# Multilingual Labels for FoodOn

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**

Data sets involving food frequently are of interest across multiple cultural contexts. Many ontologies and vocabularies related to food are monolingual. Enriching ontologies and vocabularies with multilingual data is useful for extending access to a resource to wider audiences. The WikiFCD knowledge base of FAIR food composition data includes mappings to Wikidata and mappings to FoodOn, a widely-used ontology for food. In this paper we describe a sample of six food items from WikiFCD as the basis for an exploration of strategies for sourcing multilingual labels from projects of the Wikimedia Foundation. We present five subgraphs of data related to food items sourced from Wikidata, Wikipedias and WikiFCD. We describe the advantages and disadvantages of sourcing labels from each subgraph. Each subgraph can be quickly retrieved from the Wikidata Query Service SPARQL endpoint or the SPARQL endpoint of WikiFCD. These strategies could be adapted to other domains seeking to enrich their data with multilingual labels.

**Keywords**

Food Composition, Nutri-informatics, Wikibase, Wikidata

## Introduction

WikiFCD is a knowledge base of food composition data. After importing more than three hundred thousand food items and their associated food composition measurements from the United States Department of Agriculture's Food Data Central (FDC), we began to add data from additional countries. We have added more than three thousand food items to the knowledge base along with measurements of nutritional components of those foods from published sources [1]. The purpose of this knowledge base is to provide web access to published data about food composition related to foods that are not found in FDC. For example, researchers have documented gaps in coverage of plant foods and foods from international cuisines in FDC [2, 3]. WikiFCD contains food composition information for many plant foods that are not included in FDC. Plant foods contain phytonutrients that support human health [4, 5, 6]. The nutrients found in these minimally-processed plant foods are within the range effectively utilized by

*IFOW 2023: Integrated Food Ontology Workshop, 9th Joint Ontology Workshops (JOWO 2023), co-located with FOIS 2023, 19-20 July, 2023, Sherbrooke, Québec, Canada*

[*] 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)

human metabolism, meaning we can digest and make use of these nutrients more easily [7]. Providing the food composition data for these foods allows people who track their food intake to get a more accurate estimate of their nutritional intake.

We created this knowledge base using Wikibase, and it is now part of the ecosystem of Wikibases related to Wikidata. The ecosystem of Wikibases includes all instances of Wikibase that can be federated with the Wikidata Query Service SPARQL endpoint [8]. Wikibase is software infrastructure for creating knowledge bases [9]. An advantage to creating a knowledge base is that it is possible to connect the data in the knowledge base with other data on the web. We created mappings to Wikidata for some classes of items and for some properties in WikiFCD. These mappings allow us to combine our data with subsets of data from Wikidata, increasing the complexity and types of questions we can ask of our data. We have integrated the FoodOn ontology into WikiFCD so that other projects that make use of FoodOn can also make use of data from WikiFCD [10].

The FoodOn mappings in WikiFCD allow us to connect these food items to their corresponding food items in Wikidata. This connection, stored in WikiFCD, opens up several pathways to source additional information about these food items. We can think of each of the potential graphs of food-related data in Wikidata as Wikidata subsets [11]. In this paper we describe five subsets of data from WikiFCD and Wikidata that can be used to source labels for these food items in languages other than English. Multilingual labels can then be used for search, in application development, or for other purposes. For projects that make use of FoodOn, this data set of food items could be an option for sourcing multilingual labels to expand the audiences for other projects.

Data in Wikidata and WikiFCD are FAIR data. The FORCE 11 community published the FAIR data principles in 2014 to promote data publishing practices that would support open science and open access [12]. FAIR is an acronym for Findable, Accessible, Interoperable and Reusable, four qualities of published data that promote data sharing. By populating WikiFCD with data from published sources, we made food composition data findable and accessible on the web via the WikiFCD SPARQL endpoint[1]. Every food item in Wikidata and in WikiFCD has a QID, which is a Unique Resource Identifier (URI). This means that every piece of data in our dataset has a machine-actionable URI. The mapping statements we added to our food items that connect them to Wikidata serve as our bridges with the web of linked open data, making our data interoperable with many additional datasets. Data in Wikidata and in WikiFCD are available under the terms of the Creative Commons Zero license[2]. Our selection of CC0 as the license means that anyone can freely reuse the data. These aspects of publishing data in the Wikidata and WikiFCD knowledge bases fulfills the most complete degree of FAIRness, level F, "FAIR data, Open Access, Functionally Linked", as described in [13]. We welcome others who are interested in reusing data from WikiFCD.

[1] https://wikifcd.wikibase.cloud/query/

[2] https://creativecommons.org/share-your-work/public-domain/cc0/

## 1. The WikiFCD Knowledge Base

We implemented the WikiFCD knowledge base using Wikibase, an extension of the MediaWiki software that was created for the Wikidata knowledge base[3]. We include several subsets of data from Wikidata in WikiFCD in order to establish useful mappings between the WikiFCD system and Wikidata itself. These mappings allow us to ask questions of the WikiFCD dataset in combination with data from Wikidata, increasing the breadth and complexity of the queries.

Each item in WikiFCD has a unique identifier, or Qid. The Qid for this food item is Q135853. Most of the statements on the food items in WikiFCD provide information about the amount of a nutrient in a 100 gram sample of the food item. Each statement also has room for a reference. References record the source for the statement. In WikiFCD, most reference are for a food composition table.

