SHACL4GW: SHACL Shapes for the Global Wordnet Association RDF Schema

Anas Fahad Khan
Istituto di Linguistica Computazionale
Consiglio Nazionale delle Ricerche
Pisa, Italy
fahad.khan@ilc.cnr.it

John P. McCrae
Insight Centre and ADAPT Centre
University of Galway
Galway, Ireland
john@mccr.ae

Abstract

In this article, we introduce SHACL Shapes for Global Wordnet RDF (SHACL4GW), a new resource which uses the Semantic Web SHACL standard for the validation of RDF files using the Global Wordnet Association RDF format. We begin by giving a motivation for the creation of such a resource, continue by describing the resource itself and end with our plans for future work.

1 Introduction

In the current article, we introduce a new resource for the validation of RDF wordnets produced using the Global Wordnet Association (GWA) RDF format; as we will see this resource, **SHACL Shapes for Global Wordnet RDF** (SHACL4GW), follows the well-known SHACL standard (Knublauch and Kontokostas, 2017) for validating RDF graphs. In what follows, we begin by introducing the Global Wordnet formats, this will give us a broad overall context for the current work; next, we will give an outline of the SHACL standard and describe why it is so useful, both in general and in the particular case of the GWA RDF format. We will also see how it relates to and complements the existing RDF schema for wordnets that has been made available by the Global Wordnet Association.

2 Global WordNet Formats

The Global Wordnet formats were proposed by the Global Wordnet Assocation (Vossen et al., 2016) and further extended by McCrae et al. (2021) in order to provide a common format for the inclusion of wordnets in the Collaborative Interlingual Index (Bond et al., 2016, CILI). The format supports three-plus serialization formats with the primary format being XML, based on the Kyoto-LMF model (Soria et al., 2009). In addition, the formats support serialization in JSON (with a JSON-LD schema) and an RDF data model that can be serialized in any RDF serialization format, including Turtle (Carothers and Prud'hommeaux, 2014). Note that, in the rest of the paper, we will refer to the serialization in RDF as GWA RDF (and the primary Kyoto-LMF format as GWA LMF). As with LMF and OntoLex-Lemon (Cimiano et al., 2016; McCrae et al., 2011), the main elements of the RDF data model are the *lexical entry* and the *synset* (equivalent to the lexical concept in OntoLex-Lemon). The model fully supports the relations used in Princeton WordNet (Miller, 1995; Fellbaum, 2010)as well as relations introduced by later projects such as EuroWordNet (Vossen, 2004). In addition, extra features such as pronunciation information used by resources such as Open English WordNet (McCrae et al., 2019) are also supported.

The following is a simple example of an XML entry in the GWA LMF format.

```
<LexicalEntry id="example-w1">
  <Lemma writtenForm="grandfather"
         partOfSpeech="n"/>
  <Sense id="example-sense1"
         synset="example-synset1"/>
</LexicalEntry>
...
<Synset id="example-synset1" ili="i90287"
        partOfSpeech="n"
        members="example-w1">
  <Definition>
    the father of your father or mother
  </Definition>
  <SynsetRelation relType="hypernym"
                  target="example-synset2"/>
</Synset>
```

Listing 1: Part of Speech Information

This describes a single entry, 'grandfather' which is linked to a synset with an associated definition. The members of the synset are given allowing their order to be specified. In addition, an ILI identifier is given to allow this resource to be included in the CILI (Bond et al., 2016). The synset is described with a definition and a hypernym link to another synset.

3 The Shapes Constraint Language

The **Shapes Constraint Language (SHACL)** is a W3C standard which provides a standard way of validating RDF graphs with respect to user-defined sets of constraints; such constraints, in SHACL parlance, are known as *shapes*. Thanks to its usability and flexibility SHACL has become an important component of the Semantic Web stack, complementing other well-known Semantic Web technologies such as RDF, RDFS and OWL. In this regard, it is worth noting that, in contrast with OWL and its adoption of the open world assumption, SHACL makes it simple to impose closed world constraints on RDF data – something which is often vital for the purposes of data validation. SHACL also allows for the generation of informative reports in the course of the validation of a graph which highlight and describe the violations of constraints and can also grade violations according to their seriousness (as determined by users themselves). The use of SHACL facilitates an extra level of integration and interoperability of RDF datasets in addition to that offered by other RDF technologies, standards and best practices – along with (not unrelatedly) helping to ensure a high level of data quality of RDF data. Moreover, as well as being very expressive, SHACL shapes are also reasonably simple to create, at least for those familiar with RDF syntax, thanks to the fact that they are defined using RDF triples.

