Matching Food Items with FoodOn Identifiers in a Knowledge Base

Katherine Thornton[1], Kenneth Seals-Nutt[2] and Mika Matsuzaki[3]

[1] *WikiFCD Collaborative, Olympia, WA, USA*
[2] *WikiFCD Collaborative, New York, New York, USA*
[3] *Johns Hopkins Bloomberg School of Public Health, 615 N Wolfe St, Baltimore, MD 21205, United States*

Abstract

We maintain a knowledge base of food items and food composition data, WikiFCD. We source food composition data from published sources such as food composition tables (FCTs) and scientific papers. As we ingest data sets, we need to match food items to corresponding identifiers from the Food Ontology (FoodOn). We describe the process of matching food items to FoodOn identifiers in a knowledge base of food composition data. We describe a software tool called LexMapr that can be used to bulk match food items to FoodOn identifiers. We have found that the LexMapr tool reduces the time needed to match food items to their corresponding FoodOn identifiers.

Keywords

food composition, nutri-informatics, FoodOn, knowledge base

1. Introduction

When people search for food composition data they are often looking for data available on the web. One of the largest cost-free online databases of food composition data is Food Data Central from the USDA [1]. Food Data Central contains many information about packaged food items that are sold in grocery stores. It also contains many records for a variety of fruits and vegetables typically consumed in the United States. Despite the breadth of Food Data Central, we lack food composition data for the thousands of cultivars of fruits and vegetables, and for food items that are consumed in other parts of the world. Having web access to food composition data about many plant and vegetable foods that are part of the human diet is crucial for being able to represent a diversity of dietary intakes across communities.

Research into human diets often involves asking people what they have eaten, known as food intake data [2]. When processing food intake recall data, researchers are limited to recording food items for which they have food composition data. If food composition data for a food item is not found, a substitution is made [3]. Gaps in coverage of food items reduce the quality of studies that involve food intake procedures. Increasing the number of food items for which we

The Eighth Joint Ontology Workshops (JOWO'22), August 15-19, 2022, Jönköping University, Sweden

*Corresponding author.
†These authors contributed equally.
katherine.thornton@yale.edu (K. Thornton); kenneth@seals-nutt.com (K. Seals-Nutt); mmatsuz2@jhu.edu (M. Matsuzaki)
0000-0002-4499-0451 (K. Thornton); 0000-0002-5926-9245 (K. Seals-Nutt); 0000-0002-7020-3757 (M. Matsuzaki)

have food composition data available in a public knowledge base will improve research data related to food intake.

Food composition data are available for many plant foods in the food composition tables (FCTs) published at the national level and scientific articles that describe one or more food items in detail [4]. Many of these tables are available online as PDF or CSV files. This means that researchers would need to read through the full table to discover if the food item they are interested in is included and then convert them to a different format in order to import them into their own systems. Researchers estimate that ten thousand species of plants are edible [5]. Even the largest food composition databases on the web lack food composition data for many of these foods.

Human diets that consist of plant-based foods are more environmentally sustainable than human diets that include animal-based foods [6]. We aim to provide food composition data for additional plant-based foods that we source from published literature. This goal complements work to move toward more sustainable human diets [7]. We hope to support researchers, software developers, and communities who need accurate food composition data for more of the plants that humans consume. In order to accommodate a flexible, extensible data model, provide search and SPARQL query access on the web, support multilingual data and open contributions to the community we decided to create a knowledge base for this data.

2. The WikiFCD Knowledge Base

WikiFCD is a knowledge base of structured data about food items and their nutritional composition [8]. We accept data from published sources about food items. Our aim is to improve accessibility of these data by offering an online, public, multilingual resource that consolidates food composition data[1].

The WikiFCD knowledge base is open to contributions from anyone who would like to add food composition data. We support a peer-production model for contribution[8]. The data are free for anyone to reuse, meeting the need for cost-free, web-based access to food composition data for foods.