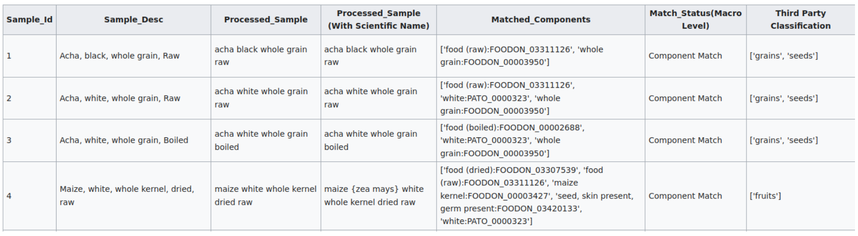
Researchers have already created knowledge graphs related to food. Fore example, FoodKG is a knowledge graph of food data [14]. WikiFCD centers food composition data for plant foods, specifically unprocessed or minimally processed plant foods. One difference between WikiFCD and FoodKG is that FoodKG does not include mappings to Wikidata. When Food KG was created there were not many food items in Wikidata. Another difference between Food KG and WikiFCD is that WikiFCD contains multilingual data.

Open Food Facts is a database of food composition data that is open to community data contribution and curation [15]. Open Food Facts makes use of the UPC codes on food packaging. While WikiFCD does contain data related to some packaged foods, we emphasize fruits, vegetables, grains and other foods that are often sold without packaging.

## 2. Identifying Food Items in Wikidata

We used several strategies to identify candidate queries that we would use to extract a subset of food items from Wikidata. Our aim was to find a subset of food items to match with FoodOn identifiers. We wrote a SPARQL query for food items in Wikidata with a corresponding article in English Wikipedia. We thought that the fact that Wikipedians had already written an article about the food might be an indicator of the popularity of the food. While this was a useful set of food items, we had to remove food brands from the query. We rejected the article query results as unsuitable and tried another approach. We wrote a SPARQL query for food items in Wikidata with a statement using Property ‘‘USDA NDB number” which is an identifier for a food item in the United States Department of Agriculture National Nutrient Database. Due to the fact that this identifier would only be found on food items in the Wikidata knowledge base, and would not include food brands, we determined that this would be a more suitable subset of food items for matching with food items from FoodOn. Our subgraph, Wikidata food items with a USDA NDB identifier, consists of one thousand three hundred and ninety-five food items. Due to the fact that some of these food items have multiple USDA NDB identifiers, the subgraph contains five hundred thirteen unique food items. This seemed like a suitable corpus of food items for which we could identify FoodOn term matches.

[3] https://wikiba.se/



| 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 1:** Layout of result data produced by LexMapr

## 3. Matching Food Items to FoodOn

When an organization wants to integrate FoodOn with their data, they need to map foods from their data to FoodOn identifiers. Researchers have found the LexMapr application to provide useful matches for food items and FoodOn identifiers [16, 17, 18]. LexMapr is a Django application that accepts CSV files as input and returns a file with information about matches and potential matches of submitted food labels with FoodOn identifiers. We used the LexMapr application to provide candidate matches between food items from our NDB number subset of Wikidata and FoodOn identifiers[4]. LexMapr is a Python-based Django application 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 1.

For some of the food items LexMapr generated full matches. For those items for which LexMapr generated component matches, we manually reviewed the results. Some of the component matches turned out to contain matches and some did not.

## 4. Adding Food Items Subset from Wikidata to WikiFCD

We wrote a bot script using WikidataIntegrator, a Python library, to create new food items in WikiFCD for the food items in the Wikidata subset. Numerous research groups use WikidataIntegrator to add data to Wikidata and to Wikibases [19]. We designed the bot script to add statements with links back to the corresponding food items in Wikidata. This set of mappings is useful for writing SPARQL queries to ask questions about the data in WikiFCD in combination with the data from Wikidata.

## 5. Leveraging Multilingual Content in Wikimedia Projects

FoodOn provides food names in English as well as alternative names in English. The FoodOn curation team would like to include multilingual content in order to make FoodOn more useful for people working in other linguistic contexts. Multilingual data is expensive to create and difficult to source. The purpose of this paper is to compare and contrast five strategies for extracting multilingual labels for food items from projects of the Wikimedia Foundation. The

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

mappings that we created between items in WikiFCD and Wikidata allow us to write federated SPARQL queries that combine data from WikiFCD with data from Wikidata. Wikidata is a multilingual knowledge base with support for more than three hundred human languages [20]. We identify five graphs of multilingual data from the Wikidata knowledge base that the FoodOn curation team may evaluate for suitability for integration into FoodOn. The five graphs are structurally distinct in that they are modeled differently in the knowledge base, and partially overlapping. In this way the subgraphs can be compared with one another to provide an agreement score for various foods.

In WikiFCD we curate multilingual labels for food items when we find them in food composition table sources. An advantage of this approach is that all multilingual labels have provenance information and include, at the statement level, references to the source from which we gathered the data. Using our subgraph of food items in WikiFCD for which we have created mappings to Wikidata, we then write SPARQL queries to extract multilingual labels for the food items from Wikidata. An advantage to writing these SPARQL queries is that we can run them again in the future and check if additional data has been added to Wikidata. With more than twelve thousand active editors each month contributing to Wikidata, we anticipate that the number of multilingual labels in Wikidata will increase over time.

### 5.1. Article Names per Language Version of Wikipedia

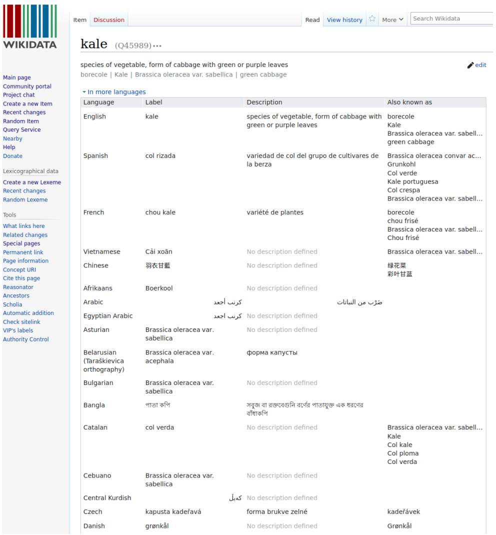
The Wikidata community has created items in Wikidata for each of the articles across the different language versions of Wikipedia [21]. There are currently more than three hundred active Wikipedias [22]. Wikidata contains information about mappings between items and the sitelinks to all of the corresponding articles across the different language versions of Wikipedia. If we look at a food item such as kale, the Wikidata item for which is Q45989, we can use a SPARQL query on the Wikidata Query Service to find links to all Wikipedias that have an article about kale. Currently sixty-two Wikipedias have an article about kale. Each of these Wikipedia articles has a title, and from these article titles, we can get a sense of what kale is called in these different languages.

