A Large Scale Corpus of Food Composition Tables

Azanzi Jiomekong[1,*], Cosmas Etoga[1], Brice Foko[1], Vadel Tsague[1], Martins Folefac[2], Sorel Kana[2], Mouhamadou Mansour Sow[3] and Gaoussou Camara[4]

[1]*Department of Computer Science, University of Yaounde I, Yaounde, Cameroon*
[2]*neuralearn.ai, Cameroon*
[3]*Pôle Science et Technologie du Numérique, Université Virtuelle du Sénégal, Dakar, Sénégal*
[4]*Unité de Formation et de Recherche en Sciences Appliquées et des TIC, Université Alioune Diop de Bambey, Bambey, Sénégal*

Abstract
In this paper, we introduce *TSOTSACorpus*, a large scale corpus of Food Composition Tables composed of more than 16,000 tables collected from scientific and Zenodo repositories. Our continuing maintenance and curation aims at growing this corpus in order to furnish good quality, up-to-date and cultural heritage of all foods information in the world. Compared to related datasets (INFOODS, LanguaL), we found that this corpus contains more information. In addition, it can be processed by humans and machines.

Keywords
Food Information Engineering, Food Composition Database, Food Composition Table, Tabular data,

1. Introduction

In recent years, many Food Composition Tables (FCT) [1] have been published in several formats (PDF, CSV, XSLX). However, these data are scattered on the Internet, making their exploitation difficult because one has to search, get data and extract information from them. On the other hand, many FCT whether it be at the country, regional or world wide level suffers from many problems: (1) Static databases sometimes in PDF or in XLSX, CSV, ODT formats; (2) Outdated data - the comparison of several FCT [2] showed that FCT should be always update because eating habit change over time; (3) Not harmonized data.

In this paper, we propose to extract, unify and link all Food Composition Tables published worldwide and accessible either in the form of scientific publication or in a free and/or open source license in a strong centralized corpus of FCT. One way to achieve this is by making each dataset accessible in a machine-readable format, which can be realized by putting these tables in CSV format and enriching them with metadata and data on their provenance. To this end, knowledge is automatically extracted from scientific literature and Zenodo repositories, curated

SemTab@ISWC 2022, October 23–27, 2022, Hangzhou, China (Virtual)
fidel.jiomekong@facsciences-uy1.cm (A. Jiomekong); etogacosmas@gmail.com (C. Etoga); fokobrice3@gmail.com (B. Foko); vadel.tsague@gmail.com (V. Tsague); martinsderick99@gmail.com (M. Folefac); jsorelkana@gmail.com (S. Kana); mouhamadoum.sow@uvs.edu.sn (M. M. Sow); gaoussou.camara@uadb.edu.sn (G. Camara)
https://sites.google.com/facsciences-uy1.cm/azanzijiomekong (A. Jiomekong)
0000-0002-0877-7063 (A. Jiomekong)

and annotated using biomedical ontologies. The work we present in this paper is an ongoing work and the next Section will present the current version of *TSOTSACorpus*.

2. *TSOTSACorpus*: a large scale corpus of FCT

Globally, *TSOTSACorpus* is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. The development version is available for download on Google Drive[1] and will be published on Zenodo as soon as the curation and annotation process is finished. The source code we are using for the extraction of tables from PDFs documents is available on GitHub[2] and Google Collaboratory[3]. A video showing how we automatically extract tables from PDFs is also available[4]. Once the tables are extracted from scientific papers, we have also considered the extraction of datasets from zenodo.org - the source is available on GitHub[5].

TSOTSACorpus construction is an extensive work of semi-automatic collection, extraction, curation and annotation of food data. Currently, more than 5,000 PDF documents acquired from scientific repositories are processed and more than 11,000 tables extracted from them. To this end, we used Neural Networks (NN) algorithms and we followed the Table detection, Text detection, Text recognition steps. Concerning the implementation, we rely on PaddleOCR which were trained with the Paddle framework in the Python programming language. On the other hand, Zenodo API[6] were used to automatically extract FCT datasets - more than 5,000 tables are currently extracted.

The current version of the corpus is composed of more than 16,000 tables of food, describing more than 60,000 foods, 200 food groups, and 800 food components. It covers the food consumed in more than 123 countries from 1987 to 2022. At this stage of this work, the extraction of additional tables, the curation and annotation process are in progress. The curation consists of linking each tabular data to the knowledge source from which it was built, identify and delete duplicate knowledge sources, arrange data in the CSV files so as to be exactly like the ones in PDF. The annotation process is being done by using biomedical ontologies (identified using ontobee.org - FoodOn, SNOMED CT and NCIT are currently used). We are also planning to consider the annotation with Wikidata and DBpedia knowledge Graphs. We expect to produce the first version, curated and annotated, composed of more than 20,000 tables during the first quarter of 2023 so that it can be used during the future editions of the SemTab challenge[7].

Acknowledgment

We are grateful to SemTab organizers for having given us the opportunity to share this work with the community. We are also grateful to Vinsight and neuralearn.ai for the training support.

[1] https://drive.google.com/drive/u/1/folders/1U2dEye_f02MhHOkmowuh2UyAKX60Ix39

[2] https://github.com/Neuralearn/pdf-to-excel

[3] https://colab.research.google.com/drive/1gOPBCVO9VtKcoIewXyr_6nNoxo1Bkqbz

[4] www.youtube.com/watch?v=HZh31OGiQRQ

[5] https://github.com/iconoyuri/zenodo-file-downloader

[6] https://zenodo.org/api/records/

[7] https://www.cs.ox.ac.uk/isg/challenges/sem-tab/

References

[1] M. Khalis, V. Garcia-Larsen, H. Charaka, M. M. S. Deoula, K. El Kinany, A. Benslimane, B. Charbotel, A. S. Soliman, I. Huybrechts, G. A. Soliman, 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.

[2] A. Jiomekong, Comparison of food composition tables/databases, 2022. URL: https://orkg.org/comparison/R206121/.