We use Wikibase, an extension of MediaWiki, as the platform for WikiFCD. Wikibase is the software that enables Wikidata, a public knowledge base that supports projects of the Wikimedia Foundation[9]. Wikibase includes a SPARQL endpoint that provides query access to the data [10].

We designed our data model to accommodate food composition data from a wide variety of sources. We have mapped some of the WikiFCD properties to properties in the Wikidata knowledge base, and our public SPARQL endpoint supports federated SPARQL queries that combine data from Wikidata with data from our own knowledge base. This allows us to benefit from the nutri-informatics approaches that mapping biomedical resources to Wikidata makes possible [11]. Wikidata contains thousands of properties dedicated to mapping Wikidata items to related information in external databases[12]. Through mapping our Wikibase to Wikidata, we gain the ability to combine our data with data from the Wikidata knowledge base as well as identifiers for millions of concepts in thousands of other information systems, databases,

[1]The WikiFCD knowledge base is available at: https://wikifcd.wiki.opencura.com/wiki/Main_Page

Table 4. Carotenoids

Food code	Food Name	No. of Regions	Lutein	Zeaxanthin	Lycopene	β-Cryptoxanthin	γ-Carotene	α-Carotene	β-Carotene	Total Carotenoids
			← μg →							
			LUTN	ZEA	LYCPN	CRYPXB	CARTG	CARTA	CARTB	CARTOID
E020	**Fig** *(Ficus carica)*	6	19.0±2.27	2.01±0.64					2.40±0.59	454±51.3
E021	**Goosberry** *(Emblica officinalis)*	5	38.7±9.78	2.86±1.61					1.58±0.09	62.01±6.45
E022	**Grapes, seeded, round, black** *(Vitis vinifera)*	4	27.8±8.85	3.59±2.20					29.36±8.37	213±29.4
E023	**Grapes, seeded, round, green** *(Vitis vinifera)*	5	31.3±8.77	4.06±1.60					30.77±11.10	208±28.8
E024	**Grapes, seeded, round, red** *(Vitis vinifera)*	5	17.2±3.04	7.58±2.28					19.94±1.29	314±99.9
E025	**Grapes, seedless, oval, black** *(Vitis vinifera)*	5	19.4±3.07	8.14±1.73					19.73±1.76	321±50.6
E026	**Grapes, seedless, round, green** *(Vitis vinifera)*	5	25.8±7.90	6.82±3.77					25.46±6.66	216±31.3
E027	**Grapes, seedless, round, black** *(Vitis vinifera)*	5	16.1±2.90	7.95±1.73					20.58±2.36	232±27.9
E028	**Guava, white flesh** *(Psidium guajava)*	5	19.7±4.59	5.92±3.06					298±32.8	996±108
E029	**Guava, pink flesh** *(Psidium guajava)*	5	90.3±11.5	8.94±2.21	2843±340				267±88.2	4078±477
E030	**Jack fruit, ripe** *(Artocarpus heterophyllus)*	5	19.3±6.91	2.42±1.27					23.53±1.24	59.61±4.90
E031	**Jambu fruit, ripe** *(Syzygium cumini)*	2	18.49	5.68					1.55	66.50
E032	**Karonda fruit** *(Carissa carandas)*	1	6.12	1.14					15.64	55.89
E033	**Lemon, juice** *(Citrus limon)*	6	9.68±2.22	1.72±0.50					2.62±0.34	85.99±3.36
E034	**Lime, sweet,pulp** *(Citrus limetta)*	6	13.2±3.54	1.89±0.44					2.54±0.30	85.13±10.71
E035	**Litchi** *(Nephelium litchi)*	4	27.33±9.81	1.61±0.62					1.47±0.21	129±15.1
E036	**Mango, ripe, banganapalli** *(Mangifera indica)*	6	3.19±1.51	1.62±0.22		2.72±1.31			1168±141	1424±308
E037	**Mango, ripe, gulabkhas** *(Mangifera indica)*	2	4.18	1.64		4.58			666	1350
E038	**Mango, ripe, himsagar** *(Mangifera indica)*	3	6.26±2.09	1.19±0.14		5.60±3.84			1181±136	1423±201
E039	**Mango, ripe, kesar** *(Mangifera indica)*	4	3.71±2.21	2.33±0.17		6.96±3.31			1264±172	1438±314
E040	**Mango, ripe, neelam** *(Mangifera indica)*	2	2.95	1.67		2.51			1291	1376

102

Figure 1: Example of a section of the India FCT

and websites. This is relevant for people interested in food and human health because projects such as the Gene Wiki initiative have added human gene information and data from the Disease Ontology to Wikidata [13]. By writing federated queries we can ask questions about food items in combination with questions about aspects of human biological processes.

