Ordia: A Web application for Wikidata lexemes

Finn Årup Nielsen

Cognitive Systems, DTU Compute, Technical University of Denmark

Abstract. Since 2018, Wikidata has had the ability to describe lexemes, and the associated SPARQL endpoint *Wikidata Query Service* can query this information and visualize the results. *Ordia* is a Web application that displays the multilingual lexeme data of Wikidata based on embedding of the responses from the Wikidata Query Service via templated SPARQL queries. Ordia has also a SPARQL-based approach for online matching of the words of a text with Wikidata lexemes and the ability to use a knowledge graph embedding as part of a SPARQL query. Ordia is available from `https://tools.wmflabs.org/ordia/`.[1]

1 Introduction

The multilingual collaboratively editable and freely-licensed knowledge base *Wikidata*[2] [7] was set up in October 2012. On this website users can describe items and links between the items via properties, as well as add qualifiers and sources to support the individual claims. The Wikidata data—originally formatted in a nested JSON-like structure—is translated to a Semantic Web representation and continuously updated and made available via a SPARQL endpoint: The *Wikidata Query Service* (WDQS)[3] and as such part of the Linked Open Data cloud.

In 2016, the Wikidata developers announced dictionary support in Wikidata [4], and in May 2018, Wikidata enabled the entering of basic data about *lexemes* and their *forms*. Later that year, Wikidata also switched on support for *senses*, and links to the Q-items from senses[4] can be established. As the rest of Wikidata, the lexeme part of Wikidata is multilingual and ontological definitions in one language are available in other languages.

Below I will describe the *Ordia* Web application that takes advantage of the Wikidata lexeme data, aggregating the information via WDQS and presenting it on a website with added functionality in the form of lexeme extraction from a text and SPARQL integration of knowledge graph embedding information.

[1] This work is funded by the Innovation Fund Denmark through the projects DAnish Center for Big Data Analytics driven Innovation (DABAI) and Teaching platform for developing and automatically tracking early stage literacy skills (ATEL).

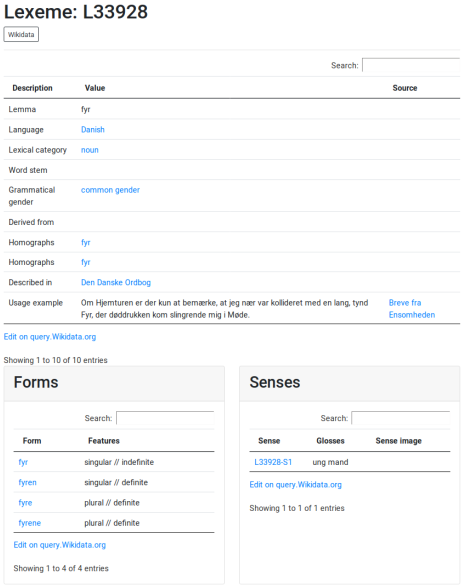
[2] `https://www.wikidata.org`.

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

[4] The "ordinary" Wikidata items are referred to by an identifier consisting of the letter 'Q' and an integer, while the properties are identified by the letter 'P' and an integer. Lexemes are identified by the initial letter 'L'.

2 Ordia Web application

Ordia is available from GitHub at http://github.com/fnielsen/ordia developed under the Apache 2.0 licens. It may be cloned from that repository and run locally. The canonical homepage for the Web application is at https://tools.wmflabs.org/ordia/ under the *Toolforge* cloud service provided by the Wikimedia Foundation.



Lexeme: L33928

Wikidata

Search:

Description	Value	Source
Lemma	fyr	
Language	Danish	
Lexical category	noun	
Word stem		
Grammatical gender	common gender	
Derived from		
Homographs	fyr	
Homographs	fyr	
Described in	Den Danske Ordbog	
Usage example	Om Hjemturen er der kun at bemærke, at jeg nær var kollideret med en lang, tynd Fyr, der døddrukken kom slingrende mig i Møde.	Breve fra Ensomheden

Edit on query.Wikidata.org

Showing 1 to 10 of 10 entries

Forms

Search:

Form	Features
fyr	singular // indefinite
fyren	singular // definite
fyre	plural // definite
fyrene	plural // definite

Edit on query.Wikidata.org

Showing 1 to 4 of 4 entries

Senses

Search:

Sense	Glosses	Sense image
L33928-S1	ung mand	

Edit on query.Wikidata.org

Showing 1 to 1 of 1 entries

Fig. 1: Screenshot of the page in Ordia for the Danish lexeme *fyr* at https://tools.wmflabs.org/ordia/L33928.

