Semantic Annotation of TSOTSATable Dataset*

Azanzi Jiomekong[1], Uriel Melie[1,*], Hippolyte Tapamo[1] and Gaoussou Camara[4]

[1]Department of Computer Science, University of Yaounde I
[4]EIR-IMTICE, University Alioune Diop de Bambey, Sénégal

Abstract

Food Composition Table (FCT) or Food Composition Databases (FCD) are composed of tables that describe foods and its composition. During SemTab 2022, we proposed a Food Composition Table dataset that we called TSOTSATable dataset. In this paper, we present how the annotation of this dataset is being done using Wikidata, FoodOn and Open Research Knowledge Graph (ORKG). The extracted tables are annotated using Wikidata and FoodOn. The scientific papers from which knowledge are extracted are annotated using Open Research Knowledge Graph. The annotation consists of matching the cells of the food tables to the Wikidata Knowledge Graph and FoodOn ontology, the matching of elements of the table describing the scientific papers to ORKG resources, the detection of the type of elements of each columns (CTA) and the matching of the relations between columns to ORKG properties. During the annotation, we found that many tables were not relevant to the food domain. Thus, we added a new annotation task that is Irrelevant Table Detection (ITD). This consists for a domain dataset, to determine the tables that are not relevant to this domain.

Keywords

Food Science and Nutrition, Food information engineering, Food Composition Tables, Semantic Table Annotation, TSOTSATable dataset,

1. Introduction

Food Composition Tables (FCT) or Food Composition Databases (FCD) are used for a variety of purposes such as clinical practices, public health/education and nutrition monitoring, food industry, food regulation, research, etc. [1]. In the rest of this paper, we'll be using FCT to designate both FCT and FCD. FCT are constructed using a direct method based on chemical analysis, an indirect method based on existing data and scientific literature or the combination of both [2]. Given the high cost of the direct method, indirect method is generally used [2, 3]. This consists of extracting food knowledge from existing resources and using these resources to build FCT. During SemTab@ISWC 2022, we proposed a dataset composed of tables extracted from several data sources [4]. To extract tables from Food Composition documents in PDF format, we used Neural Networks (NN) algorithms. The source code we're using for the extraction of tables from PDFs documents is licensed under Apache License, Version 2.0 and is available on

SemTab'23: Semantic Web Challenge on Tabular Data to Knowledge Graph Matching 2023, co-located with the 22nd International Semantic Web Conference (ISWC), November 6-10, 2023, Athens, Greece
*Corresponding author.
fidel.jiomekong@facsciences-uy1.cm (A. Jiomekong); uriel.melie@facsciences-uy1.cm (U. Melie); hippolyte.tapamo@facsciences-uy1.cm (H. Tapamo); gaoussou.camara@uadb.edu.sn (G. Camara)
0000-0002-0877-7063 (A. Jiomekong)

GitHub[1] and Google Collaboratory[2]. a video showing how we automatically extract tables from PDFs is also available[3].

A study of the TSOTSATable dataset showed that the data extracted from scientific papers were more complete than the ones extracted from Zenodo. On the other hand, Zenodo contains a lot of tables not relevant to the domain of food science and nutrition. Thus, we decided to start the annotation of tables extracted from scientific papers.

In this paper, we describe how this dataset is being curated using Wikidata, FoodOn [5] and Open Research Knowledge Graph (ORKG) [6]. Raw data is available on GitHub[4] and an excerpt of this dataset, containing data extracted from papers of the "Journal of Food Composition Tables" are already annotated[5] and published on Zenodo repository [7]. The source code of the TSOTSATable system used for the annotation is published under MIT license. We provided two source code: one for the annotation of the tables using Wikidata and FoodOn[6] and one for the annotation of scientific papers[7]. The dataset is published under Creative Commons Attribution-ShareAlike 4.0 International License[8].

In the rest of this paper, we present the annotation in Section 2, the overview of the dataset in Section 3 and the conclusion in Section 4.

2. TSOTSATable annotation

The curation started by the annotation of an excerpt of the TSOTSATable dataset, containing data extracted from papers of the "Journal of Food Composition and Analysis". To this end, two tools were designed: one for the automatic annotation using Wikidata and FoodOn and one for automatic annotation using Open Research Knowledge Graph. In the following sections, a short overview of the ontology and Knowledge Graphs used for the annotation is presented (see Section 2.1). Thereafter, the different annotation tasks are presented in Section 2.2 and finally, the annotation process is presented 2.3.

