MULTILINGUAL LABELS FOR DIGITAL PRESERVATION

Katherine Thronton
Yale University Library
United States
katherine.thornton@yale.edu
0000-0002-4499-0451

Kenneth Seals-Nutt
Yale University Library
United States
kenneth.seals-nutt@yale.edu
0000-0002-5926-9245

Abstract – We introduce a technique for finding multilingual translations for lists of words using technologies of the Semantic Web. We present four subsets of data from Wikidata and Wikipedia as sources of multilingual labels. Our sample dataset consists of seven terms related to digital preservation. We compare the number of labels we can source for these terms from other human languages via SPARQL queries using the Wikidata Query Service. After discussing the composition of each subset, we detail their advantages and disadvantages. Providing multilingual la- bels as additional access points for resources such as on- tologies, vocabularies and user interfaces for applications increases the relevance of these resources to a larger per- centage of the global population. Increasing multilingual access promotes inclusion for a broader range of people, which leads to greater diversity in the digital preservation community.

Keywords – Wikidata, Semantic Web, multilingual data, knowledge graph subsets

Themes – Digital Accessibility, Inclusion, and Diversity, We're All in this Together

I. INTRODUCTION

Digital preservation is an international field made up of practitioners from all parts of the globe. Resources such as ontologies, vocabularies, and applications relevant to the work of digital preservation are frequently monolingual. English is used as the primary language for many resources. Such monolingual resources restrict their audience to people who have knowledge of English, while all others are excluded. Increasing the number of multilingual access points within such resources promotes equity and broadens the diversity of audiences who can benefit from the field of digital preservation. Maintainers of monolingual resources are faced with budgetary constraints, and may argue that expanding multilingual access is too expensive to be practical. Translations created by human experts are expensive, perhaps we can leverage the technologies of the Semantic Web to source multilingual labels as a cost-effective alternative.

We describe four subsets of Wikidata that people may find useful for sourcing multilingual labels. We created a sample data set to test for multilingual label coverage, and we describe the results of consulting four subsets of Wikidata for each term in the sample data set. After describing each subset, we discuss advantages and disadvantages of each. We introduce an interactive application that we created to browse each set of multilingual labels. We offer this work as a demonstration of how communities of editors who contribute to the projects of the Wikimedia Foundation have created a valuable multilingual knowledge graph. The fact that all of the data in Wikidata is free for anyone to reuse for any purpose makes this a shared international resource. Times of international crises such as global pandemic, or armed conflict, reinforce the importance of striving to make resources more equitably available. Providing cost-effective strategies for sourcing multilingual labels is a pathway to promoting equity through increasing multilingual access.

II. WIKIDATA

Wikidata is a project of Wikimedia Deutschland, the German chapter of the Wikimedia Foundation. The Wiki- data community launched this public knowledge base of structured data in 2012. The

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121100

architecture of Wiki- data was designed from the outset to support multilingual content [1]. The Wikidata knowledge base contains Items that can be connected to literal values, or to other Items, through the use of Properties [2]. The Wikidata community has added more labels in English than any other supported language, but there are dozens of additional languages for which the Wikidata community has also added many labels [3]. The work of the members of the Wikidata community to add multilingual labels to Items and Properties has resulted in a corpus of equivalent labels across hundreds of human languages.

III. Related Work

Wikidata is an example of a collaboratively-created knowledge graph [4]. After ten years of existence, Wiki- data is well-recognized as valuable source for reusable data [5]. Researchers have leveraged multilingual content from Wikidata for various applications. For example, multilingual content from Wikidata has been used to power a question-answering platform [6], and has been used to generate article placeholders for encyclopedias [7]. The challenge of organizing access terms for multi- lingual digital content has been addressed by language- independent mappings drawn from the Semantic Web [8]. Multilingual access is necessary for national con- texts in which multiple languages are supported [9]. Due to the fact that the digital preservation community is an international community, it is clear that we need to pro- vide access to our applications and resources in a wide range of human languages [10].

We sampled several subsets of Wikidata for this work. Wikidata subsets are portions of the Wikidata knowledge graph [11]. The size of Wikidata makes it desirable to reuse a subset, as it is time-consuming to process and host the entire Wikidata graph. Often subsets are focused around a particular type of data, or a specific domain. We identified subsets of Wikidata related to seven sample terms, and wrote SPARQL queries to ex- tract associated data from Wikidata. Subsets of Wikidata may be extracted by a variety of software tools, or via the Wikidata Query Service. An overview of tools available to extract subsets from Wikidata is provided by [12]. Researchers have also explored memory-efficient techniques that allow for larger subsets to be extracted more quickly than techniques that use Wikidata's SPARQL end- point [13].

Researchers and practitioners approach the translation of ontologies, vocabularies, and other term-based resources using a variety of methods. One approach is to extend an ontology with multilingual labels [14]. A successful tri-lingual project is described in [15]. Others have explored using Wikipedias to generate translations [16]. Our approach differs in that we combine translations from Wikipedias along with additional multilingual content from Wikidata, thus extending coverage from additional human languages.

