

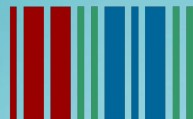
E-Schemas in Evaluating Wikidata References

Referencing Quality Scoring System (RQSS)

Seyed Amir Hosseini Beghaeiraveri

Semantic Web PhD researcher

Heriot-Watt University



WIKIDATA

DATA QUALITY
DAYS 2022

Referencing Quality Scoring System (RQSS)

A Comprehensive Framework for Assessing the Quality of Referencing in Linked Data

6

CATEGORIES

21

DIMENSIONS

40

METRICS

4

SHEX-RELATED
METRICS

Metric 1: Syntactic Validity of Reference Triples

Examines the **Wikidata Referencing reification ShEx schema** over all statement-reference nodes

Statement Node



`prov:wasDerivedFrom`



Reference Node

Simple Value



`prv:opq`

`pr:xyz`



`wikibase:quantityAmount`
`Wikibase:timeValue`

...

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
START=@<statement_node>
<statement_node> {
  prov:wasDerivedFrom @<reference_node>* ;
}
<reference_node> {
  <ref-property> xsd:decimal OR xsd:integer OR
  xsd:dateTime OR xsd:string OR IRI* ;
  <ref-property> @<value>* ;
}
<value> {
  wikibase:quantityAmount xsd:decimal OR xsd:integer?;
  wikibase:timeValue xsd:dateTime?;
  #...
}
```

Metric 1: Syntactic Validity of Reference Triples

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
```

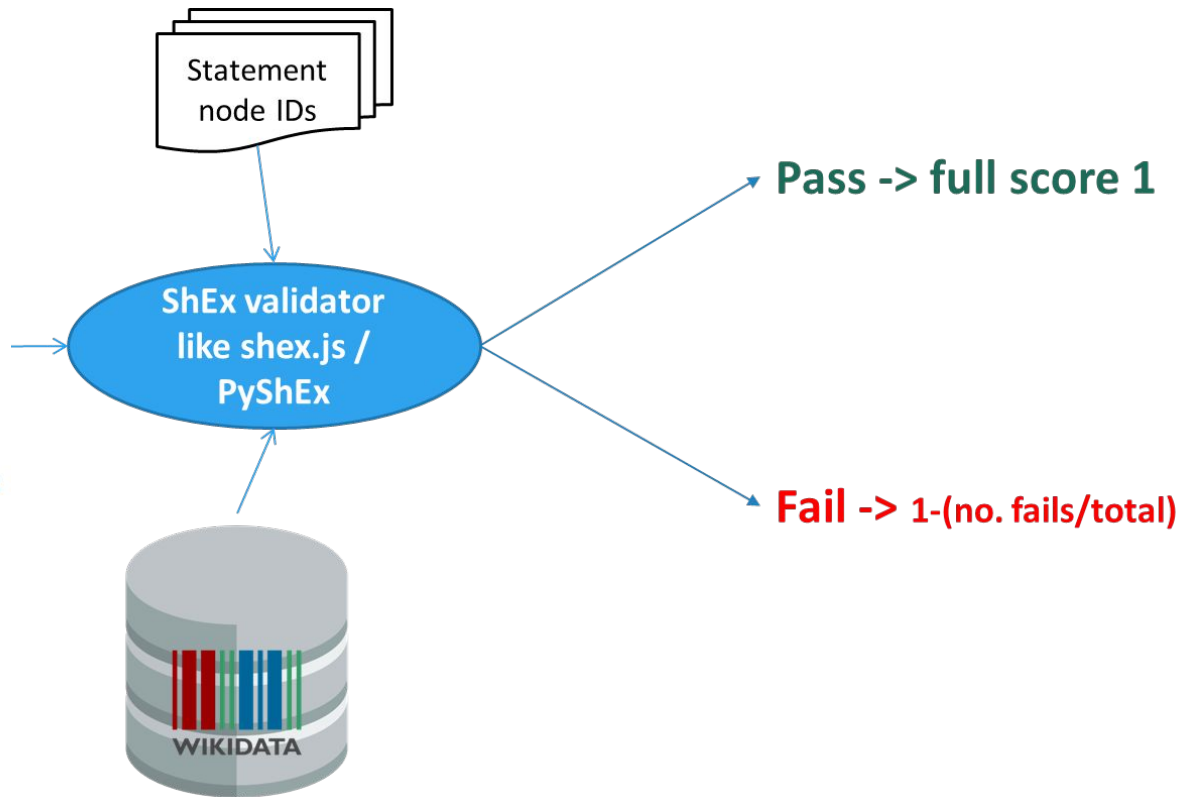
```
START=@<statement_node>
```

```
<statement_node> {  
  prov:wasDerivedFrom @<reference_node>* ;  
}
```

```
<reference_node> {  
  <ref-property> xsd:decimal OR xsd:integer OR  
  xsd:dateTime OR xsd:string OR IRI* ;  
  <ref-property> @<value>* ;  
}
```

```
<value> {  
  wikibase:quantityAmount xsd:decimal OR xsd:integer?;  
  wikibase:timeValue xsd:dateTime?;  
  #...  
}
```

**Wikidata RDF model
of references Shape**



Metric 1: Syntactic Validity of Reference Triples

Subset	Num of Statement Nodes	Num of Fails	Score
Gene Wiki	97,062,660	Under computation	Under computation
Music	12,743,480	2,798	99.98%
Ships	1,116,976	51	99.99%
Random 100K #1	1,225,313	580	99.95%
Random 100K #2	1,226,097	624	99.94%
Random 500K	6,117,915	2,482	99.95%
Random 1M	12,231,380	4,945	99.95%

Metric 2, 3: ShEx in Completeness

Schema completeness of references

How many classes and fact types have a defined schema for references?