### 5.2. Multilingual Labels from Wikidata Food Items

Another option for finding multilingual data is to consult the labels on Wikidata items. The Wikidata data model was designed with multilingual content in mind [21]. More than three hundred human languages are supported in Wikidata [20]. If we return to the example of kale, the Wikidata item for kale currently has labels in seventy-nine languages, a sample of which can be seen in Figure 2. This means that there are seventeen additional labels beyond the number of articles across the different Wikipedia language versions. We can compare these labels with those we found from the Wikipedias to check for consistency and accuracy.

### 5.3. Common Names from Taxon Items in Wikidata

Wikidata has a property “Taxon common name” (P1843) that can be used on taxon items to list common names for the organism. The common names listed using this property are another source of multilingual labels for food items. The property has a required qualifier which indicates



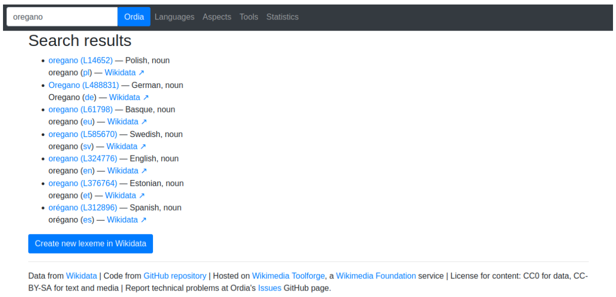
kale (Q45989)

species of vegetable, form of cabbage with green or purple leaves
borecole | Kale | Brassica oleracea var. sabellica | green cabbage

In more languages

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | kale | species of vegetable, form of cabbage with green or purple leaves | borecole<br>Kale<br>Brassica oleracea var. sabell...<br>green cabbage |
| Spanish | col rizada | variedad de col del grupo de cultivares de la berza | Brassica oleracea convar ac...<br>Grunkohl<br>Col verde<br>Kale portuguesa<br>Col crespa<br>Brassica oleracea var. sabell... |
| French | chou kale | variété de plantes | borecole<br>chou frisé<br>Brassica oleracea var. sabell...<br>Chou frisé |
| Vietnamese | Cải xoăn | No description defined | Brassica oleracea var. sabell... |
| Chinese | 羽衣甘藍 | No description defined | 绿花菜<br>彩叶甘蓝 |
| Afrikaans | Boerkool | No description defined | |
| Arabic | كرنب أجعد | ضَرْب من النباتات | |
| Egyptian Arabic | كرنب اجعد | No description defined | |
| Asturian | Brassica oleracea var. sabellica | No description defined | |
| Belarusian (Taraškievica orthography) | Brassica oleracea var. acephala | форма капусты | |
| Bulgarian | Brassica oleracea var. sabellica | No description defined | |
| Bangla | পাতা কপি | সবুজ বা রক্তবেগুনি বর্ণের পাতাযুক্ত এক ধরণের বাঁধাকপি | |
| Catalan | col verda | No description defined | Brassica oleracea var. sabell...<br>Kale<br>Col kale<br>Col ploma<br>Col verda |
| Cebuano | Brassica oleracea var. sabellica | No description defined | |
| Central Kurdish | کەیڵ | No description defined | |
| Czech | kapusta kadeřavá | forma brukve zelné | kadeřávek |
| Danish | grønkål | No description defined | Grønkål |

**Figure 2:** Wikidata item for 'kale' showing some of the labels in a variety of languages



Search results

- oregano (L14652) — Polish, noun
  oregano (pl) — Wikidata ↗
- Oregano (L488831) — German, noun
  Oregano (de) — Wikidata ↗
- oregano (L61798) — Basque, noun
  oregano (eu) — Wikidata ↗
- oregano (L585670) — Swedish, noun
  oregano (sv) — Wikidata ↗
- oregano (L324776) — English, noun
  oregano (en) — Wikidata ↗
- oregano (L376764) — Estonian, noun
  oregano (et) — Wikidata ↗
- orégano (L312896) — Spanish, noun
  orégano (es) — Wikidata ↗

Data from Wikidata | Code from GitHub repository | Hosted on Wikimedia Toolforge, a Wikimedia Foundation service | License for content: CC0 for data, CC-BY-SA for text and media | Report technical problems at Ordia's Issues GitHub page.

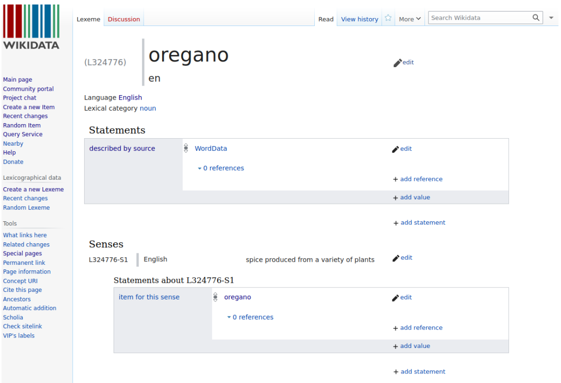
**Figure 3:** Using the Ordia application to search for 'oregano' in Wikidata's L namespace

that the language of the common name must also be provided in statements using this property. In this way we not only know the label, but also the language in which it is found.