3. Adding Data to WikiFCD

People contributing to WikiFCD often start from food composition tables (FCTs) that have been published, many of which include data about hundreds of food items. Figure 1 is an example of a page from the India FCT. Research teams that publish FCTs typically include the name of the food items in English, common names of the food items in one or more languages spoken in the region covered by the dataset, and a selection of macro- and micro-nutrient values for the food items. Occasionally the FCT authors include the Latin binomial name of the plant species from which the food item is derived. We then review all of the columns in the dataset and prepare a mapping from the column names to relevant properties in WikiFCD.

The WikiFCD team recently completed an integration with the Food Ontology (FoodOn) [14]. For new data to be added to WikiFCD we now need a process through which we can match food items to their FoodOn identifiers. Some FCTs include hundreds of food items, so we require an effective way to undertake food items matching that will scale to the size of the incoming data sets.

4. FoodOn

The Food Ontology (FoodOn) is an ontology of food items, relationships to species, food processes, and cross references to other ontologies [15]. FoodOn is built upon LanguaL, and

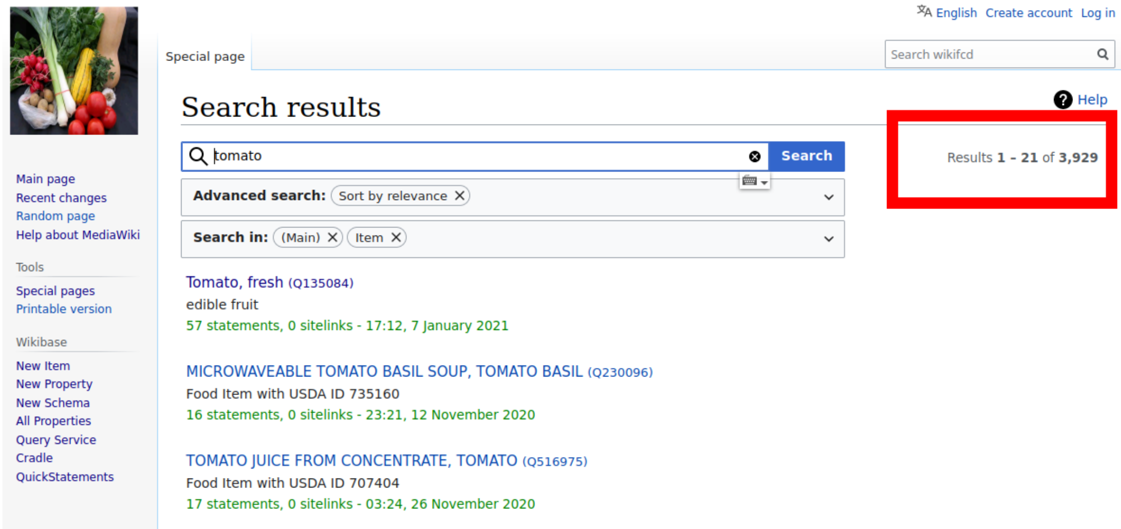

Figure 2: Screenshot of search for 'tomato' in WikiFCD showing 3,929 results.

incorporates the food items from LanguaL [16]. The FoodOn curating team actively extends the ontology and continues to add new terms as needed. We find FoodOn to be a valuable resource that is used across multiple domains from agriculture, to food science, health sciences among others. By adding FoodOn identifiers to food items in our knowledge base we expect that our data will be easier for others to reuse for their own needs.