IV. Sample Data Set

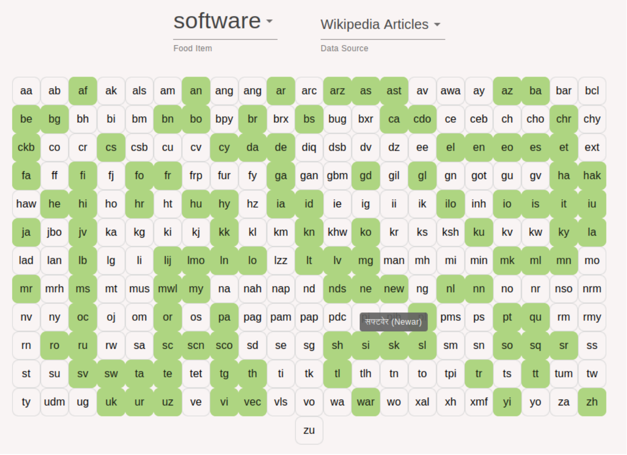
We selected seven terms related to digital preservation to create a sample data set[1]. The terms we included are: file format, checksum, operating system, data integrity, software, license, and reproducibility. We chose these terms because of their relevance to digital preservation work activities. We then searched the Wikidata knowledge base to gather the Qids for the relevant Wiki- data items. Using the Qids for these terms, we wrote SPARQL queries to identify multilingual labels for these terms. The Wikidata items served as the basis for three of the subsets: the Wikidata Item Labels, the Wikipedia Article Titles, and the Wikidata Lexemes. To find our fourth subset, we searched the Property namespace for our terms to retrieve the Property Labels.

We created an interactive application that presents the labels available in each of the four subsets for each of the words in our sample dataset[1]. This application allows anyone to quickly compare the language cover- age per subset for each term. For example, in Table 1, we see a visualization of the languages (represented by their ISO codes[2]) color-coded if we have a label in that language, and without color if we do not. As the user hovers over a language, the label itself will be dis- played alongside the name of the language. There are drop-down menus that users can select from in order to switch between terms from the sample data set and to switch between the four subsets. The layout of languages is consistent across the different views, allowing visual comparison of the overlap between

[1] The webapp that includes the interactive table is available at https://wikidp-research.k2.services/multi-lingual-table.

[2] https://www.iso.org/iso-639-language-codes.html

subsets. In Figure 2, we see the labels available from the Wikidata item for ‘software’ (Q7397).

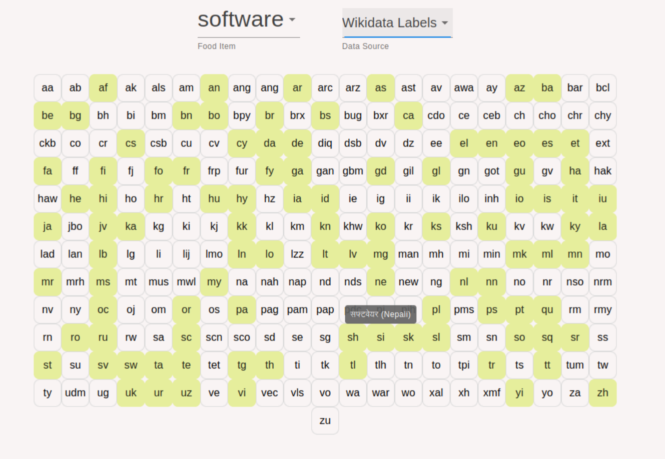


software ▾ — Food Item

Wikipedia Articles ▾ — Data Source

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	सफ्टवेर (Newar)			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 1: Labels for ‘software’ from Wikipedia Article titles, as seen in https://wikidp- research.k2.services/multi-lingual-table



software ▾ — Food Item

Wikidata Labels ▾ — Data Source

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	सफ्टवेयर (Nepali)			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 2: Labels for ‘software’ from Wikidata, as seen in https://wikidp-research.k2.services/multi-lingual-table

V. Reusing Multilingual Content from Wikimedia Projects

The human editors working to create and extend content in Wikimedia projects are constantly working to improve the quality of information across the projects. The large number of people who view and edit this con- tent help to remove errors and ensure accuracy. Content in Wikidata is published under the Creative Com- mons Zero license, meaning that data in Wikidata is free for anyone to reuse for any purpose. The Wikidata SPARQL endpoint[3] is a public endpoint that anyone can use to request data from Wikidata [17]. No credentials are needed to run queries on Wikidata’s SPARQL end- point, making this a convenient method of data retrieval. We introduce four subsets in this section: Wikipedia Article Titles, Wikidata Item Labels, Wikidata Property Labels and Wikidata Lexemes. Data from each of these subsets is available from the Wikidata Query Service.

