Validating Danish Wikidata lexemes

Finn Årup Nielsen[1], Katherine Thornton[2], and Jose Emilio Labra Gayo[3]

[1] Cognitive Systems, DTU Compute, Technical University of Denmark, Denmark
faan@dtu.dk
[2] Yale University Library, New Haven, CT, USA
katherine.thornton@yale.edu
[3] University of Oviedo, Spain
labra@uniovi.es

Abstract. Two of the newest features of Wikidata are support for lexicographic data (lexemes), and support for Shape Expressions (ShEx). We demonstrate the first application of ShEx for validation of entity data for Wikidata lexemes. Validation of entity data in Wikidata against ShEx schemas allows editors to discover missing or incorrect information. It may also form a basis for discussion of the data models implicitly used in Wikidata. We present a use case and benchmark for ShEx and discuss its current limitations.

1 Introduction

Since 2018, it has been possible to represent lexeme data with associated information about forms and senses in Wikidata [10]. Statistics from June 2019, indicate that more than 47,000 lexemes for more than 330 human languages have already been added to Wikidata. For Danish alone, there are entries for over 1,700 lexemes, including nouns, verbs, adjectives, adverbs and words from several other word classes.[4] These lexemes can be described by properties specifying forms, senses, language, lexical categories, grammatical features, hyphenation, etc. We may also link lexemes to external linguistic resources, e.g., DanNet [7] is a wordnet for Danish. Words found in version 2.2 of DanNet have an associated Wikidata property. Thus, each lexeme has associated structured data suitable for machine consumption.

ShEx (Shape Expressions) is a concise, formal language for modeling and validating RDF graphs [8]. The ShEx compact syntax allows users to rapidly write schemas to capture data models as they evolve. ShEx is actively being used to validate data in Wikidata [4]. In May 2019, Wikidata enabled a new namespace, which allows Wikidata editors to store and collaboratively edit ShEx schemas. Editors can subsequently use these schemas for validation of Wikidata items.

[4] Updated statistics are available from the Ordia tool: https://tools.wmflabs.org/ordia/statistics. Language statistics are available at https://tools.wmflabs.org/ordia/language/.

Here we describe our initial experiences with using ShEx for validation of linguistic data on Wikidata, focusing on Danish Wikidata lexemes. This is a convenient method for a user to quickly gain an overview of the current status of the data with regard to a specific schema, entirely using tools in the Wikidata ecosystem. Apart from discovering missing or incorrect data, we identify cases where ShEx pinpoints issues for discussion about a 'data model' for the Danish lexicographic data, as well as cases where ShEx validation is difficult.

2 Related work

Numerous mechanisms and tools are available that allow Wikidata users to validate entered lexeme data [4]. Some of the mechanisms that work across Wikidata entities are: 1) Datatype restrictions per-property: The definition of properties specifies which datatype is allowed in the value field, such that, e.g., a free-form text (literal) cannot be used where a Wikidata item ('IRI') is required. 2) Literal value restrictions: Wikidata also sets up constraints on literal data values via regular expressions defined in the P1793 Wikidata property. 3) Property constraints: Each property may also be associated with constraints that provide hints to editors about what values are expected, and what values are not. 4) Identifying patterns using SPARQL: With the SPARQL-based *Wikidata Query Service* (WDQS), users can formulate queries that can show inconsistency in the lexeme data, e.g., 'find every lexeme without any form' as suggested on the *Ideas of queries* wiki page.[5]

Scholia, a SPARQL-based web application and Wikidata frontend focusing on scholarly data [6], dynamically creates so-called subaspects with information about (possibly) missing data in Wikidata. For instance, the 'missing' subaspect for an author displays author name strings that could be resolved to Wikidata items and authored publications that are missing specification of one or more main subjects.

Ordia is a web application for Wikidata lexemes [5]. Currently it has no validation, but browsing its various dynamically generated web pages a user may discover unusual patterns, e.g., the web page `https://tools.wmflabs.org/ordia/lexical-category/` displays values used as lexical categories for lexemes where rare categories can contain errors. An example is L45350 which has the lexical category *centralized version control system*, — an obvious error.

In addition to [9], ShEx and Wikidata have also been described in connection with schema inference [11]. The Wikidata ShEx Inference tool can be used to automatically suggest a schema for a type of resource based on properties that have been used on similar items in Wikidata.