FoodOn is a member of the Open Biomedical Ontologies (OBO). This umbrella organization ensures the interoperability of member ontologies and provides tools, such as the Dashboard and automated tests, which help improve the quality of each individual ontology [17].

Adding a property for the FoodOn identifier to food items is a helpful way for us to group items. Using FoodOn as a pivot ontology is a useful way to bring together food items from disparate published sources [18]. Because we integrate data from many sources, we have many composition profiles for the same food item. For example, in Figure 2 we see a screenshot of a search for 'tomato' in WikiFCD with a red rectangle around the results of the search showing that 3,929 items in the knowledge base contain the string 'tomato'.

5. Using LexMapr to Identify Food Items

We used the LexMapr tool to automatically match food items with corresponding FoodOn identifiers[2]. This use builds on the work of other groups who have successfully used LexMapr [19, 20, 21]. LexMapr is a Python program that reads a list of food item labels and then returns the following columns: Sample_Desc, Processed_Sample, Processed_Sample (With Scientific Name), Matched_Components, Match_Status(Macro Level), and Third Party Classification, as seen in Figure 3.

[2]https://lexmapr.cidgoh.ca/user-guide/

Sample_Id	Sample_Desc	Processed_Sample	Processed_Sample (With Scientific Name)	Matched_Components	Match_Status(Macro Level)	Third Party Classification
1	Acha, black, whole grain, Raw	acha black whole grain raw	acha black whole grain raw	['food (raw):FOODON_03311126', 'whole grain:FOODON_00003950']	Component Match	['grains', 'seeds']
2	Acha, white, whole grain, Raw	acha white whole grain raw	acha white whole grain raw	['food (raw):FOODON_03311126', 'white:PATO_0000323', 'whole grain:FOODON_00003950']	Component Match	['grains', 'seeds']
3	Acha, white, whole grain, Boiled	acha white whole grain boiled	acha white whole grain boiled	['food (boiled):FOODON_00002688', 'white:PATO_0000323', 'whole grain:FOODON_00003950']	Component Match	['grains', 'seeds']
4	Maize, white, whole kernel, dried, raw	maize white whole kernel dried raw	maize {zea mays} white whole kernel dried raw	['food (dried):FOODON_03307539', 'food (raw):FOODON_03311126', 'maize kernel:FOODON_00003427', 'seed, skin present, germ present:FOODON_03420133', 'white:PATO_0000323']	Component Match	['fruits']

Figure 3: Layout of result data produced by LexMapr

The first column of the results from LexMapr, Sample_Desc, is the original string supplied in the uploaded CSV. The second column, Processed_Sample, is the string that LexMapr uses for searching. The third column, Processed_Sample (With Scientific Name), is the string from the second column with a Latin binomial scientific name for the species from which the food item is derived. The fourth column, Matched_Components, contains the label and identifier for any FoodOn terms that match components of the string from column two. The fifth column, Match_Status(Macro Level), contains an indication of the type of match. The sixth column, Third Party Classification, provides a food type for the food item.

LexMapr emits a column of data that indicates if each row of data is a a 'Component Match' a 'Full Term Match' or a 'No Match'. We have found that if LexMapr finds a 'Full Term Match' for a food item, that match is likely to be accurate. When LexMapr finds a 'Component Match' we need to review the match to see if it is a correct match.

For example, in the Nigeria Food Composition Table one of the food items is 'acha'. LexMapr proposed a component match '['food (raw):FOODON_03311126', 'whole grain:FOODON_00003950']'. We reviewed this result with members of the FoodOn curation team and found that 'acha' needs to be added to the list of synonyms for 'fonio'. After this addition, future users of LexMapr will find that 'acha' will match with 'black fonio grain' with FOODON_03540018 as the identifier.

The column 'Processed_Sample (With Scientific Name)' is produces useful data related to the scientific name of the organism from which the food item is produced. Due to the fact that some of the FCTs we integrate into our knowledge base do not provide this information, we can leverage results from FoodOn to enrich our data.