## 5.4. Wikidata Lexemes

Wikidata introduced support for creating and editing lexemes in 2018 [23]. The Wikidata community creates lexemes, forms and senses in the L namespace [24]. Editors have already added more than half a million lexical entries [25]. The words described in the L namespace include words related to food. A fourth pathway for sourcing multilingual labels related to foods is to leverage Wikidata's lexeme data.

Editors use the property 'item for this sense' (P5137) to connect senses to items in Wikidata. In Figure 3 we see the result of searching for 'oregano' in the Ordia application [26]. In Figure 4 we see the lexeme 'oregano' (L324776 ) in English. Under the first sense, which has the identifier 'L324776-S1', we see a statement that uses 'item for this sense' (P5137) connecting the sense to



(L324776) oregano
en

Language English
Lexical category noun

Statements

described by source WordData
0 references

Senses

L324776-S1 English spice produced from a variety of plants

Statements about L324776-S1

item for this sense oregano
0 references

**Figure 4:** Lexeme L324776 with L324776-S1 and the statement that connects to Q92581152, the Wikidata item for 'oregano'

the Wikidata item for oregano.

## 6. Food Items and Multilingual Labels

We selected six food items as a demonstration data set: purslane, kale, apple, rice, oregano and chocolate. We wrote SPARQL queries to describe subsets of Wikidata or WikiFCD for each technique for sourcing multilingual labels described above. This sample allows us to compare and contrast the suitability of each subgraph of multilingual labels, and to identify advantages and disadvantages of the different techniques. We also created a web application that allows users to browse the dataset, available here[5].

Table 1: Count of labels from 3 Wikidata subgraphs

| food item | Count WD labels | Count WP articles | Count Lexeme Senses |
|---|---|---|---|
| purslane | 2 | 40 | 2 |
| oregano | 25 | 95 | 3 |
| kale | 64 | 62 | 4 |
| apple | 142 | 182 | 37 |
| rice | 130 | 179 | 39 |
| chocolate | 118 | 144 | 24 |

In Table 1 we list the count of labels that we found in three of the subgraphs: Wikidata labels, titles of articles in different language versions of Wikipedia, and lexeme senses. For the Wikidata labels subgraphs we find that apple, rice, and chocolate have more labels than kale, oregano or purslane. This distribution is likely due to the popularity of these foods. We anticipate that over time more labels will be added for all of these food items. It is interesting that Wikidata editors have not yet added labels from some of these language versions of Wikipedia to Wikidata.

[5] https://wikifcd.k2.services/multi-lingual-table

In terms of the historical layers of information in Wikidata, creating items for all Wikipedia articles and adding sitelinks between Wikipedias and Wikidata was a very early step in adding data to Wikidata [21]. The order in which data layers were added to Wikidata is likely a factor in why there are fewer lexeme senses that have been connected to Wikidata items. The Lexeme namespace was added to Wikidata in 2018, six years after Wikidata launched, so editors have had less time to contribute data [23].

Table 2: Count of labels from taxon and taxon common name subgraph of Wikidata

| taxon name | Count taxon common name | Count distinct languages |
|---|---|---|
| Portulaca oleracea | 172 | 57 |
| Origanum vulgare | 148 | 59 |
| Brassica oleracea | 22 | 15 |
| Malus domestica | 41 | 41 |
| Oryza sativa | 17 | 4 |
| Theobroma cacao | 5 | 2 |

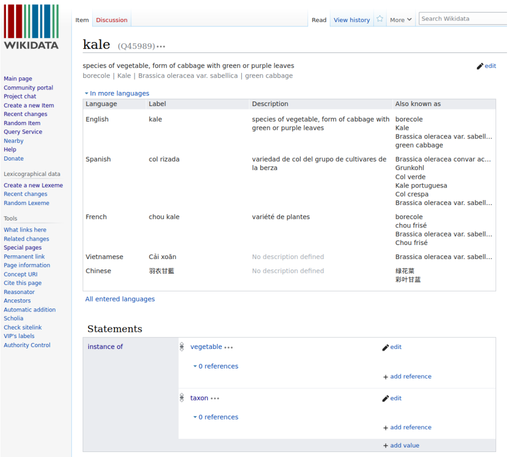
In Table 2 we see counts for the number of statements using the property "taxon common name" on items for the taxa related to the food items from our sample set. The taxon to food item relationships are: Portulaca oleracea / purslane, Origanum vulgare / oregano, Brassica oleracea / kale, Malus domestica / apple, Oryza sativa / rice, and Theobroma cacao / chocolate. We note that the relationships between taxa and food items are not consistently modeled in Wikidata. For example, Brassica oleracea encompasses many cultivars of food items. We were not able to find a closer match for kale in Wikidata at the time of writing.

## 6.1. Wikidata Labels Subgraph

An advantage of sourcing translations from the subgraph that consists of Wikidata food items and their labels is that many editors see these labels. It is likely that this subgraph will grow the most quickly out of the five discussed in this paper in the near term, due to the ease of adding labels to the Wikidata knowledge base. Adding labels in additional languages is work that some editors monitor [20]. Labels, descriptions and aliases are parts of Wikidata items that receive frequent editor attention [27]. A disadvantage of sourcing translations from this subgraph is that labels do not have references, thus we do not know the provenance of the labels. There is some inconsistency in the subgraph of food items that they are sometimes confused with taxon items. For example, in Figure 5, we see that the Wikidata item for 'kale' is both a food item (vegetable) as well as the taxon Brassica oleracea var. sabellica. It would be preferable to have two separate Wikidata items, one for the food item and one for the taxon as is the case for most food items.

While there are two taxon common names listed the Wikidata item for 'kale' (Q45989), there are more taxon common names related to 'kale' listed on the Wikidata item for 'Brassica oleracea' (Q146212), thus we chose to use the item 'Brassica oleracea' for the count of multilingual labels in Table 2.

