was developed as a metadata representation framework.

b) The ONE consists of 339 classes. It reuses terms from 22 existing ontologies. The main referred ontologies are the NCIT (National Cancer Institute Thesaurus, 43 classes) and the MeSH (Medical Subject Headings, 33 classes). Recommendations were given to the required terms that could not be found in the corresponding ontologies of other subjects. ONE proposes 79 new classes to describe nutrition data and 24 new classes to describe manuscript content. The structures of the three referred standards were used to build taxonomic hierarchies of the ontology. For instance, the STROBE-nut recommendations were arranged under their corresponding STROBE reporting items.

c) Authoritative ontologies were suggested to construct nodes and edges of the DIKW pyramid for nutritional epidemiology: 1) Digital Object Identifiers (DOIs) were suggested to identify research manuscripts; 2) the Ontology for Nutritional Epidemiology (ONE) was suggested to annotate research manuscripts; 3) the Food Ontology (FoodOn) and the Human Disease Ontology (DOID) were suggested to annotate diet and disease, respectively; 4) the OBO Relations Ontology (RO) and the Dublin Core Metadata Initiative (DCMI) Metadata Terms were suggested to clarify relationships (edges); 5) the class “foaf:Person” from the FOAF ontology (Friend Of A Friend) was suggested to identify recruited participants. Finally, information findability of the graph database was showcased by using both SPARQL and SQWRL query languages.

d) A sample of 15 scientific articles was annotated according to the STROBE-nut reporting guidelines. The extracted article metadata as well as their STROBE-nut annotations were stored in a Neo4j graph database. The content structure, presence of essential study characteristics as well as the reporting completeness of the articles was visualized automatically from the graph database. The archived linked data were interoperable through their annotations and relations.

Discussion

We introduced an ontology for nutritional epidemiologic research. The present work is an extension of the Ontology for Nutritional Studies (ONS) that was developed as a comprehensive ontology to describe research in a broader human nutrition domain. The present work adds value to FoodOn, which formulates descriptions of nutrients, food items and diets.

To the best of our knowledge, it is the first time that an ontology is developed based on reporting guidelines. Reporting guidelines are widely endorsed and applied by nutrition journals to improve reporting completeness of articles. However, currently, reporting guidelines remain a paper-based initiative. Converting paper-based guidelines into a machine-readable framework can expand the use of reporting guidelines to knowledge tracking and logical reasoning.

For instance, the ONE could be applied to monitor article reporting completeness. It can help to identify the frequently and rarely reported STROBE-nut items and where they are reported in manuscripts. Other potential applications include visualising trends in nutritional epidemiology and identification of neglected research areas. This would, however, require development and integration of relevant ontologies and a collaborative effort of nutrition scientists, data scientists and computer scientists.

Keywords: nutritional epidemiology, the STROBE-nut reporting guidelines, ontology, research semantics, data standardization, data retrieval, data visualization