The current work is novel for introducing the use of SHACL shapes in a linguistic linked data context. Although SHACL has been widely used for semantic data validation in other domains, with numerous online tutorials and tools available for working with the language[1], there are few (publicly available) resources that show the use (and usefulness) of SHACL in the context of linguistic linked data. The GWA RDF format presents an excellent case study for demonstrating the utility of SHACL for validating RDF language resources. This additional means of validating RDF wordnets provides an extra, much needed level of interoperability to such resources – over and above that offered by the OntoLex-Lemon ontology (on which the GWA RDF format is based) and the wordnet-specific RDF vocabulary `wn` made available by the Global Wordnet Association[2] – and thus helps to contribute to the growth of the Global Wordnet Grid[3]. Up until now only the DTD schema[4], made available by the GWA, has offered this functionality and only for the GWA LMF XML format; the use of SHACL shapes for GWA RDF will allow for the direct validation of RDF files (that is, without the need to first convert RDF graphs to the LMF XML format). Moreover, it does this by using standard Semantic Web technologies in a way that is easily shareable and can be easily built upon in the case of extensions to the GWA schema. The idea of the present work is both to argue for the *use* of SHACL shapes for validating GWA RDF graphs, as well to propose *a specific set of* SHACL shapes, which we describe below and which can be downloaded at the following link: `https://github.com/anasfkhan81/SHACL4GW`.

4 Creating SHACL shapes for GWA RDF

4.1 SHACL4GW

In the rest of this article, we will describe the SHACL "Shapes Graph" which we have developed for GWA RDF and which we refer to as SHACL4GW; this graph is available at `https://github.com/anasfkhan81/SHACL4GW`. It can be used to validate individual GWA RDF files via the excellent SHACL playground site[5]. In particular, we will explain some of the thinking behind the design decisions we have taken.

It is important to emphasise that the work we present here (SHACL4GW) is intended as a proposal to be shared and discussed with the wider wordnet community[6] with a view to gathering feedback and, if needed, modifying our proposal in collaboration with others. In a number of cases, we have left things open since we were not aware of there being a settled best practice for how to represent such cases in RDF (this is most notably the case with *LexicalResource*, see below), with the intention once again to open a discussion with the wider community as to what the best approach might be.

We began the process of putting together our

[1] SHACL is also the subject of a forthcoming book by Veronika Heimsbakk `https://veronahe.wordpress.com/shacl-for-the-practitioner/`

[2] `https://globalwordnet.github.io/schemas/wn#`

[3] `http://globalwordnet.org/resources/global-wordnet-grid/`

[4] `https://globalwordnet.github.io/schemas/WN-LMF-1.3.dtd`

[5] `https://shacl-playground.zazuko.com/`

[6] The Global Wordnet Conference is obviously an excellent venue for this.

SHACL graph by analysing the original DTD file for the GWA LMF format. Several of the declarations in the DTD could, it turned out, be easily converted into SHACL shapes using classes and properties from the wn vocabulary and the OntoLex vocabulary on which it is based. In other cases the conversion wasn't so straightforward, as we shall see. In general, the DTD was our primary guide to which elements should be obligatory and which to make optional. Our priority throughout was to maintain interoperability between formats, and indeed to make it even simpler to convert, and to 'roundtrip', between the different GWA formats (LMF XML, JSON, and RDF). In addition, we also sought to emphasise interoperability between wordnets in RDF without making the constraints overly restrictive.

4.2 Methodology

One fairly indicative example of the kinds of decisions we had to make in drafting our SHACL graph is given by cases in which we associate language metadata with individual URI resources. This is required (obligatory) in the case of the *Lexicon*, but implied (non-obligatory) in the case of *Definition*. Here we decided to limit the user to the choice of two linked data properties the `dc:language` property or the OntoLex lime metadata module property `lime:language` using the SHACL `sh:or` logical constraint. This choice allows a certain level of flexibility, since the DC property is very frequently used in general, but the *lime* property is commonly used in the context of OntoLex; at the same time this limitation helps to make GWA RDF graphs much more interoperable than otherwise.

```
sh:or ([
      sh:name "Language" ;
      sh:description "Ensure there is one
           single language assigned to the
           Wordnet, via DC:language" ;
      sh:path dc:language ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:nodeKind sh:IRIOrLiteral ;]
 [    sh:name "Language" ;
      sh:description "Ensure there is one
           single language assigned to the
           Wordnet, via lime:language" ;
      sh:path lime:language ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:nodeKind sh:Literal ;])
```

Listing 2: Use of logical sh:or constraint.