A disadvantage of sourcing translations from the subgraph that consists of Wikidata food items and their labels are that there are no references for labels. Labels, descriptions, and aliases



kale (Q45989)

species of vegetable, form of cabbage with green or purple leaves
borecole | Kale | Brassica oleracea var. sabellica | green cabbage

| Language | Label | Description | Also known as |
|---|---|---|---|
| English | kale | species of vegetable, form of cabbage with green or purple leaves | borecole<br>Kale<br>Brassica oleracea var. sabell...<br>green cabbage |
| Spanish | col rizada | variedad de col del grupo de cultivares de la berza | Brassica oleracea convar ac...<br>Grunkohl<br>Col verde<br>Kale portuguesa<br>Col crespa<br>Brassica oleracea var. sabell... |
| French | chou kale | variété de plantes | borecole<br>chou frisé<br>Brassica oleracea var. sabell...<br>Chou frisé |
| Vietnamese | Cải xoăn | No description defined | Brassica oleracea var. sabell... |
| Chinese | 羽衣甘藍 | No description defined | 绿花菜<br>彩叶甘蓝 |

Statements

| | | |
|---|---|---|
| instance of | vegetable<br>0 references | |
| | taxon<br>0 references | |

**Figure 5:** Wikidata item for 'kale' is both a taxon and a vegetable

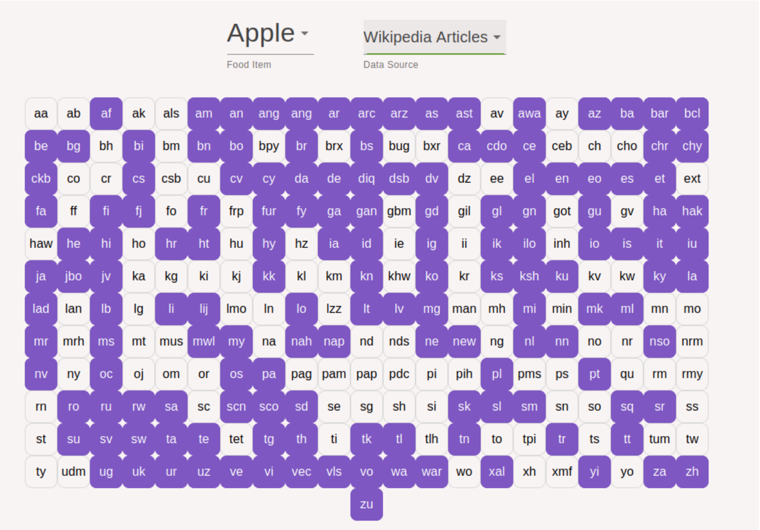
are structured differently than statements on items in Wikidata. So if we are curious about a particular label, we can't turn to a reference for more information. The inconsistency in data modeling practices across editors results in some food items to also describe taxa is another disadvantage. There should be separate Wikidata items for food items and taxa. This will likely be cleaned up by the Wikidata community, but it will require time. Looking at the labels for 'apple' from this subgraph, we find labels in Faroese, 'súrepli', and Navaho, 'Bilasáana' among the 142 labels available in Wikidata.

## 6.2. Article Titles from Sitelinks to Wikipedias

Sourcing labels from the subgraph of food items and their sitelinks to different language versions of Wikipedia involves the connections between Wikidata items and Wikipedia articles, if the food item is well-known, there may be articles in many different language versions of Wikipedia, as seen in Figure 6. Using this subgraph has the advantage of precise food item matches to articles about those foods, thus the article title is usually a reliable source for multilingual data. Article titles are high-visibility in that they are frequently seen by readers, and thus errors are corrected more quickly. Sourcing labels from the subgraph of food items and their sitelinks with different language versions of Wikipedia has the disadvantage that new article titles get added when new articles are written, which is not a fast process. This contrasts with the ease of adding additional labels to Wikidata. Rather than typing a string and pressing save to add a label in Wikidata, more effort is required for an editor to create a new article in Wikipedia, the title of which would then become an additional multilingual label candidate.

## 6.3. Taxon Common Name Subgraph

Some of the advantages of sourcing translations from the subgraph that consists of Wikidata taxon items and their taxon common names include the fact that editors can contribute references on these statements, and that this subgraph is likely to grow over time. As of October, 2022



Food Item: Apple

Data Source: Wikipedia Articles

aa ab af ak als am an ang ang ar arc arz as ast av awa ay az ba bar bcl
be bg bh bi bm bn bo bpy br brx bs bug bxr ca cdo ce ceb ch cho chr chy
ckb co cr cs csb cu cv cy da de diq dsb dv dz ee el en eo es et ext
fa ff fi fj fo fr frp fur fy ga gan gbm gd gil gl gn got gu gv ha hak
haw he hi ho hr ht hu hy hz ia id ie ig ii ik ilo inh io is it iu
ja jbo jv ka kg ki kj kk kl km kn khw ko kr ks ksh ku kv kw ky la
lad lan lb lg li lij lmo ln lo lzz lt lv mg man mh mi min mk ml mn mo
mr mrh ms mt mus mwl my na nah nap nd nds ne new ng nl nn no nr nso nrm
nv ny oc oj om or os pa pag pam pap pdc pi pih pl pms ps pt qu rm rmy
rn ro ru rw sa sc scn sco sd se sg sh si sk sl sm sn so sq sr ss
st su sv sw ta te tet tg th ti tk tl tlh tn to tpi tr ts tt tum tw
ty udm ug uk ur uz ve vi vec vls vo wa war wo xal xh xmf yi yo za zh
zu

**Figure 6:** Screenshot of the visualization of labels for ‘apple’ from article titles from Wikipedia language versions in our webapp available at https://wikifcd.k2.services/multi-lingual-table.

there are more than 780,000 uses of P1843 ‘taxon common name’.