2.1. Ontology and Knowledge Graphs Annotations

To annotate the TSOTSATable dataset, we choose different vocabularies corresponding to the different domains covered by the dataset. The dataset is composed of two types of tables: the tables describing foods and their composition (food science and nutrition domain) and a table containing the list of scientific papers (digital library domain) from which the tables are extracted.

The first step of the annotation consists of the selection of relevant KGs or ontologies to use. Given that this is a food domain, we compared several excerpts of the dataset with several food

[1] https://github.com/Neuralearn/pdf-to-excel
[2] https://colab.research.google.com/drive/1gOPBCVO9VtKcoIewXyr_6nNoxo1Bkqbz
[3] www.youtube.com/watch?v=HZh31OGiQRQ
[4] https://github.com/jiofidelus/tsotsa/tree/main/TSOTSATable_dataset/rawData
[5] https://github.com/jiofidelus/tsotsa/tree/main/TSOTSATable_dataset/annotatedData
[6] https://github.com/jiofidelus/tsotsa/tree/main/TSOTSATable
[7] https://github.com/jiofidelus/SemTabTable-Papers/tree/main/sourceCode
[8] https://creativecommons.org/licenses/by-sa/4.0/

Figure 1: An example of comparison of an excerpt of the TSOTSATable dataset with several food ontologies

ontologies hosted on Bioportal. We used the ontology recommender of Bioportal to search for the most appropriate ontology in the food domain that can be used to annotate the dataset. Fig. 1 presents an example of the use of ontology recommender to search for the most appropriate ontology.

From the ontology recommender, we found that FoodOn is the most appropriate ontology to annotate the dataset. FoodOn[9] [5] is an OBO Foundry ontology used to describe domestical animal food, animal and plant food sources, food categories and products, etc. The FoodOn ontology can be explored using several ontology lookup services. In our case, we used the Ontology Lookup Service[10] (OLS). OLS is a repository of several biomedical ontologies. We used the OLS API to search for relevant annotations and annotate the dataset. To this end, we first search for the list of all CEA given a cell. Thereafter, we identify for each CEA their CTA. Finally, we identify the entity to which the majority of the cells are linked to and vote this as the CTA. The entities linked to the entity voted as the CTA and corresponding to the entities

[9] http://foodon.org
[10] https://www.ebi.ac.uk/ols/docs/api

found during the lookup are designated as the CEA of the different cells of the table. To improve the results obtained after the automatic annotation, a PhD in Food Science and Nutrition is currently checking the annotated dataset.

On the other hand, we manually searched for a set of terms in the Wikidata KG using its search engine and we found that Wikidata contains a lot of relevant annotations. Wikidata[11] is amongst the most popular KGs in the world. It is involved in the SemTab challenge since the challenge was launched in 2019. Once we found that Wikidata contain relevant annotations, we built an automatic tool for the annotation of the dataset using the Wikidata MediaWiki API[12]. The same disambiguation process used during the annotation of the dataset by FoodOn ontology was used to select amongst the entities the ones that may match to the elements of the table.

Concerning the annotation of scientific papers from which tables are extracted, we rely on ORKG because we have a great experience on the use of this KG for annotating scientific papers. Open Research Knowledge Graph[13] (ORKG) is a scholarly KG used to acquire, publish and process structured scholarly knowledge published in the scholarly literature. It is built according to the principle of Open Science, Open Data, and Open Source. We used the automatic annotation feature of ORKG to annotate all the scientific papers from which the tables were extracted.

2.2. Annotations tasks

During the annotation process, we found that many tables were not relevant to the domain of nutrition. On the other hand, ORKG is based on an ontology. This ontology describes a research paper as composed of paper metadata and its semantic description. The semantic description consists of (1) assigning ORKG classes to the different key-insights extracted, (2) defining several properties for comparing research contributions, (3) and comparison tables of research contributions dealing with the same research problem. From this ontology, instances are instantiated during the paper annotation. Based on this, and the annotations tasks generally proposed by SemTab challenge, we defined the following annotations tasks:

- **Column Entity Annotation (CEA):** This is to match each cell of the tables to the ontology/KG entity. The entities in the tables extracted were matched to Wikidata and FoodOn. Concerning scientific papers, we used ORKG resources, which can be a class, an instance, or a property.
- **Column Type Annotation (CTA):** this consists of the assignment of classes from the ontology and KGs to columns of the tables.
- **Column Property Annotation (CPA):** This is the assignment of a property to the relationship between two columns in tables. We found it difficult to identify properties amongst columns of the tables. In effect, the majority of these tables contain numbers in the cells and sometimes in the headers, abbreviations of food components (for instance, k=potassium, Fe=Fer, Mn=Manganese, etc.) The fact that the columns are filled with only

[11] https://www.wikidata.org/
[12] https://www.wikidata.org/w/api.php
[13] https://orkg.org/

numbers make it difficult to build an automatic tool for determining the relations between two columns. Thus, in the current version of the dataset, this annotation task concerns only the scientific papers.
- **Irrelevant Table Detection (ITD):** this task consists of the detection of tables that are not relevant to the domain of Food and Nutrition. It should be noted that this task is currently manual.