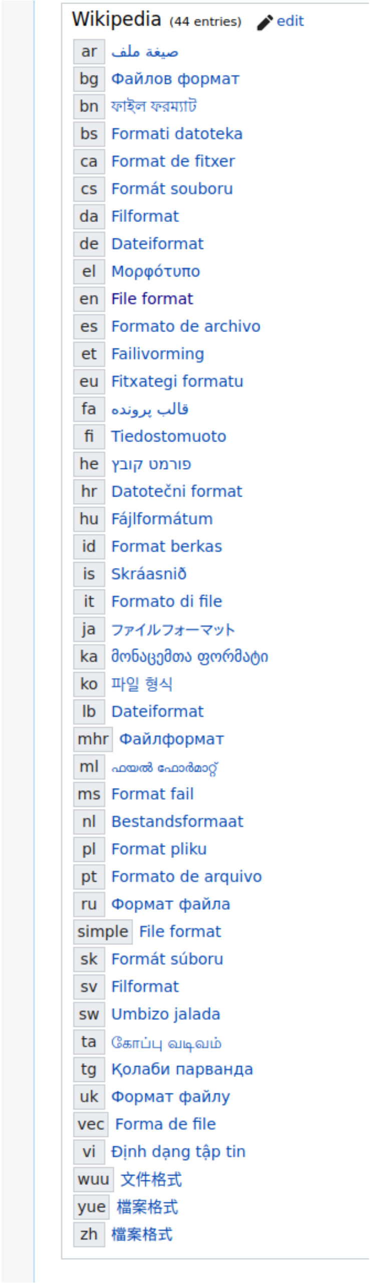
A. *Article Names per Language Version of Wikipedia*

One early layer of data in Wikidata is that of interwiki links. Interwiki links connect articles that describe a topic among the different language versions of Wikipedia. These interwiki links are now stored in Wikidata, meaning that Wikidata items are connected to corresponding Wikipedia articles [1]. The titles of the articles in the different language versions are a potential source of multilingual labels for these terms. New language versions of Wikipedia are still being created. There are more than three hundred versions of Wikipedia [18]. Hypothetically, if every language version were to have an article about file formats, we would then have hundreds of multilingual labels from the set of article titles. We wrote SPARQL queries to return the article titles from each of these language versions of articles about file formats.

For example, there are 44 versions of Wikipedia that have an article about file formats, as seen in Figure 3. We can retrieve all of these article titles and consider them multilingual label candidates. The largest number of la- bels from the Wikipedia Articles subset is available for ‘operating system’ with 150 potential labels. This is due to the fact that more Wikipedia communities have writ- ten articles about ‘operating system’ than about any of the other terms from our sample set. Only twenty-seven language versions of Wikipedia have articles about ‘reproducibility’. This is likely due to the frequency of usage of these terms, and thus relevance for an average contributor to Wikipedia. The Google Books Ngram Viewer[4], which presents occurrence data for search terms as seen in the corpus of Google Books, demonstrates that ‘operating system’ is found more frequently than ‘reproducibility’ between the years 1960-2019, as seen in Figure 4.

[3] https://query.wikidata.org/

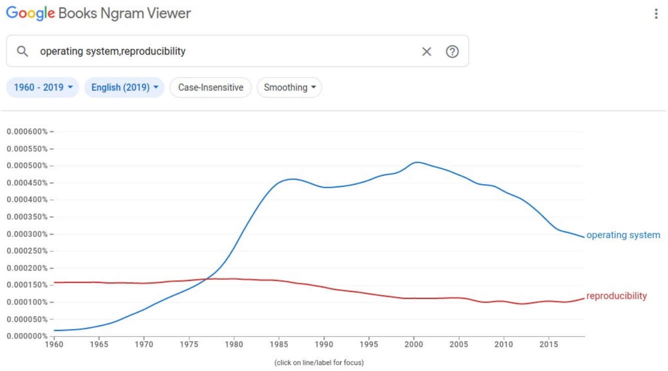
[4] https://books.google.com/ngrams/



Wikipedia (44 entries) edit

ar	صيغة ملف
bg	Файлов формат
bn	ফাইল ফরম্যাট
bs	Formati datoteka
ca	Format de fitxer
cs	Formát souboru
da	Filformat
de	Dateiformat
el	Μορφότυπο
en	File format
es	Formato de archivo
et	Failivorming
eu	Fitxategi formatu
fa	قالب پرونده
fi	Tiedostomuoto
he	פורמט קובץ
hr	Datotečni format
hu	Fájlformátum
id	Format berkas
is	Skráasnið
it	Formato di file
ja	ファイルフォーマット
ka	მონაცემთა ფორმატი
ko	파일 형식
lb	Dateiformat
mhr	Файлформат
ml	ഫയൽ ഫോർമാറ്റ്
ms	Format fail
nl	Bestandsformaat
pl	Format pliku
pt	Formato de arquivo
ru	Формат файла
simple	File format
sk	Formát súboru
sv	Filformat
sw	Umbizo jalada
ta	கோப்பு வடிவம்
tg	Қолаби парванда
uk	Формат файлу
vec	Forma de file
vi	Định dạng tập tin
wuu	文件格式
yue	檔案格式
zh	檔案格式

Figure 3: Interwiki links for the different language ver- sions of Wikipedia that contain an article titled ‘file for- mat’.



Google Books Ngram Viewer

operating system,reproducibility

1960 - 2019, English (2019), Case-Insensitive, Smoothing

Figure 4: Google Books Ngram Viewer results for ‘oper- ating system’ and ‘reproducibility’ from 1960-2019

In Table 1 we see the count of labels available in the Wikipedia Article Title subgraph for each of the terms in our sample data set. An advantage of sourcing labels from this subset is that article titles have high visibility within Wikipedias, thus these labels are likely to be corrected very quickly if they are vandalized or require improvement. A disadvantage of sourcing labels from this subset is that new articles are created on a relatively slow timeline, thus this subset is likely to grow slowly. If we compare the work involved in writing a new article in Wikipedia with adding a label to a Wikidata item, writing a new article requires substantially more effort.