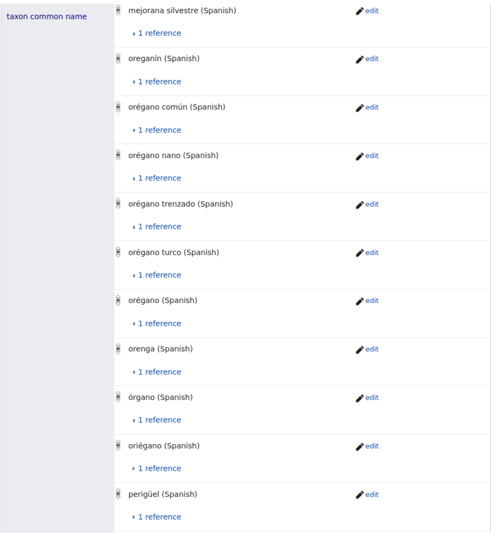
Sourcing labels from the subgraph of Wikidata taxon items and their taxon common names has the disadvantage that sometimes when multiple common names are provided per language, it will require manual review to determine which of this would be appropriate for consideration from FoodOn. In the cases where multiple common names are provided for a language, it is not always clear if the labels provided all refer to the same species. For example, in Figure 7 we see that several common names are listed on the Wikidata item for Origanum vulgare are from Spanish. Without further research it is not clear which of these might be the best fit for FoodOn. Some of them could even be for specific subspecies of Origanum vulgare, and may need to be moved to the subspecies items as they are added to Wikidata.

We anticipate that more Wikidata editors will continue to contribute taxon common name statements, and that this subgraph will continue to grow over time.

## 6.4. Lexeme Senses Subgraph

The advantages of sourcing multilingual labels from the subgraph of Wikidata Lexeme Senses that connect to food items include the fact that there is room for references on these statements, and that the subgraph is likely to grow as more editors contribute to the L namespace. In Figure 8, we can see that there are a small number of labels that we can source for ’apple’ from the L namespace. The Wikidata community creates Lexeme challenges periodically to encourage participation in the L namespace. For example, a recent challenge related to vegetables resulted in the addition of more of the connections between food items and lexeme senses [6]. With the creation of the Abstract Wikipedia community, we believe it is likely that more editors will be

[6] https://dicare.toolforge.org/lexemes/challenge.php?id=64



taxon common name

mejorana silvestre (Spanish) edit
› 1 reference

oreganín (Spanish) edit
› 1 reference

orégano común (Spanish) edit
› 1 reference

orégano nano (Spanish) edit
› 1 reference

orégano trenzado (Spanish) edit
› 1 reference

orégano turco (Spanish) edit
› 1 reference

orégano (Spanish) edit
› 1 reference

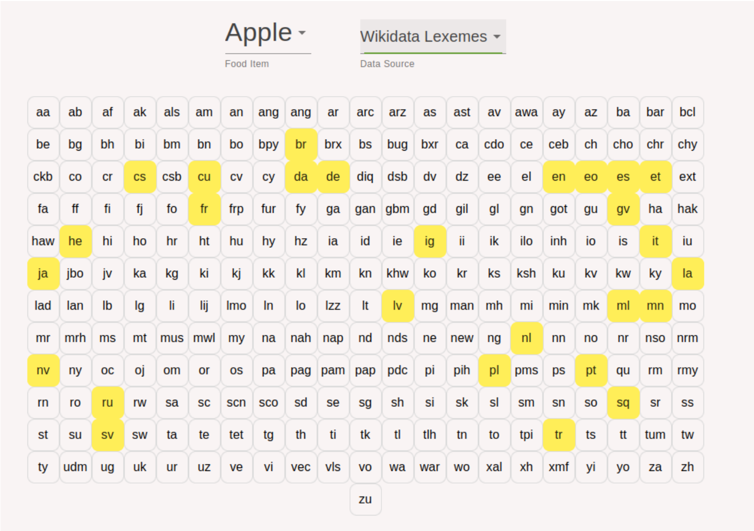
orenga (Spanish) edit
› 1 reference

órgano (Spanish) edit
› 1 reference

oriégano (Spanish) edit
› 1 reference

perigüel (Spanish) edit
› 1 reference

**Figure 7:** Multiple common names from Spanish listed for Origanum vulgare



| | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| aa | ab | af | ak | als | am | an | ang | ang | ar | arc | arz | as | ast | av | awa | ay | az | ba | bar | bcl |
| be | bg | bh | bi | bm | bn | bo | bpy | **br** | brx | bs | bug | bxr | ca | cdo | ce | ceb | ch | cho | chr | chy |
| ckb | co | cr | **cs** | csb | **cu** | cv | cy | **da** | **de** | diq | dsb | dv | dz | ee | el | **en** | **eo** | **es** | **et** | ext |
| fa | ff | fi | fj | fo | **fr** | frp | fur | fy | ga | gan | gbm | gd | gil | gl | gn | got | gu | **gv** | ha | hak |
| haw | **he** | hi | ho | hr | ht | hu | hy | hz | ia | id | ie | **ig** | ii | ik | ilo | inh | io | is | **it** | iu |
| **ja** | jbo | jv | ka | kg | ki | kj | kk | kl | km | kn | khw | ko | kr | ks | ksh | ku | kv | kw | ky | **la** |
| lad | lan | lb | lg | li | lij | lmo | ln | lo | lzz | lt | **lv** | mg | man | mh | mi | min | mk | **ml** | **mn** | mo |
| mr | mrh | ms | mt | mus | mwl | my | na | nah | nap | nd | nds | ne | new | ng | **nl** | nn | no | nr | nso | nrm |
| **nv** | ny | oc | oj | om | or | os | pa | pag | pam | pap | pdc | pi | pih | **pl** | pms | ps | **pt** | qu | rm | rmy |
| rn | ro | **ru** | rw | sa | sc | scn | sco | sd | se | sg | sh | si | sk | sl | sm | sn | so | **sq** | sr | ss |
| st | su | **sv** | sw | ta | te | tet | tg | th | ti | tk | tl | tlh | tn | to | tpi | **tr** | ts | tt | tum | tw |
| ty | udm | ug | uk | ur | uz | ve | vi | vec | vls | vo | wa | war | wo | xal | xh | xmf | yi | yo | za | zh |
| | | | | | | | | | | zu | | | | | | | | | | |

**Figure 8:** Labels for apple from the Lexeme namespace in Wikidata

drawn to contribute to the L namespace so that more data from the lexeme namespace will be available for reuse by Abstract Wikipedia [28].