- The richness of input in terms of schema

Schema-based property completeness of references

If a reference schema is defined for class/fact of type X, how many instances of X have got a reference with the property mentioned in the schema?

- Adjustment of the actual data with the schema

Metric 2, 3: ShEx in Completeness

In Wikidata: Entity Schemas (Eids)

```
<#reference> { } # Any reference will suffice.
```

```
<#disease-ontology-reference> { # reference to a term from the disease ontology term
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P699 xsd:string ; # Disease Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}
```

```
<#mondo-disease-reference> { # reference to a term from the MonDo ontology
  pr:P248 [ wd:Q27468140 ] ; # stated in [P248] Mondo disease ontology [Q27468140]
  pr:P5270 xsd:string ; # Mondo ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}
```

```
<#symptom-ontology-reference> { # reference to a term from the Symptom Ontology
  pr:P248 [ wd:Q5282129 ] ; # stated in [P248] Symptom ontology [Q27468140]
  pr:P8656 xsd:string ; # Symptom Ontology ID
  pr:P813 xsd:dateTime ; # Date of retrieval
}
```

<https://www.wikidata.org/wiki/EntitySchema:E273>

How many disease instances has got reference using P248/P699/P813

How many symptom instances has got reference using P248/P8656/P813

Metric 2, 3: ShEx in Completeness

Needed information:

In Total:

- How many EntitySchemas do we have?
- What classes do each EntitySchemas cover? (Related classes)

In the text of each schema:

- What facts in the EntitySchema text are specified to have referenced and what kind of references should be used for them?

Metric 2, 3: ShEx in Completeness

In Total:

- How many EntitySchemas do we have?
- What classes do each EntitySchemas cover? (Related classes)

arts [\[edit \]](#)

label ↕	description ↕	aliases ↕	class/Property ↕	dependencies ↕
actor (E25)	basic scheme for actor		occupation (P106)= actor (Q33999), Actress (Q21169216)	
album (E248)	simple schema for albums	audio album music album record album	album (Q482994)	
author (E42)	A Schema for authors		occupation (P106)= writer (Q36180)	
camera operator (E29)	basic scheme for camera operators	cameraman camerawoman	occupation (P106)= camera operator (Q1208175)	
cinematographer (E28)	basic schema for cinematographer		occupation (P106)= cinematographer (Q222344)	

https://www.wikidata.org/wiki/Wikidata:Database_reports/EntitySchema_directory

Metric 2, 3: ShEx in Completeness

In the text of each schema:

- What facts in the EntitySchema text are specified to be referenced and what kind of references should be used for them?

Challenges:

- Some E-ids are empty (e.g. E9) or invalid (e.g. E368, E93)
- ShEx is like SPARQL: required parsing
- References shapes might be in different E-ids:

```
## Statement details
<#P31_instance_of_gene> {
  ps:P31 [wd:Q7187] ;      # Instance of [P31] gene
  prov:wasDerivedFrom @E265:ncbi-gene-reference OR @E265:ensembl-gene-reference ;
}
```

Metric 2, 3: ShEx in Completeness

A Python [script based on PyShEx](#) to parse E-ids

Number of E-ids: 319

- On 9 Jul 2022

Number of E-ids with referenced facts: 13

Subset	class schema completeness for references	property schema completeness for references	Schema-based property completeness
Ships	24% of 989 classes	21% of 28,018 properties	6% of 490,748 property-reference pairs
Random 100K #1	44% of 2552 classes	6 % of 42,164 properties	25% of 2,754,860 property-reference pairs




Metric 2, 3, 4: Completeness for ShEx

If a fact of type X
has a reference
using the
ref-property Y, how
many other type X
facts have a
reference using
property Y?

Populating E-ids

Albert Einstein (Q937)

German-born theoretical physicist; developer

<u>relative</u> X	 Lina Einstein kinship to subject ▼ 1 reference <u>reference URL</u> Y
	 Elsa Einstein kinship to subject ▼ 1 reference
languages spoken, written or signed	 German ▼ 0 references

no Y!

Ernest Rutherford (Q9123)

New Zealand-born British chemist and physicist (1871-1937)

<u>relative</u> X	 Ralph H. Fowler kinship to subject ▼ 0 references
languages spoken, written or signed	 English no Y! ▼ 2 references stated in Bibliothèque nationale de France ID reference URL retrieved reference URL

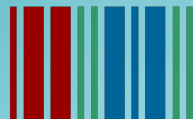
Y

Thanks for your attention!

Seyed Amir Hosseini Beghaeiraveri

Semantic Web PhD researcher
Heriot-Watt University

© Seyed Amir Hosseini Beghaeiraveri
All rights waived via CC0



WIKIDATA

DATA QUALITY
DAYS 2022