Term	Count Wikipedia Article Titles
file format	44
checksum	37
operating system	150
data integrity	30
software	131
license	65
reproducibility	27

Table 1: Count of Labels from Wikipedia Article Titles

B. *Multilingual Item Labels from Wikidata*

The designers of the data model for Wikidata intended it to be a multilingual knowledge base [1]. Each item in Wikidata has a Qid identifier composed of the letter Q and numbers. These Qids are designed to avoid privileging one human language over others supported by the knowledge base. The Wikidata data model sup- ports labels in more than three hundred languages [3]. Wikidata editors add labels in many different languages. Over time, the set of all labels for a particular item becomes a very useful set of translations.

An advantage of sourcing labels from this subset is that Wikidata labels are added at a faster pace than new articles are created, this this subset is likely to grow more quickly over time. In order to add a label, users type the string into the user interface in the area designated for the language of choice, and then press ‘publish’ to contribute the content. Wikidata item labels are seen by many editors, as well as by many people who reuse data from Wikidata, thus these labels are likely to be updated quickly if they require improvement. Multilingual labels are an aspect of Wikidata that some editors monitor closely [3].

In Table 2 we see the count of labels available in the Wikidata Items subgraph. To date the terms from our sample data set with the largest number of available la- bels are ‘software’ and ‘operating system’, each with la- bels in more than one hundred human languages. The terms ‘checksum’ and ‘reproducibility’ have fewer available labels.

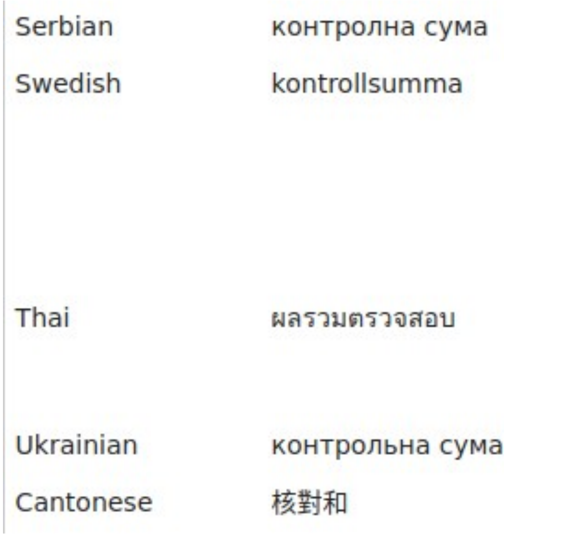
C. *Multilingual Property Labels from Wikidata*

Four of the terms in our sample dataset are related to properties in Wikidata. Wikidata properties are predicates that describe how items are related to one another. Properties are also modeled to accommodate la- bels in all languages supported by Wikidata. Some members of the Wikidata community specialize in working on property labels [19]. We can consult the subgraph of property labels for our sample dataset to see if there are any additional labels in languages not yet covered by the other subsets. While this may result in some redundant labels, as we would expect the labels to be the same for the item and the property, there could be some additional languages that have coverage in the property label subgraph. For example, in Figure 5, we see some of the labels available for the Wikidata item ‘checksum’ include labels from Thai and Ukrainian, but not Turkish. In contrast, in Figure 6, we see that a Turkish label is available.

Term	Count Wikidata Item Labels
file format	55
checksum	39
operating system	105
data integrity	33
software	103
license	70
reproducibility	35

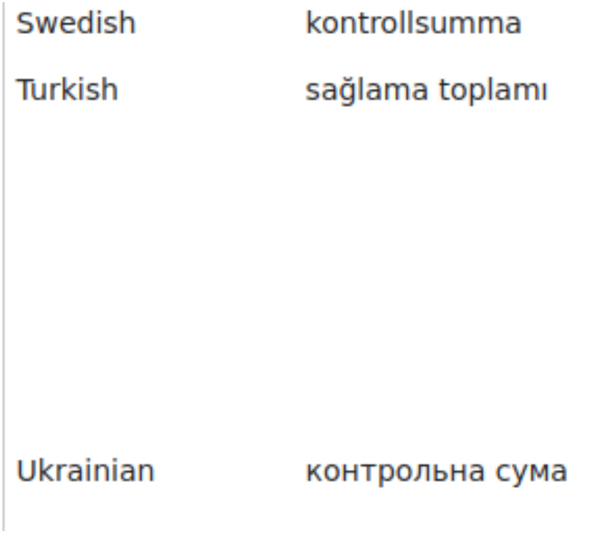
Table 2: Count of Labels from Wikidata Item Subgraph

A disadvantage of sourcing labels from this subgraph is that there are a limited number of properties in Wikidata. There are currently more than 10,000 proper- ties in Wikidata, but more than 100,000,000 items. Thus there are many terms that will not be found among properties. In Table 3 we see the count of labels available in the Wikidata Property label subgraph. For terms in our sample data set that are not related to a Wikidata property, we recorded N/A in the table.