This is the first application of ShEx validation to the lexeme namespace in Wikidata. Validation using ShEx provides more comprehensive validation than has previously been possible with Wikidata tools. Datatype restrictions and literal value restrictions can be expressed in ShEx and identified alongside property

[5] `https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Ideas_of_queries`.

constraints for multiple properties at the same time. This offers more complete information, by allowing users to test situations that previously could only be approached piece by piece according to the above rules. The constraint system is useful for communicating expectations and potential issues to users on a per-property basis. Validating entity data against a schema describing a data shape can span multiple properties.

3 Validation of Danish lexemes

We wrote ShEx schemas for Danish lexemes with the identifiers E15 (Danish lexeme), E34 (Danish noun), E54 (lexeme), E56 (Danish verb), E62 (Danish pronoun) and E65 (Danish numeral) as well as a ShEx for Danish hyphenation E68.[6]

Many rules exist for Danish words and grammar, see, e.g., [1,3], and we coded several in ShEx wrt. forms, senses, hyphenation, conjugation class, DanNet identifier, grammatical gender, etc. Here we will discuss a few:

1. All Danish Wikidata lexemes should have one unique value for DanNet words, — either one unique identifier or no value. Proper Danish nouns, adverbs, pronouns and words from a number of other word classes should *not* have an associated DanNet identifier.
2. A Danish noun should have one single grammatical gender, either common gender or neuter.
3. Each part of a hyphenated representation should contain a vowel.
4. The grammatical gender of a compound should have the same grammatical gender as the final lexeme of the compound, except for compounds suffixed *-fuld*.

For the first rule, related to DanNet identifiers, Wikidata has the ability to explicitly state 'no value' and ShEx can test for the presence of this 'no value' in the Wikidata property for DanNet (P6140) with "`a [ wdno:P6140 ]`". In the cases where the DanNet identifier is present with a single value, we can test its format with a regular expression. Current DanNet identifiers conform to a format with 8 digits. The combined ShEx constraint for Wikidata DanNet statements then reads `a [ wdno:P6140 ] | ps:P6140 /^[0-9]{8}$/`". A test for the uniqueness of DanNet identifiers is not currently possible in ShEx, but it is possible that this feature will be added [2].

The second rule for two grammatical genders is represented by the following ShEx schema: `wdt:P5185 [ wd:Q1305037 wd:Q1775461 ]`. Danish does not reveal grammatical gender in plural, so *plurale tantum* (word with only plural forms) and its derived lexemes have no obvious gender and no gender is specified in Wikidata. The same lack of obvious gender occurs for a word such as a *druk*

[6] The Wikidata page with editing capabilities are, e.g., `https://www.wikidata.org/wiki/EntitySchema:E15` while the URL for ShEx import is `https://www.wikidata.org/wiki/Special:EntitySchemaText/E68` for E15

and for proper nouns. Either we need to define the gender or modify the ShEx schema to accommodate the exceptions. We also note that a few nouns have multiple genders, and these nouns also require exceptions.

For the third rule, we note that a Danish hyphenation rule can apply across most, if not all, word classes, thus we can write a general schema for hyphenation and use ShEx's `IMPORT` feature to include shapes of the schema as part of another schema. The essential part in the current Danish hyphenation schema (E68) is: “`a [ wdno:P5279 ] | ps:P5279 /.*[aeiouyæøå] .*\u2027.*[aeiouyæøå].*/ ;`”. Here `wdno:P5279` catches Wikidata's ‘no value’ for forms without possible hyphenation. The regular expression uses the hyphenation point unicode character. It could be further refined and extended to accommodate accented vowels and words with multiple possible hyphenation points.

The fourth rule, related to the gender of compounds, requires a more elaborate ShEx schema. Our current implementation enumerates over grammatical gender and over the number of compound parts as well as creates exceptions for the *-fuld* suffix and words that are not compounds. The result is a verbose ShEx.