We now look at some of the main classes covered in SHACL4GW. In this first version of our graph, we decided not to create constraints corresponding to the *LexicalResource* declaration in the original DTD since there isn't a standard way of representing a Lexical Resource defined container for one or more lexicons in OntoLex[7]. A number of elements in the original DTD have the same set of Dublin Core metadata elements as potential attributes. Instead of adding these to individual shapes, we created a `MetadataElementShape` which is associated with individual classes via the property `sh:TargetClass`.

Lexicon

When it came to creating shapes for the Lexicon class, there were no major surprises (except possibly for the addition of a `sh:or` clause for language information as mentioned above) and the conversion from the DTD was fairly straightforward. We defined a `sh:NodeShape` called `LexiconShape` with target class `lime:Lexicon`, in addition to creating property shapes using the following properties and classes to add relevant constraints regarding label, email, license, version, URL, citation, status, note and confidence information: `rdfs:label`, `schema:email`, `cc:license`, `owl:versionInfo`, `wn:status`, `wn:note`, `wn:confidenceScore` respectively.

Lexical Entry

The creation of the shape corresponding to *LexicalEntry*, `LexicalEntryShape`, was, once again, fairly straightforward. We associate each *LexicalEntry* with exactly one lemma by making use of `ontolex:canonicalForm` and targeting the `FormShape` node (described elsewhere in the file) in order to ensure that this has the correct shape. Similarly, we make sure that senses have the correct shape via another property shape with path `ontolex:sense` and which targets the `SenseShape` node (which again is described elsewhere in the file). Part-of-speech information (obligatory for *LexicalEntry* elements) is described by the following shape:

```
    sh:property [
    sh:name "Part of Speech" ;
        sh:path wn:partOfSpeech  ;
        sh:minCount 1 ;
        sh:maxCount 1 ;
    sh:in (wn:noun wn:verb wn:adjective wn:adverb
         wn:adjective_satellite wn:named_entity
    wn:conjunction wn:adposition wn:other_pos wn:
        unknown_pos ) ;
    ]
```

Listing 3: Part of Speech Information

[7] One possible candidate for a class corresponding to LexicalResource could be the Data Catalog Vocabulary class *dataset*. However, it may also be that there is no need to explicitly cover this in our SHACL graph.

Note how SHACL allows us to guarantee that each Lexical Entry has exactly one part of speech as well as specifying what values this can have. Although we can encode similar information as axioms in OWL, it is complicated to use such axioms for the purposes of validation because of the Open World Assumption.

Senses and Sense Relations

For the *Sense* element, we were able to formulate SHACL constraints that corresponded fairly closely to almost all of the declarations in the DTD, aside, that is, from those referring to the lexicalized status of a *Sense* element and adjposition information (adjectival position); we couldn't find elements in standard pre-existing RDF vocabularies corresponding to these declarations[8]. For the rest, we were able to make use of OntoLex and `wn` vocabularies to determine our shapes. In order to ensure that *SenseRelation*s belonged to the list found in the `wn` vocabulary we used the `sh:in` property as follows:

```
    sh:targetClass vartrans:SenseRelation ;
    sh:property [
    sh:name "Category" ;
    sh:description "Make␣sure␣the␣Sense␣Relation␣
         belongs␣to␣the␣correct␣category" ;
        sh:path vartrans:category ;
        sh:minCount 1 ;
        sh:maxCount 1 ;
        sh:in (wn:antonym wn:also wn:participle
             wn:pertainym wn:derivation wn:
             domain_topic wn:has_domain_topic wn:
             domain_region wn:has_domain_region
             wn:exemplifies wn:is_exemplified_by
             wn:similar wn:other wn:
             simple_aspect_ip wn:
             secondary_aspect_ip wn:
             simple_aspect_pi wn:
             secondary_aspect_pi wn:feminine wn:
             has_feminine wn:masculine wn:
             has_masculine wn:young wn:has_young
             wn:diminutive wn:has_diminutive wn:
             augmentative wn:has_augmentative wn:
             anto_gradable wn:anto_simple wn:
             anto_converse)];
```