Serbian	контролна сума
Swedish	kontrollsumma
Thai	ผลรวมตรวจสอบ
Ukrainian	контрольна сума
Cantonese	核對和

Figure 5: Some of the labels available for the Wikidata item ‘checksum’ (Q218341)



Swedish	kontrollsumma
Turkish	sağlama toplamı
Ukrainian	контрольна сума

Figure 6: Some of the labels available for the Wikidata property ‘checksum’ (P4092)

Term	Count Wikidata Property Labels
file format	38
checksum	26
operating system	69
data integrity	N/A
software	N/A
license	76
reproducibility	N/A

Table 3: Count of Labels from Wikidata Property Sub- graph

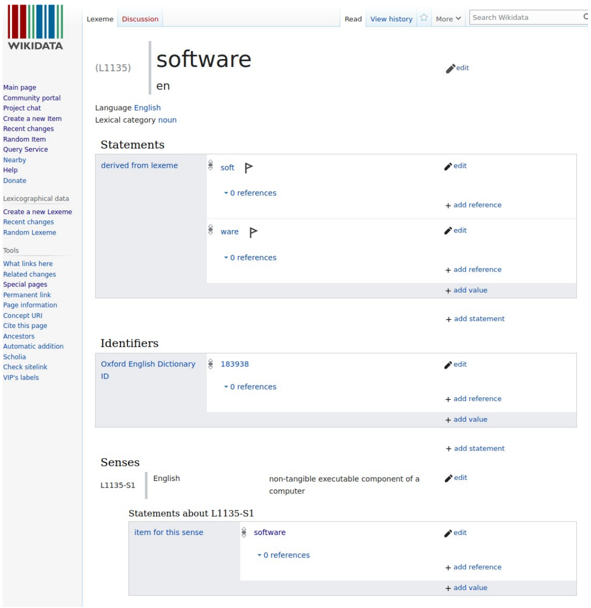
D. *Multilingual Property Labels from Wikidata*

Wikidata also contains detailed linguistic data in the Lexeme namespace. Community members create lexemes, forms, and senses in the Lexeme namespace following the data model for lexicographical data [20]. The Lexeme namespace, namespace L, was created in 2018 [21]. Wikidata has a property that is used to connect Lexeme senses to corresponding Wikidata items. The property has the English label ‘item for this sense’ and is P5137. Through the use of this property, the Wiki- data items from our sample data set can be connected to lexeme senses. In Figure 7, we see the lexeme ‘software’ (L1135). In the section of the page with the heading ‘Senses’ we see that the property ‘item for this sense’ has the value ‘software’ which is the Wikidata item identified with Q7397.

The graph of lexeme senses and their connections to Wikidata items is likely to increase in size over time. Currently, for this sample data set there are zero lexeme senses for ‘checksum’ and ‘data integrity’. This is likely due to the fact that these concepts are domain-specific, and relatively infrequently used by people who are not engaged with the domain of computing.

A useful tool for searching for lexemes is Ordia [22]. Ordia can be used to search for lexemes and provides overviews of connections between lexemes and other content. Wikidata editors to the Lexeme namespace have already contributed more than half a million lexical entries [23].

An advantage of sourcing labels from this subset is that it is likely to grow in the future. As more editors use the property ‘item for this sense’ P5137 to connect Lexeme senses to Wikidata items, this subgraph will grow. Lexemes are connected to external identifiers related to etymology, dictionaries and other linguistic resources. Depending on the type of resource for which you are sourcing multilingual labels, pointers to additional linguistic information may also be helpful. Another advantage is that the data model for lexicographic data in the L namespace accommodates the use of references. Lexemes can also be connected to authoritative sources from which information was sourced. For example, in Figure 9, we see that the Swedish noun ‘licens’ is sourced back to Svenska Akademiens Ordbok using the property ‘described by source’ in Wikidata’s Lexeme namespace. This increases the value of labels sourced from the lexeme subset as they may also include provenance information.