As a Web application and Python package, Ordia is heavily inspired from our other current Wikidata Web application projects: Scholia [3], cvrminer[5] and Wembedder [2]: Ordia uses the Flask web framework together with Javascript and SPARQL templates in Jinja[6] to dynamically build webpages. The constructed SPARQL queries are sent to the WDQS SPARQL endpoint and the responses are added to the generated HTML pages, either with HTML embedding or via Javascript and the *DataTables* library.[7] The library provides means for sorting table rows and for drill down via a search field. The SPARQL queries used to generate the tables in Ordia are all linked from an anchor in the lower left corner of the tables, making a SPARQL-knowledgeable user able to inspect and modify the queries.

Ordia creates separate pages for Q-items, lexemes, forms and senses, and makes panels with tables on each of them. Fig. 1 shows an example for a lexeme. Ordia uses a URL scheme for Q-items inspired from Scholia's notion of *aspects* and shows aspects for language, lexical category, grammatical features, propeties and references, e.g., the link /language/Q809 will show Polish (Q809) lexemes, while /Q809 shows Polish as a semantic concept in its own right. Some of the aspects show graphs for the ontology, e.g., the page for noun as a lexical category at /lexical-category/Q1084

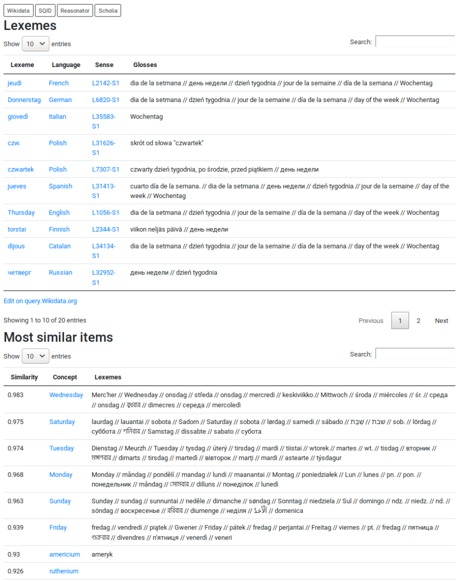
For searching after lexemes and forms, Ordia uses the MediaWiki API of Wikidata: The user types in a search in the Ordia interface and Ordia makes an API call to Wikidata and presents the results in the Ordia interface.

[5] Descriptions of cvrminer at https://tools.wmflabs.org/cvrminer/ has not been published. The Web application displays information about organization as listed in Wikidata.

[6] http://flask.pocoo.org/ and http://jinja.pocoo.org/

[7] https://datatables.net/.

Wembedder, a Wikidata-based knowledge graph embedding Web service [2], works with a simplified RDF2Vec approach implemented with Gensim's word2vec model [5,8,1]. As Ordia, Wembedder runs as part of Toolforge.[8] The current implementation only handles the Q-items and properties of Wikidata, — not lexemes, forms nor senses. The only functionality implemented in the Wembedder Web service so far is a *most similar* service that returns the most similar items and properties based on a query item. Wembedder has no SPARQL endpoint capability, so federated SPARQL queries cannot be made. Instead Ordia calls the REST interface of Wembedder via a Javascript Ajax call from the server side and formats the received JSON with Wikidata identifier and similarity values as two-tuple values for the SPARQL VALUES construct. The VALUES construct is then interpolated into a SPARQL template and sent off to WDQS with the response formatted in Ordia in a table with the DataTable library.



Lexemes

Show 10 entries Search:

Lexeme	Language	Sense	Glosses
jeudi	French	L2142-S1	dia de la setmana // день недели // dzień tygodnia // jour de la semaine // día de la semana // Wochentag
Donnerstag	German	L6820-S1	dia de la setmana // dzień tygodnia // jour de la semaine // día de la semana // day of the week // Wochentag
giovedì	Italian	L35583-S1	Wochentag
czw.	Polish	L31626-S1	skrót od słowa "czwartek"
czwartek	Polish	L7307-S1	czwarty dzień tygodnia, po środzie, przed piątkiem // день недели
jueves	Spanish	L31413-S1	cuarto día de la semana. // dia de la setmana // день недели // dzień tygodnia // jour de la semaine // day of the week // Wochentag
Thursday	English	L1056-S1	dia de la setmana // dzień tygodnia // jour de la semaine // día de la semana // day of the week // Wochentag
torstai	Finnish	L2344-S1	viikon neljäs päivä // день недели
dijous	Catalan	L34134-S1	dia de la setmana // dzień tygodnia // jour de la semaine // día de la semana // day of the week // Wochentag
четверг	Russian	L32952-S1	день недели // dzień tygodnia

Edit on query.Wikidata.org

Showing 1 to 10 of 20 entries

Most similar items

Show 10 entries Search:

Similarity	Concept	Lexemes
0.983	Wednesday	Merc'her // Wednesday // onsdag // středa // onsdag // mercredi // keskiviikko // Mittwoch // środa // miércoles // śr. // среда // onsdag // বুধবার // dimecres // середа // mercoledì
0.975	Saturday	laurdag // lauantai // sobota // Sadorn // Saturday // sobota // lørdag // samedi // sábado // שַׁבָּת // שבת // sob. // lördag // суббота // শনিবার // Samstag // dissabte // sabato // субота
0.974	Tuesday	Dienstag // Meurzh // Tuesday // tysdag // úterý // tirsdag // mardi // tiistai // wtorek // martes // wt. // tisdag // вторник // মঙ্গলবার // dimarts // tirsdag // martedì // вівторок // marți // mardi // astearte // týsdagur
0.968	Monday	Monday // måndag // pondělí // mandag // lundi // maanantai // Montag // poniedziałek // Lun // lunes // pn. // pon. // понедельник // måndag // সোমবার // dilluns // понеділок // lunedì
0.963	Sunday	Sunday // sundag // sunnuntai // neděle // dimanche // søndag // Sonntag // niedziela // Sul // domingo // ndz. // niedz. // nd. // söndag // воскресенье // রবিবার // diumenge // неділя // ٱلْأَحَد // domenica
0.939	Friday	fredag // vendredi // piątek // Gwener // Friday // pátek // fredag // perjantai // Freitag // viernes // pt. // fredag // пятница // শুক্রবার // divendres // п'ятниця // venerdì // veneri
0.93	americium	ameryk
0.926	ruthenium	

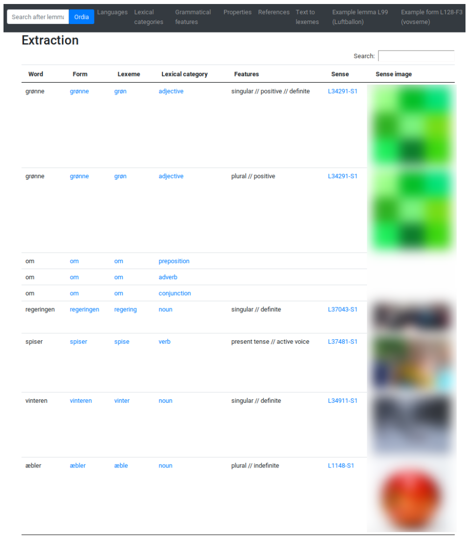
Fig. 2: Screenshot of Ordia's page for the *Thursday* Wikidata concept (Q129). The top panel shows the associated lexemes and senses in the languages that link to the concept, while the lower panel displays the result from the Wembedder knowledge graph embedding similarity computation.

Ordia uses the Wembedder queries on pages for Q-items, where a table displays related Q-items sorted according to similarity and augmented with information from the lexeme part of Wikidata. Fig. 2 shows an example of the output on the page for the concept *Thursday* corresponding to the page https://tools.wmflabs.org/ordia/Q129, where the top panel displays lexemes for languages linked (e.g., jeudi, Donnerstag, Thursday) and the lower panel shows the result from WDQS with Wembedder-included results. Here Wednesday and Saturday are the most related concepts to Thursday.

The **text-to-lexeme** facility in Ordia at https://tools.wmflabs.org/ordia/text-to-lexemes enables the user to write a short text into an HTML text area on the client side, and send it off to Ordia. Ordia then makes a simple sentence detection and lowercases the first letter of the sentences before word tokenization with a simple regular expression pattern. Identified words are interpolated into a WDQS query via the VALUES keyword to search for matching forms, and the response from the SPARQL endpoint is shown in a table in the Ordia interface. The

[8] https://tools.wmflabs.org/wembedder/

language of the input sentence must be specified. Currently, Ordia only handles a small number of selected languages, but in principle every language in Wikidata lexemes could be supported.



Extraction

Word	Form	Lexeme	Lexical category	Features	Sense	Sense image
grønne	grønne	grøn	adjective	singular // positive // definite	L34291-S1	
grønne	grønne	grøn	adjective	plural // positive	L34291-S1	
om	om	om	preposition			
om	om	om	adverb			
om	om	om	conjunction			
regeringen	regeringen	regering	noun	singular // definite	L37043-S1	
spiser	spiser	spise	verb	present tense // active voice	L37481-S1	
vinteren	vinteren	vinter	noun	singular // definite	L34911-S1	
æbler	æbler	æble	noun	plural // indefinite	L1148-S1	

Fig. 3: Screenshot of Ordia's text-to-lexeme facility, where the sentence "Regeringen spiser grønne æbler om vinteren" ("The government is eating green apples during winter"): https://tools.wmflabs.org/ordia/text-to-lexemes?text=Regeringen+spiser+gr%C3%B8nne+%C3%A6bler+om+vinteren&language=da.