A ShEx schema is submitted to the ShEx2 Simple Online Validator[7] with a list of Wikidata items to be tested, — most conveniently found by a SPARQL query to WDQS. The conformance reports produced in the ShEx validation process provide detailed feedback about which statements on which items have issues. Testing a subgraph for conformance to the ShEx schema, users can survey multiple items in a single validation session. This allows users to address issues across a set of items rather than reviewing them individually. By reading these conformance reports we discovered numerous issues. Most of the non-conformant items we discover are errors of omission, rather than errors of commission. In the former case, e.g., nouns missing grammatical gender, while in the latter case, e.g., nouns with a wrong gender. We corrected a number of the erroneous entries and added many new statements.

4 Discussion

ShEx conformance reports can be used as a basis for concrete discussions among editors about approaches for improving the data. Communicating via a schema language allows editors to be explicit about their expectations and intentions.

We note that some of the rules may be contested by Wikidata users. For instance, there has been discussion on which character should indicate hyphenation and how Wikidata should indicate that a word form cannot be hyphenated.[8]

For people preparing to write schemas for use in the Wikidata context, we offer the following practical recommendations. While writing schemas is an investment of effort, the ability to quickly gain an overview of a subset of the Wikidata graph is useful. It may be possible to reuse shapes from existing schemas,

[7] https://tools.wmflabs.org/shex-simple/wikidata/packages/shex-webapp/doc/shex-simple.html.

[8] The hyphenation discussions have taken place on the so-called talk page of the property on Wikidata: https://www.wikidata.org/wiki/Property_talk:P5279.

thus it may be helpful to explore existing schemas. Reading existing schemas is also a helpful way to get a sense of how others have expressed data models that could serve as examples.

The opportunity to provide a link to a schema in the ShEx namespace of Wikidata supports collaborative refinement of data models, such as those for Danish lexemes. Thus, if some of the rules discussed in this paper require refinement, editors of Wikidata will be able to point to schemas, to query maps indicating relevant sub-graphs of Wikidata, and to individual lexeme items which may need to be revised.[9]

References

1. Allan, R., Holmes, P., Lundskær-Nielsen, T.: Danish. Routledge (1995)
2. Baker, T., Prud'hommeaux, E.: Shape Expressions (ShEx) Primer (July 2017), http://shex.io/shex-primer-20170713/
3. Hansen, E., Heltoft, L.: Grammatik over det Danske Sprog. University Press of Southern Denmark (February 2019)
4. Labra Gayo, J.E., Prud'hommeaux, E.G., Boneva, I., Kontokostas, D.: Validating RDF Data, vol. 7 (September 2017), http://book.validatingrdf.com/
5. Nielsen, F.Å.: Ordia: A Web application for Wikidata lexemes. In: ESWC 2019 Posters & Demos (May 2019), http://www2.compute.dtu.dk/pubdb/views/edoc_download.php/7137/pdf/imm7137.pdf
6. Nielsen, F.Å., Mietchen, D., Willighagen, E.: Scholia, Scientometrics and Wikidata. In: The Semantic Web: ESWC 2017 Satellite Events (October 2017), http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/7010/pdf/imm7010.pdf
7. Pedersen, B.S., Nimb, S., Asmussen, J., Sørensen, N.H., Trap-Jensen, L., Lorentzen, H.: DanNet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. Language Resources and Evaluation 43, 269–299 (August 2009)
8. Prud'hommeaux, E.G., Labra Gayo, J.E., Solbrig, H.: Shape expressions: an RDF validation and transformation language. SEM '14: Proceedings of the 10th International Conference on Semantic Systems (2014)
9. Thornton, K., Solbrig, H., Stupp, G., Labra Gayo, J.E., Mietchen, D., Prud'hommeaux, E.G., Waagmeester, A.: Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation. In: The Semantic Web. pp. 606–620 (May 2019)
10. Vrandečić, D., Krötzsch, M.: Wikidata: a free collaborative knowledgebase. Communications of the ACM 57, 78–85 (October 2014), http://cacm.acm.org/magazines/2014/10/178785-wikidata/fulltext
11. Werkmeister, L.: Schema Inference on Wikidata (October 2018), https://github.com/lucaswerkmeister/master-thesis/releases/download/final/master-thesis-Lucas-Werkmeister.pdf

[9] This work is funded by the Innovation Fund Denmark through DABAI and ATEL projects. The third author is partially funded by the Spanish Ministry of Economy and Competitiveness (Society challenges: TIN2017-88877-R). We appreciate feedback from Eric Prud'hommeaux on the ShEx schemas discussed in this paper.