(L1135) software
en

Language English
Lexical category noun

Statements

derived from lexeme: soft — 0 references; ware — 0 references

Identifiers

Oxford English Dictionary ID: 183938 — 0 references

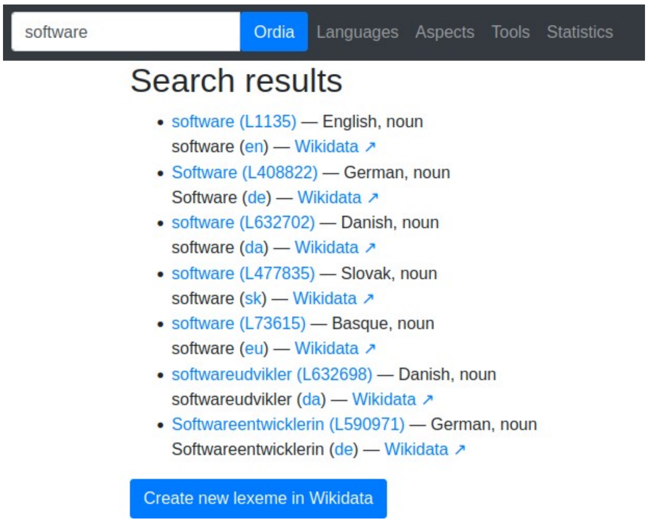
Senses

L1135-S1 English — non-tangible executable component of a computer

Statements about L1135-S1

item for this sense: software — 0 references

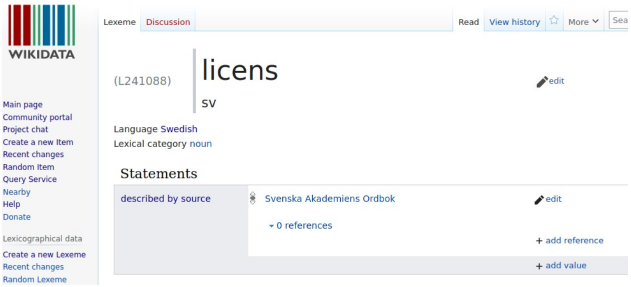
Figure 7: Lexeme L1135 ‘software’ in Wikidata



Search results

- software (L1135) — English, noun
  software (en) — Wikidata ↗
- Software (L408822) — German, noun
  Software (de) — Wikidata ↗
- software (L632702) — Danish, noun
  software (da) — Wikidata ↗
- software (L477835) — Slovak, noun
  software (sk) — Wikidata ↗
- software (L73615) — Basque, noun
  software (eu) — Wikidata ↗
- softwareudvikler (L632698) — Danish, noun
  softwareudvikler (da) — Wikidata ↗
- Softwareentwicklerin (L590971) — German, noun
  Softwareentwicklerin (de) — Wikidata ↗

Figure 8: Screenshot from the Ordia application showing a search for ‘software’ in Wikidata’s L namespace.



(L241088) licens
sv

Language Swedish
Lexical category noun

Statements

described by source: Svenska Akademiens Ordbok — 0 references

Figure 9: The Swedish noun ‘licens’ is connected to Sven- ska Akademiens Ordbok using the property ‘described by source’ in Wikidata’s Lexeme namespace.

The disadvantage of sourcing labels from this subset is that there are not as many editors contributing edits to the Lexeme namespace in Wikidata as there are editors who contribute to other namespaces.

In Table 4 we see the count of labels available in the Wikidata Lexemes subgraph. The terms ‘software’ and ‘license’ currently have the largest number of lexeme senses that have been connected to their Wikidata items. As encouragement for more Wikidata editors to familiarize themselves with the Lexeme namespace, people organize weekly challenges with a topical focus. For example, one recent lexeme challenge was focused on software[5] and another on computing[6]. These challenges are announced via Wikidata-related communication channels. We anticipate that as more editors learn about the lexeme namespace this subgraph is likely to increase in size.

Term	Count Wikidata Lexeme Senses
file format	1
checksum	0
operating system	3
data integrity	0
software	16
license	23
reproducibility	1

Table 4: Count of Labels from Wikidata Lexeme Sub- graph

VI. Discussion

The multilingual labels available via the Wikidata Query Service could be of value to people who are looking to source translations for terms in an ontology, vocabulary, glossary or for text in the user-interface of an application. While the number of available labels varies across terms, the open licensing of the data and the accessibility of the data via the Wikidata Query Service make this an attractive cost-free alternative to hiring translators for groups with limited budgets.

Looking at the count of labels available for the terms in our sample data set it is clear that editors have added more labels for ‘software’ and ‘operating system’ than the other terms. This is likely due to the high levels of awareness many editors have for these terms. The other terms in the sample data set are more specialized, and thus may be less familiar to editors. To date, editors have added the fewest number of labels for the terms ‘reproducibility’ and ‘data integrity’. Fewer editors may be familiar with these terms, or have use cases that would lead them to edit these items.

The webapp[7] we created to complement this paper allows viewers to see each label in the context of the set of supported languages. Not only can you get a sense of how many labels are available per ter for each subset, it is also possible to see each label if you hover over the colored language blocks in the webapp.

Members of the Wikidata community are motivated to contribute for many different reasons. There is no group or individual dictating how others should con- tribute to the project [24]. Different subsets of Wiki- data have different numbers of labels for the terms in our sample data set because there is no coordination of how work is accomplished, other than ad hoc decisions among editors to collaborate. This is consistent with the theoretical work describing peer-productions systems [25], [26].

As more people with digital preservation expertise decide to become editors of projects of the Wikimedia Foundation, it is possible that editors from our international community of practice could contribute more la- bels in additional languages to items, properties, and lexeme senses to Wikidata or contribute new articles in additional language versions of Wikipedia. Such contributions would benefit anyone interested in reusing data from Wikidata or Wikipedia. Guidance related to contributing to Wikidata tailored to the digital preservation community is described in [27]. Leveraging the infrastructure of the projects of the Wikimedia Foundation for collaboration is a strategy for that supports users from many different language contexts to benefit [28].