## 6.5. WikiFCD Common Names Subgraph

In WikiFCD we created a property, "common name" (P76), that we use to record common names in multiple languages. We add these statements to WikiFCD whenever additional food item names are provided in the source food composition tables. An advantage of sourcing labels from WikiFCD is that all common name statements have a reference back to a publication. For example, one of the food composition tables we integrated into WikiFCD is the SMILING Food composition table for Vietnam. Food item labels are provided by the team of authors who prepared the food composition measurements for each food item. In this food composition table both the food item name in Vietnamese and the food item name in English are provided.

Currently there are five hundred sixty three food items in this subgraph. A disadvantage of sourcing labels from WikiFCD is that the corpus grows more slowly than the Wikidata subgraphs because a smaller number of people contribute to WikiFCD than Wikidata. We anticipate that as more people contribute data to WikiFCD this subgraph will grow more quickly. We chose to exclude some of the labels that the SPARQL query for common names from WikiFCD returned because they were the names of dishes rather than food items. Some food composition tables include data for dishes of multiple ingredients as well as individual food items in their datasets. We also found that there were differences in practices among the authors of food composition tables. For example, some authors provide taxon information at the species level and some provide taxon information at the varietal or subspecies levels.

## 7. Creating a Dataset of Multilingual Labels for Food Items

Anyone can reuse data from WikiFCD or Wikidata for any purpose. One or more of the subgraphs described in this paper could be reused as a source of labels for any food-related application that requires multilingual access terms for food items. Additionally these subsets could be periodically monitored for updates from the communities. We anticipate that all of these subgraphs will be extended by editors adding new information to these wikis. As Wikidata editors notice errors, they will address them and improve them [29].

We explored five distinct subgraphs in order to understand which of these would return relevant label candidates. Another advantage of collecting labels from multiple subgraphs is that we can cross-check labels with the results of another subgraph. For example, if we look at the leafy green called 'kale' in English, for the Wikipedia article subset we find that the German language article is titled 'Grünkohl'. The Wikidata item has the German label 'Grünkohl'. Seeing the same label in different subsets increases our confidence that the label is accurate.

Each of the food items in our sample data set had dozens of distinct language labels across the sets of relevant subgraphs. In total we counted 173 distinct languages with labels for 'apple', 134 distinct languages with labels for 'chocolate', 68 distinct languages with labels for 'kale', 82 distinct languages with a label for 'oregano', 65 distinct languages with labels for 'purslane' and 151 distinct languages with labels for 'rice'. Currently the labels in Wikidata are unevenly distributed across the supported languages. We anticipate that as more editors join the Wikidata community they will contribute many more labels in additional languages. Reusing multilingual content from Wikidata in open scientific projects will increase the accessibility of data produced. For domains such as food and human nutrition, multilingual data can be shared more widely which could impact more people.

## 8. Creating Wikibases for Interoperability

We chose to create WikiFCD so that we could design a data model to accommodate food composition data sourced from a diverse range of published sources. Our decision to use Wikibase allowed us to provide web-based access to WikiFCD, meaning anyone can find this data online. We had an explicit data model in mind, inspired by the structure of the published food composition tables we drew from as our data sources. The data in WikiFCD are not all appropriate for Wikidata. By connecting some items and some properties in WikiFCD to their Wikidata correlates through mapping statements, we are able to treat the two resources as if they were a single knowledge base through writing federated SPARQL queries. The SPARQL endpoint for WikiFCD supports federated queries that allow us to combine data from WikiFCD with data from Wikidata.

In the future, if certain subsets of WikiFCD are of interest to the Wikidata community, we are well-positioned to quickly contribute data to Wikidata itself. We will be able to leverage the mappings that we created between food items with FoodOn identifiers and appropriate Wikidata items. As more and more people create Wikibase instances for specialized data, we all have more data to combine with different subsets of Wikidata through federated queries.

## 9. Conclusion

The five subgraphs we describe in this paper were created by different communities of editors across multiple projects of the Wikimedia Foundation and the ecosystem of Wikibases. Editors add content to different language versions of Wikipedia, to Wikidata, and to WikiFCD. As more people contribute multilingual content, all of these projects become more accessible to additional language communities.

Some decisions about the suitability of sourcing labels from any combination of the subgraphs described in this paper will require manual review. In some cases there will be multiple candidate labels per languages, and a curator will need to evaluate them for suitability. The breadth of languages covered by these label subgraphs and the fact that the data is free to reuse make this an attractive source of multilingual content.

Enriching WikiFCD with multilingual labels is a priority for our project because we want to be sure that FAIR food composition data for a broad range of foods from diverse cuisines are easily accessible on the web at no cost. As nutrition plays a part in maintaining health, people who need data about foods that are not found in popular sources like Food Data Central will be able to find food composition data in WikiFCD. We created an interactive webapp for our sample dataset so that others can quickly compare the availability of labels across the subsets we discussed[7].

Developers of applications, ontologies, vocabularies, and other resources may need a free-to-reuse source for multilingual content. As a multilingual knowledge base, Wikidata contains labels in more than three hundred human languages. This means that anyone who needs to source multilingual labels for words could explore Wikidata to see how complete the current label inventory is. Depending on the domain, one or more of these subgraphs may contain

[7] https://wikifcd.k2.services/multi-lingual-table

enough multilingual labels to increase coverage for a specific project or use case. The purpose of demonstrating five distinct subgraphs in this paper is to emphasize that there is multilingual content in different layers of the projects of the Wikimedia Foundation.