After running LexMapr, we conduct a manual review of all rows that have 'Component Match' in column four. The component matches require human review because sometimes the correct FoodOn identifier is provided, sometimes multiple possible FoodOn identifiers are provided, and sometimes multiple potential matches are provided but are not relevant.

LexMapr provides automated matching of food items to FoodOn identifiers that will save us time when compared with manually searching FoodOn to find identifiers. We anticipate that LexMapr will gain accuracy as we add synonyms to FoodOn when we find evidence for them in our inventory of FCTs from all over the world[3].

[3]The LexMapr output for an entire FCT in the WikiFCT knowledge base is available at https://wikifcd.wiki.open-

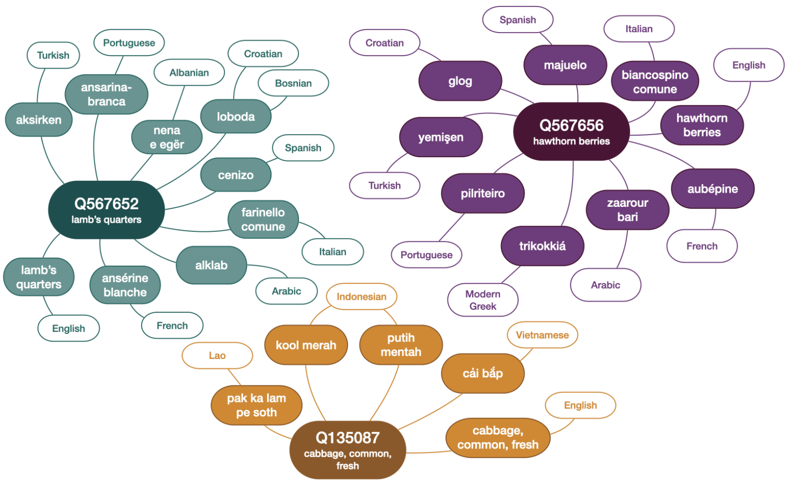

Figure 4: Three food items from our knowledge base that contain multilingual data.

6. Proposing Terms for FoodOn

Due to the fact that we integrate food composition tables from all over the world, we have a diverse pool of food items in WikiFCD. Some of these food items are not yet represented in FoodOn, and we would like to offer this information to the FoodOn curation team for consideration. Many of the FCTs we have integrated into WikiFCD provide English language labels for food items as well as labels in one or more languages from the region. When we have labels in languages other than English, such as the examples seen in Figure 4, we can also contribute these to FoodOn to help extend multilingual coverage for FoodOn terms. We found the food names seen in 4 in published FCTs, and each of these is referenced with the source where we found it.

Expanding the number of labels for food items in FoodOn will allow more people to search for food items in their own languages.

7. Conclusion

Using LexMapr to match food items with FoodOn identifiers allows the WikiFCD team to save time that would be spend manually matching. Due to the international sources of data for WikiFCD we are also generating a lot of potential synonyms for food items already in FoodOn. Over time, once these synonyms are reviewed and accepted by the FoodOn curation team, they will likely improve the performance of LexMapr. For food items that are new for FoodOn (and unmatched by LexMapr) we now have the opportunity to present these items to the FoodOn curation team for consideration.

---

cura.com/wiki/Property_talk:P317

We wanted to integrate with FoodOn to ensure that our knowledge base would be well-positioned for reuse by other organizations that make use of FoodOn. An example of a relevant community is that of the Ontology for Nutritional Studies (ONS) [22]. The ONS is a formal ontology framework for the description of nutritional studies. Because FoodOn and the ONS are already integrated, the use of FoodOn identifiers in WikiFCD makes it easier to also integrate food composition data from our knowledge base with ONS data.

People who are considering integrating FoodOn with their data will need to match foods to FoodOn identifiers. LexMapr will help ease integration with FoodOn. Once you have mapped your data to FoodOn you also gain connections with all of the OBO ontologies that FoodOn reuses.