Listing 4: Part of Speech Information

Synsets and Synset Relations

Finally, in this brief summary, we will look at the constraints which we have defined for senses and synsets, the latter of which, it should be noted, are encoded in GWA RDF using the OntoLex class `LexicalConcept`. As for senses, we were able to capture all of the constraints found in the DTD declarations apart from those pertaining to the so called 'lex file', since we were unable to find relevant pre-existing vocabularies to encode this in RDF. With regards to the relationships between synsets, encoded in the GWA LMF format as *SynsetRelation*s, we define a `ConceptualRelationShape` which allows us to restrict the relationships between synsets to those proposed by the GWA.

```
    ex:ConceptualRelationShape a sh:NodeShape ;
    sh:targetClass vartrans:ConceptualRelation ;
    sh:property [
    sh:name "Category" ;
sh:description "Make␣sure␣the␣Synset␣Relation␣
     belongs␣to␣the␣correct␣category" ;
      sh:path vartrans:category ;
      sh:minCount 1 ;
      sh:maxCount 1 ;
      sh:in (wn:agent wn:also wn:attribute wn:
           be_in_state wn:causes wn:
           classified_by wn:classifies wn:
           co_agent_instrument wn:
           co_agent_patient wn:co_agent_result
           wn:co_instrument_agent wn:
           co_instrument_patient wn:
           co_instrument_result wn:
           co_patient_agent wn:
           co_patient_instrument wn:
           co_result_agent wn:
           co_result_instrument wn:co_role wn:
           direction...  wn:ir_synonym wn:
           similar)];
```

Listing 5: Part of Speech Information

5 Summary and Conclusions

In this paper, we have discussed the creation of a SHACL shapes graph for the GWA RDF format. We have motivated the need for such a resource and detailed our first (and fairly comprehensive) attempt at such a graph. In summary, we have covered the following classes mentioned in the original LMF DTD:

- *Lexicon*, *Lexical Entry*, *Form*, *Pronunciation*, *Tag*, *Definition*, *ILI Definition*, *Example*, *Sense*, *Synset*, *Sense Relation*, *Synset Relation*

In addition we have partially covered *SyntacticBehaviour*. The following classes that are present in the original LMF are not explicitly covered as we were not aware of a settled best practice for representing this information in RDF using the OntoLex vocabulary and, moreover, this information is not common in wordnet resources:

- *Lexicon Extension*, *Requires*, *Extends*, *External Lexical Entry*, *External Lemma*, *External Form*, *External Sense*, *External Synset*

In the case of the External prefixed elements (e.g. External Lexical Entry), it may turn out that given the linking mechanism in RDF there is no need to define specific shapes here. In any case we have managed to to cover all of the commonly used parts of the schema used by the Global Wordnet Association.

[8] Note that in Ontolex, senses lexicalize concepts rather than being lexicalized themselves.

Acknowledgements

John P. McCrae is supported by Research Ireland under Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and Grant Number 13/RC/2106_P2, ADAPT SFI Research Centre.

References

Francis Bond, Piek Vossen, John McCrae, and Christiane Fellbaum. 2016. CILI: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 50–57, Bucharest, Romania. Global Wordnet Association.

Gavin Carothers and Eric Prud'hommeaux. 2014. RDF 1.1 Turtle. W3C recommendation, W3C.

Philipp Cimiano, John P. McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Community report. W3C community report.

Christiane Fellbaum. 2010. *WordNet*, pages 231–243. Springer Netherlands, Dordrecht.

Holger Knublauch and Dimitris Kontokostas. 2017. Shapes constraint language (SHACL). W3C recommendation, W3C.

John McCrae, Dennis Spohr, and Philipp Cimiano. 2011. Linking lexical resources and ontologies on the semantic web with lemon. In *Proc. of the 8th Extended Semantic Web Conference*, pages 245–249.

John P. McCrae, Michael Wayne Goodman, Francis Bond, Alexandre Rademaker, Ewa Rudnicka, and Luis Morgado Da Costa. 2021. The GlobalWordNet formats: Updates for 2020. In *Proceedings of the 11th Global Wordnet Conference*, pages 91–99, University of South Africa (UNISA). Global Wordnet Association.

John P. McCrae, Alexandre Rademaker, Francis Bond, Ewa Rudnicka, and Christiane Fellbaum. 2019. English WordNet 2019 – an open-source WordNet for English. In *Proceedings of the 10th Global Wordnet Conference*, pages 245–252, Wroclaw, Poland. Global Wordnet Association.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Claudia Soria, Monica Monachini, and Piek Vossen. 2009. Wordnet-LMF: fleshing out a standardized format for wordnet interoperability. In *Proceedings of the 2009 International Workshop on Intercultural Collaboration*, pages 139–146.

Piek Vossen. 2004. EuroWordNet: a multilingual database of autonomous and language-specific wordnets connected via an inter-lingualindex. *International Journal of Lexicography*, 17(2):161–173.

Piek Vossen, Francis Bond, and John McCrae. 2016. Toward a truly multilingual GlobalWordnet grid. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pages 424–431, Bucharest, Romania. Global Wordnet Association.