## Acknowledgments

We thank the Joint Food Ontology Working Group for productive discussions about FoodOn. We thank the Wikidata community for continuing to improve the Wikidata knowledge base.

## References

[1] 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.

[2] Tidball, Tidball, Curtis, The absence of wild game and fish species from the usda national nutrient database for standard reference: Addressing information gaps in wild caught foods 53 (2014). doi:10.1080/03670244.2013.792077.

[3] A. Durazzo, E. Camilli, S. Marconi, S. Lisciani, P. Gabrielli, L. Gambelli, A. Aguzzi, M. Lucarini, J. Kiefer, L. Marletta, Nutritional composition and dietary intake of composite dishes traditionally consumed in italy, Journal of Food Composition and Analysis 77 (2019) 115–124.

[4] N. Monjotin, M. J. Amiot, J. Fleurentin, J. M. Morel, S. Raynal, Clinical evidence of the benefits of phytonutrients in human healthcare, Nutrients 14 (2022). URL: https://www.mdpi.com/2072-6643/14/9/1712. doi:10.3390/nu14091712.

[5] 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.

[6] H. Mechchate, A. El Allam, N. El Omari, N. El Hachlafi, M. A. Shariati, P. Wilairatana, M. S. Mubarak, A. Bouyahya, Vegetables and their bioactive compounds as anti-aging drugs, Molecules 27 (2022). URL: https://www.mdpi.com/1420-3049/27/7/2316. doi:10.3390/molecules27072316.

[7] G. Menichetti, A.-L. Barabási, Nutrient concentrations in food display universal behaviour, Nature Food 3 (2022) 375–382.

[8] 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.

[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] K. Thornton, K. Seals-Nutt, M. Matsuzaki, D. Damion, Reuse of the foodon ontology in a knowledge base of food composition data, Semantic Web Journal (2023).

[11] J. E. Labra-Gayo, A. Ammar, D. Brickley, D. F. Álvarez, A. G. Hevia, A. J. Gray, E. Prud'hommeaux, D. Slater, H. Solbrig, S. A. H. Beghaeiraveri, et al., Knowledge graphs and wikidata subsetting, BioHackathon Europe 2020 (2021). URL: https://biohackrxiv.org/wu9et/.

[12] "FORCE11", The fair data principles (2014). https://www.force11.org/group/fairgroup/fairprinciples.

[13] B. Mons, C. Neylon, J. Velterop, M. Dumontier, L. O. B. da Silva Santos, M. D. Wilkinson, Cloudy, increasingly fair; revisiting the fair data guiding principles for the european open science cloud, Information Services & Use (2017) 1–8.

[14] S. Haussmann, O. Seneviratne, Y. Chen, Y. Ne'eman, J. Codella, C.-H. Chen, D. L. McGuinness, M. J. Zaki, Foodkg: a semantics-driven knowledge graph for food recommendation, in: International Semantic Web Conference, Springer, 2019, pp. 146–162.

[15] E. Chazelas, M. Deschasaux, B. Srour, E. Kesse-Guyot, C. Julia, B. Alles, N. Druesne-Pecollo, P. Galan, S. Hercberg, P. Latino-Martel, et al., Food additives: distribution and co-occurrence in 126,000 food products of the french market, Scientific reports 10 (2020) 1–15.

[16] 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).

[17] 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.

[18] 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.

[19] 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.

[20] L.-A. Kaffee, A. Piscopo, P. Vougiouklis, E. Simperl, L. Carr, L. Pintscher, A Glimpse into Babel: An Analysis of Multilinguality in Wikidata, in: Proceedings of the 13th International Symposium on Open Collaboration, OpenSym '17, ACM, New York, NY, USA, 2017, pp. 14:1–14:5. URL: https://doi.org/10.1145/3125433.3125465. doi:10.1145/3125433.3125465.

[21] D. Vrandečić, Wikidata: A new platform for collaborative data collection, in: Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 1063–1064.

[22] Meta, List of wikipedias meta, discussion about wikimedia projects, 2022. URL: https://meta.wikimedia.org/w/index.php?title=List_of_Wikipedias&oldid=23800107, [Online; accessed 8-October-2022].

[23] B. Cartoni, D. C. Aros, D. Vrandečić, S. Lertpradit, Introducing lexical masks: a new representation of lexical entries for better evaluation and exchange of lexicons, in: Proceedings of the 12th Language Resources and Evaluation Conference, 2020, pp. 3046–3052.

[24] F. Nielsen, Lexemes in wikidata: 2020 status, in: Proceedings of the 7th Workshop on

Linked Data in Linguistics (LDL-2020), 2020, pp. 82–86.

[25] Ordia, Ordia statistics, 2022. URL: https://ordia.toolforge.org/statistics/, [Online; accessed 13-October-2022].

[26] F. Å. Nielsen, Ordia: A web application for wikidata lexemes, in: European Semantic Web Conference, Springer, 2019, pp. 141–146.

[27] T. Pellissier Tanon, L.-A. Kaffee, Property label stability in wikidata: evolution and convergence of schemas in collaborative knowledge bases, in: Companion Proceedings of the The Web Conference 2018, 2018, pp. 1801–1803.

[28] D. Vrandečić, Building a multilingual wikipedia, Communications of the ACM 64 (2021) 38–41.

[29] K. Shenoy, F. Ilievski, D. Garijo, D. Schwabe, P. Szekely, A study of the quality of wikidata, Journal of Web Semantics 72 (2022) 100679. URL: https://www.sciencedirect.com/science/article/pii/S1570826821000536. doi:https://doi.org/10.1016/j.websem.2021.100679.