VII. Conclusion

People who create or maintain vocabularies, ontologies, applications or other projects may require multilingual labels for concepts in their systems. The cost of paying for translations into multiple languages can quickly add up, and may be beyond the budgetary constraints of many projects. Not only is there a multilingual knowledge graph that is free to reuse, exploring the multilingual data in the projects of the Wikimedia Foundation is an approachable task using the Wikidata Query Service. The Wikidata Query Service provides multiple options for downloading results in formats such as

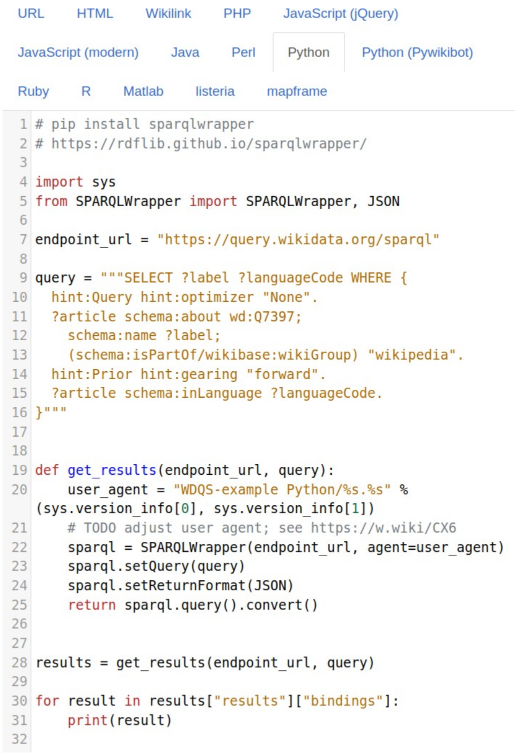
[5] https://dicare.toolforge.org/lexemes/challenge.php?id=52

[6] https://dicare.toolforge.org/lexemes/challenge.php?id=28

[7] The webapp that includes the interactive table is available at https://wikidp-research.k2.services/multi-lingual-table.

CSV, JSON, or HTML, they also provide code snipits for reusing queries within external applications, as seen in Figure 10. Once a subset has been identified, either through SPARQL queries or a ShEx schema in the Entity Schema namespace, results may be consulted again at a later point to determine if additional data is added by the Wikidata community over time. Subsets can be reused in other applications, to enrich ontologies, vocabularies, or within other resources where multilingual labels are needed.

Figure 10: Python code snipit available from the Wiki- data Query Service



```
# pip install sparqlwrapper
# https://rdflib.github.io/sparqlwrapper/

import sys
from SPARQLWrapper import SPARQLWrapper, JSON

endpoint_url = "https://query.wikidata.org/sparql"

query = """SELECT ?label ?languageCode WHERE {
  hint:Query hint:optimizer "None".
  ?article schema:about wd:Q7397;
    schema:name ?label;
    (schema:isPartOf/wikibase:wikiGroup) "wikipedia".
  hint:Prior hint:gearing "forward".
  ?article schema:inLanguage ?languageCode.
}"""


def get_results(endpoint_url, query):
    user_agent = "WDQS-example Python/%s.%s" %
(sys.version_info[0], sys.version_info[1])
    # TODO adjust user agent; see https://w.wiki/CX6
    sparql = SPARQLWrapper(endpoint_url, agent=user_agent)
    sparql.setQuery(query)
    sparql.setReturnFormat(JSON)
    return sparql.query().convert()


results = get_results(endpoint_url, query)

for result in results["results"]["bindings"]:
    print(result)

```

Sourcing labels from multiple subsets of Wikidata increases the breadth of languages that can be covered. People who are committed to holding themselves ac- countable to the values of accessibility, inclusion, and diversity may want to consider sourcing multilingual labels for resources in the domain of digital preservation from the projects of the Wikimedia Foundation. Some members of the digital preservation community may wish to contribute labels in their own languages for these terms, or for any other items or properties in Wikidata related to digital preservation, in order to improve and extend the knowledge graph.

We offer the techniques described in this paper for identifying potential subsets of multilingual data as strategies that others in the digital preservation community may find helpful. Investigating the multilingual label inventory from projects of the Wikimedia Foundation via the Wikidata Query Service could reduce or eliminate the need to source multilingual translations from other, more expensive, sources. As we are all in this together, let's support one another in our shared goals of increasing multilingual access points in projects and tools used by the digital preservation community.

VIII. ACKNOWLEDGEMENTS

We thank the communities of contributors to the projects of the Wikimedia Foundation. Contributions to projects of the Wikimedia Foundation made this re- search possible.

REFERENCES

[1] D. Vrande?i?, "Wikidata: A new platform for collab- orative data collection," in Proceedings of the 21st International Conference Companion on World Wide Web, ACM, 2012, pp. 1063-1064.

[2] W. Community, Wikidata:data model, 2023. [On- line]. Available: https : / / www . wikidata . org / wiki/Wikidata:Data_model.

[3] L.-A. Kaffee, A. Piscopo, P. Vougiouklis, E. Sim- perl, L. Carr, and L. Pintscher, "A Glimpse into Babel: An Analysis of Multilinguality in Wikidata," in Proceedings of the 13th International Symposium on Open Collaboration, ser. OpenSym '17, Galway, Ireland: ACM, 2017, 14:1-14:5, isbn: 978-1-4503- 5187-4. doi: 10.1145/3125433.3125465. [Online]. Available: https://doi.org/10.1145/3125433. 3125465.