Creating WikiFCD allows us to provide web-based access to food composition data for many plant-based foods. This knowledge base makes the FCTs we integrate much easier to work with because we provide machine-actionable data that can be integrated with other applications. For groups that work with FoodOn identifiers, by using WikiFCD, it is now easy to integrate food composition data for many food items.

Acknowledgments

We thank the Joint Food Ontology Working Group for productive discussions about FoodOn. We appreciate the assistance that Damion Dooley provided relating to the use of LexMapr. We thank the Hsiao Lab of Simon Fraser University for funding the creation of LexMapr and hosting the web application. We thank Adam Shorland for creating and maintaining WBstack. We thank the Wikidata community for continuing to improve the Wikidata knowledge base.

References

[1] B. Larrick, A. Kretser, K. McKillop, Update on “a partnership for public health: Usda global branded food products database”, Journal of Food Composition and Analysis 105 (2022) 104250. URL: https://www.sciencedirect.com/science/article/pii/S0889157521004506. doi:https://doi.org/10.1016/j.jfca.2021.104250.

[2] K. J. Morgan, S. Johnson, R. L. Rizek, R. Reese, G. L. Stampley, Collection of food intake data: an evaluation of methods., Journal of the American Dietetic Association 87 (1987) 888–896.

[3] C. Byker Shanks, B. Izumi, C. A. Parks, A. L. Yaroch, Measurement of fruit and vegetable intake incorporating a diversity, equity, and inclusion lens. comment on di noia, j.; gellermann, w. use of the spectroscopy-based veggie meterreg; to objectively assess fruit and vegetable intake in low-income adults. nutrients 2021, 13, 2270, Nutrients 14 (2022). URL: https://www.mdpi.com/2072-6643/14/4/809. doi:10.3390/nu14040809.

[4] J. R. Lupien, The fao/unu food composition initiative, Food chemistry 57 (1996) 171–173.

[5] S. Baldermann, L. Blagojević, K. Frede, R. Klopsch, S. Neugart, A. Neumann, B. Ngwene, J. Norkeweit, D. Schröter, A. Schröter, F. J. Schweigert, M. Wiesner, M. Schreiner, Are neglected plants the food for the future?, Critical Reviews in Plant Sciences 35 (2016)

106–119. URL: https://doi.org/10.1080/07352689.2016.1201399. doi:10.1080/07352689.2016.1201399. arXiv:https://doi.org/10.1080/07352689.2016.1201399.

[6] J. Gibbs, F. P. Cappuccio, Plant-based dietary patterns for human and planetary health, Nutrients 14 (2022). URL: https://www.mdpi.com/2072-6643/14/8/1614. doi:10.3390/nu14081614.

[7] N. M. Holden, E. P. White, M. Lange, T. L. Oldfield, et al., Review of the sustainability of food systems and transition using the internet of food, npj Science of Food 2 (2018) 1–7.

[8] K. Thornton, K. Seals-Nutt, M. Matsuzaki, Introducing wikifcd: Many food composition tables in a single knowledge base, in: CEUR Workshop Proceedings, volume 2969, CEUR-WS, 2021.

[9] D. Diefenbach, M. D. Wilde, S. Alipio, Wikibase as an infrastructure for knowledge graphs: The eu knowledge graph, in: International Semantic Web Conference, Springer, 2021, pp. 631–647.

[10] L. Zhou, C. Shimizu, P. Hitzler, A. M. Sheill, S. G. Estrecha, C. Foley, D. Tarr, D. Rehberger, The enslaved dataset: A real-world complex ontology alignment benchmark using wikibase, in: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, 2020, pp. 3197–3204.

[11] L. Chan, N. Vasilevsky, A. Thessen, J. McMurry, M. Haendel, The landscape of nutri-informatics: a review of current resources and challenges for integrative nutrition research, Database 2021 (2021). URL: https://doi.org/10.1093/database/baab003. doi:10.1093/database/baab003. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab003/36110502/baab003.pdf, baab003.