Fig. 3 displays the result after the Danish sentence "Regeringen spiser grønne æbler om vinteren" ("The government is eating green apples during winter") has been submitted to Ordia. The result of the WDQS query here shows the word and—if matched—the form, lexeme, lexical category, lexical feature, sense and image associated with the sense. If a word matches several forms in a language, they are all shown, i.e., no word sense disambiguation is performed. Ordia's text-to-lexemes responds within seconds. Usually Ordia responds with the HTML within 300 milliseconds for a sentence like the above. The SPARQL query sent by the client to WDQS completes typically between 1.5 and 2 seconds after the user submitted the original query. The further download and rendering of the images from Wikimedia Commons—as shown in Fig. 3—may take an extra second. If the SPARQL query does not find a matching form, a link is created to Ordia's search page, which links further on to lexeme creation to ease the setup of new lexemes.

3 Discussion

I have shown Ordia, a Web service that uses the WDQS SPARQL endpoint to build a site with lexicographic data from Wikidata. Compared to the Wiktionary-based DBnary [6], Ordia needs no extractor and presents the lexicographic information graphically and up-to-date via WDQS as changes occur in Wikidata.

The conceptual choices that has been made in designing Ordia are: 1) A user should easily be able to perform powerful SPARQL queries by navigating the Ordia interface and without using any knowledge of SPARQL; 2) the URL pattern for each page should be easy to understand and predict; 3) Each page should link to other pages and in such a way let the user discover new lexemes, concepts, forms etc., and 4) the interface should use graphics when possible, e.g., graphs of word and concept relations and for displaying images associated with senses of words.

Other Wikidata lexeme Web applications beyond Ordia are available: Lucas Werkmeister has created *Wikidata Lexeme Forms* which enables easy HTML-form-based set up of lexemes and their lexical forms for a range of languages. In November 2018, Werkmeister found that this tool has been used for the creation of 10'827 lexemes out of a total of 37'886.[9] Alicia Fagerving has created *Wikidata Senses* which eases the setup of senses associated with lexemes. While the above tools focuses on input, Léa Lacroix' *DerDieDas* game, tasks a language learner to guess and learn the grammatical gender of presented nouns. It uses the data in Wikidata via a WDQS query. Originally in German, it now has derived versions in French and Danish. Another of Werkmeister's online tools, *Wikidata Lexeme Graph Builder*, constructs a graph based on a specified Wikidata item and a Wikidata property. These and further Wikidata lexicographical tools are listed at https://www.wikidata.org/wiki/Wikidata:Tools/Lexicographical_data.

Ordia can be used in a variety of ways: A copy-and-paste of a text into Ordia's text-to-lexeme tool will quickly return an overview of missing lexeme data in Wikidata. Most words from a typical English news article are usually matched to a Wikidata lexeme, — except for proper nouns. Another useful overview that Ordia gives is the ontology of lexical categories. For instance, https://tools.wmflabs.org/ordia/lexical-category/Q36224 shows a graph with subconcepts and superconcepts of the pronoun concept independent of language. Such an overview is convenient to consult when entering lexeme data in Wikidata. The use of Ordia as, e.g., a translation or synonymy dictionary is still constrained by the yet low number of lexemes that have been entered and linked.

References

1. Mikolov, T., Chen, K., Corrado, G.S., Dean, J.: Efficient Estimation of Word Representations in Vector Space (January 2013), https://arxiv.org/pdf/1301.3781v3
2. Nielsen, F.Å.: Wembedder: Wikidata entity embedding web service (October 2017), https://arxiv.org/pdf/1710.04099
3. Nielsen, F.Å., Mietchen, D., Willighagen, E.: Scholia, Scientometrics and Wikidata. The Semantic Web: ESWC 2017 Satellite Events (October 2017).
4. Pintscher, L.: Let's move forward with support for Wiktionary. Wikidata mailing list (September 2016), https://lists.wikimedia.org/pipermail/wikidata/2016-September/009541.html
5. Ristoski, P., Paulheim, H.: RDF2Vec: RDF Graph Embeddings for Data Mining. The Semantic Web – ISWC 2016 pp. 498–514 (2016)
6. Sérasset, G.: DBnary: Wiktionary as a Lemon-Based Multilingual Lexical Resource in RDF. Semantic Web: interoperability, usability, applicability (2014).
7. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57, 78–85 (October 2014).
8. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. New Challenges For NLP Frameworks Programme pp. 45–50 (May 2010).

[9] https://quarry.wmflabs.org/query/28791