[4] A. Hogan, E. Blomqvist, M. Cochez, et al., "Knowl- edge graphs," Synthesis Lectures on Data, Seman- tics, and Knowledge, vol. 12, no. 2, pp. 1-257, 2021.

[5] L. Jarnac and P. Monnin, "Wikidata to bootstrap an enterprise knowledge graph: How to stay on topic?" In Proceedings of the 3rd Wikidata Workshop 2022. [Online]. Available: https://ceur-ws.org/ Vol-3262/paper16.pdf.

[6] T. Pellissier Tanon, M. D. de Assunção, E. Caron, and F. M. Suchanek, "Demoing platypus-a multilin- gual question answering platform for wikidata," in The Semantic Web: ESWC 2018 Satellite Events: ESWC 2018 Satellite Events, Heraklion, Crete, Greece, June 3- 7, 2018, Revised Selected Papers 15, Springer, 2018, pp. 111-116.

[7] L.-A. Kaffee, H. ElSahar, P. Vougiouklis, et al., "Mind the (language) gap: Generation of multilingual wikipedia summaries from wikidata for article- placeholders," in The Semantic Web: 15th Inter- national Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings 15, Springer, 2018, pp. 319-334.

[8] J. Gracia, E. Montiel-Ponsoda, P. Cimiano, A. Gómez-Pérez, P. Buitelaar, and J. McCrae, "Chal- lenges for the multilingual web of data," Journal of Web Semantics, vol. 11, pp. 63-71, 2012.

[9] M. Bremer-Laamanen and J. Stenvall, "Selection for digital preservation: Dilemmas and issues," in Managing preservation for libraries and archives, Routledge, 2018, pp. 53-65.

[10] T. Evens and L. Hauttekeete, "Challenges of digi- tal preservation for cultural heritage institutions," Journal of

Librarianship and Information Science, vol. 43, no. 3, pp. 157-165, 2011.

[11] J. E. L. Gayo, Creating knowledge graphs subsets us- ing shape expressions, 2021. arXiv: 2110 . 11709 [cs.DB].

[12] J. E. Labra-Gayo, A. C. G. Cavazos, A. Waagmeester, et al., "Enhancement and reusage of biomedical knowledge graph subsets," 2022.

[13] P. Nguyen and H. Takeda, "Wikidata-lite for knowl- edge extraction and exploration," arXiv preprint arXiv:2211.05416, 2022. [Online]. Available: https://arxiv.org/pdf/2211.05416.

[14] M. Espinoza, A. Gómez-Pérez, and E. Mena, "En- riching an ontology with multilingual information," in The Semantic Web: Research and Applications: 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1-5, 2008 Pro- ceedings 5, Springer, 2008, pp. 333-347.

[15] S. Niininen, S. Nykyri, and O. Suominen, "The fu- ture of metadata: Open, linked, and multilingual- the yso case," Journal of Documentation, 2017.

[16] A. Conde, A. Arruarte, M. Larrañaga, and J. A. Elor- riaga, "How can wikipedia be used to support the process of automatically building multilingual do- main modules? a case study.," Information Process- ing & Management, vol. 57, no. 4, p. 102 232, 2020.

[17] S. Malyshev, M. Krötzsch, L. González, J. Gonsior, and A. Bielefeldt, "Getting the most out of wikidata: Semantic technology usage in wikipedia's knowl- edge graph," in International Semantic Web Confer- ence, Springer, 2018, pp. 376-394.

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

[19] T. Pellissier Tanon and 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.

[20] W. community, Wikidata:lexicographical data/documentation, Online; accessed 9- March-2023, 2023. [Online]. Available: https :/ / www . wikidata . org / wiki / Wikidata : Lexicographical_data/Documentation.

[21] F. Nielsen, "Lexemes in wikidata: 2020 status," in Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020), 2020, pp. 82-86.

[22] F. Å. Nielsen, "Ordia: A web application for wiki- data lexemes," in European Semantic Web Confer- ence, Springer, 2019, pp. 141-146.

[23] Ordia, Ordia statistics, [Online; accessed 13- October-2022], 2022. [Online]. Available: {https://ordia.toolforge.org/statistics/}.

[24] C. Müller-Birn, B. Karran, J. Lehmann, and M. Luczak-Rösch, "Peer-production system or collab- orative ontology engineering effort: What is wiki- data?" In Proceedings of the 11th International Symposium on Open Collaboration, ACM, 2015, p. 20.

[25] Y. Benkler, "Coase's penguin, or, linux and the nature of the ?rm," Yale Law Journal, pp. 369-446, 2002.

[26] Y. Benkler, A. Shaw, and B. M. Hill, "Peer production: A modality of collective intelligence," Collective Intelligence, 2013.

[27] K. Thornton, "Wikidata for Digital Preservation- ists," en, Dec. 2, 2021, 9 pp. doi: 10.7207/TWGN21-19. [Online]. Available: http://dx.doi.org/10. 7207/twgn21-19.

[28] J. Samuel, "Towards understanding and improving multilingual collaborative ontology development in wikidata," in Companion of the The Web Confer- ence 2018 on The Web Conference, 2018, pp. 23-27