[12] J. Neubert, Wikidata as a linking hub for knowledge organization systems? integrating an authority mapping into wikidata and learning lessons for KOS mappings, in: Proceedings of the 17th European Networked Knowledge Organization Systems Workshop co-located with the 21st International Conference on Theory and Practice of Digital Libraries 2017 (TPDL 2017), Thessaloniki, Greece, September 21st, 2017., 2017, pp. 14–25. URL: http://ceur-ws.org/Vol-1937/paper2.pdf.

[13] A. Waagmeester, G. Stupp, S. Burgstaller-Muehlbacher, B. M. Good, M. Griffith, O. L. Griffith, K. Hanspers, H. Hermjakob, T. S. Hudson, K. Hybiske, S. M. Keating, M. Manske, M. Mayers, D. Mietchen, E. Mitraka, A. R. Pico, T. Putman, A. Timothy, N. Queralt-Rosinach, L. M. Schriml, T. Shafee, D. Slenter, R. Stephan, K. Thornton, G. Tsueng, R. Tu, S. Ul-Hasan, E. Willighagen, C. Wu, A. I. Su, Wikidata as a knowledge graph for the life sciences, Elife 9 (2020) e52614. URL: https://doi.org/10.7554/ELIFE.52614.

[14] K. Thornton, K. Seals-Nutt, M. Matsuzaki, D. Damion, Reuse of the foodon ontology in a knowledge base of food composition data, under review (2022).

[15] D. M. Dooley, E. J. Griffiths, G. S. Gosal, P. L. Buttigieg, R. Hoehndorf, M. C. Lange, L. M. Schriml, F. S. Brinkman, W. W. Hsiao, Foodon: a harmonized food ontology to increase global food traceability, quality control and data integration, npj Science of Food 2 (2018) 1–10.

[16] J. D. Ireland, A. Møller, Langual food description: a learning process, European journal of clinical nutrition 64 (2010) S44–S48.

[17] R. Jackson, N. Matentzoglu, J. A. Overton, R. Vita, J. P. Balhoff, P. L. Buttigieg, S. Carbon, M. Courtot, A. D. Diehl, D. M. Dooley, W. D. Duncan, N. L. Harris, M. A.

Haendel, S. E. Lewis, D. A. Natale, D. Osumi-Sutherland, A. Ruttenberg, L. M. Schriml, B. Smith, C. J. Stoeckert Jr., N. A. Vasilevsky, R. L. Walls, J. Zheng, C. J. Mungall, B. Peters, OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies, Database 2021 (2021). URL: https://doi.org/10.1093/database/baab069. doi:10.1093/database/baab069. arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baab069/40854912/baab069.pdf, baab069.

[18] P. Buche, J. Cufi, S. Dervaux, J. Dibie, L. Ibanescu, A. Oudot, M. Weber, How to manage incompleteness of nutritional food sources?: A solution using foodon as pivot ontology, International Journal of Agricultural and Environmental Information Systems (IJAEIS) 12 (2021) 1–26.

[19] M. Balkey, M. Batz, G. Gopinath, G. Gosal, E. Griffiths, H. Tate, R. Timme, (v) standardizing the isolation source metadata for the genomic epidemiology of foodborne pathogens using lexmapr, IAFP 2021 (2021).

[20] D. Dooley, L. Andres-Hernandez, G. Bordea, L. Carmody, D. Cavalieri, L. Chan, P. Castellano-Escuder, C. Lachat, F. Mougin, F. Vitali, et al., Obo foundry food ontology interconnectivity, in: CEUR Workshop Proceedings, volume 2969, 2021.

[21] J. Pires, J. S. Huisman, S. Bonhoeffer, T. P. Van Boeckel, Increase in antimicrobial resistance in escherichia coli in food animals between 1980 and 2018 assessed using genomes from public databases, Journal of Antimicrobial Chemotherapy 77 (2022) 646–655.

[22] F. Vitali, R. Lombardo, D. Rivero, F. Mattivi, P. Franceschi, A. Bordoni, A. Trimigno, F. Capozzi, G. Felici, F. Taglino, et al., Ons: an ontology for a standardized description of interventions and observational studies in nutrition, Genes & nutrition 13 (2018) 1–9.