2.3. Annotations process

The raw data contained the following types of files:

- TSOTSATable source: this is a CSV file containing information on scientific papers from which the tables were extracted. It is named $0 - KNSR.csv$ in the dataset.
- TSOTSATable files: these are the CSVs files containing the tables extracted from the scientific papers. Each file is named using a unique identifier. The latter allows linking the file to the corresponding source file in the knowledge source file. The file name of each table is obtained using his ID in the data source plus a number denoting the order of its apparition in the data source. For instance, the 3^{rd} table in a scientific paper that has the $ID = KNSR12$ is named $KNSR\_3$.

Concerning the annotation, we created three folders corresponding to the three vocabularies used to annotate the TSOTSATable dataset. Each folder contains different target annotations:

- TSOTSATable_CEA: this is the file containing the CEA of the tables.
- TSOTSATable_CTA: this is the file containing the CTA of the tables.
- TSOTSATable_CPA: this is the file containing the CPA of the tables.

3. Annotated Dataset overview

A subset containing 251 tables were annotated and published on Zenodo repository [7]. This subset contains:

- 38 irrelevant tables,
- 212 relevant tables,
- One table corresponding to the scientific reference from which the tables have been extracted.

Food Composition tables were annotated using Wikidata and FoodOn and the scientific papers from which data is extracted was annotated using Open Research Knowledge Graph. Table 1 presents the number of entities and types annotated using Wikidata and FoodOn. An expert in Food Science and Nutrition was invited to select these annotations randomly and verify their relevance. Concerning scientific papers, around 500 terms were annotated using ORKG.

Table 1
Statistics on the dataset

	# Entities	# Types	# NIL entities	# NIL types
Wikidata	2371	848	9619	678
FoodOn	2928	1166	8498	360

4. Conclusion

In a recent work, we extracted Food Composition data from scientific papers and we built a tabular dataset with it [1]. This paper presents how this dataset is being annotated using Wikidata, FoodOn and Open Research Knowledge Graph. To this end, Cell Entity Annotation (CEA), Column Type Annotation (CTA), Column Property Annotation (CPA) and Relevant Table Detection (RTD) tasks are considered. The first three tasks are well known Semantic Table Annotation tasks. However, the last one were found during the annotation process. In fact, the table extraction tool extracts all the tables that the scientific paper contains. However, some tables are not relevant to the Food Science and nutrition domain. Thus, we introduce this new task. We found many NULL annotation, due to the fact that many entities does not have reference to Wikidata and FoodOn. It should be noted that the detection of irrelevant tables is still done manually. We are planning to develop an additional module which allow to automatically detect the tables that are relevant to the Food and nutrition domain before their annotation.

Future work consists of finalizing the annotation and using this dataset to build a TSOTSA-Graph, a Food Composition Knowledge Graph.

References

[1] A. Jiomekong, B. Foko, Towards an approach based on knowledge graph refinement for tabular data to knowledge graph matching, 2022, pp. 111–122.

[2] H. Greenfield, D. A. Southgate, Food composition data: production, management, and use, Food & Agriculture Org., 2003.

[3] M. Khalis, et al., Update of the moroccan food composition tables: Towards a more reliable tool for nutrition research, Journal of Food Composition and Analysis 87 (2020) 103397.

[4] J. Azanzi, et al., A large scale corpus of food composition tables, Semantic Web Challenge on Tabular Data to Knowledge Graph Matching (SemTab), CEUR-WS. org (2022).

[5] D. Dooley, et al., Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, npj Science of Food 2 (2018) 23–. doi:10.1038/s41538-018-0032-6.

[6] S. Auer, et al., Improving access to scientific literature with knowledge graphs, BIBLIOTHEK – Forschung und Praxis (2020). doi:http://dx.doi.org/10.18452/22049.

[7] A. Jiomekong, U. Melie, TSOTSATable dataset: a dataset of food and its composition, 2023. doi:10.5281/zenodo.8169063.