Writing Schemas for Wikidata

Kat Thornton

30 October, 2021

Spaceport

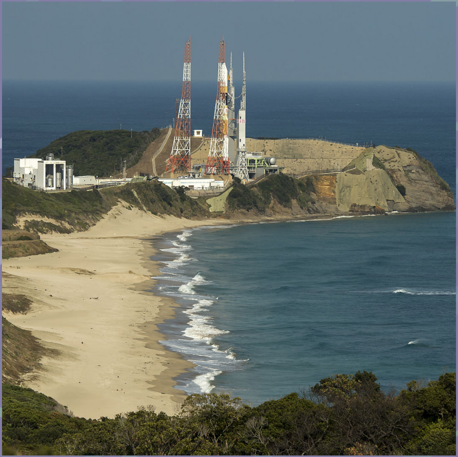
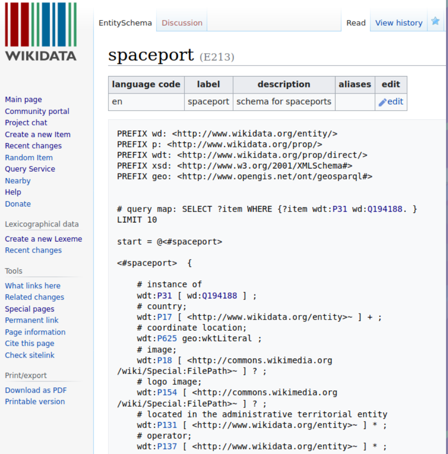


Figure 1: Tanegashima Space Center, Photo Credit: NASA/Bill Ingalls, via Wikimedia Commons

Finding a Schema



spaceport (E213)

language code	label	description	aliases	edit
en	spaceport	schema for spaceports		edit

```
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>


# query map: SELECT ?item WHERE {?item wdt:P31 wd:Q194188. }
LIMIT 10

start = @<#spaceport>

<#spaceport>  {

    # instance of
    wdt:P31 [ wd:Q194188 ] ;
    # country;
    wdt:P17 [ <http://www.wikidata.org/entity>~ ] + ;
    # coordinate location;
    wdt:P625 geo:wktLiteral ;
    # image;
    wdt:P18 [ <http://commons.wikimedia.org
/wiki/Special:FilePath>~ ] ? ;
    # logo image;
    wdt:P154 [ <http://commons.wikimedia.org
/wiki/Special:FilePath>~ ] ? ;
    # located in the administrative territorial entity
    wdt:P131 [ <http://www.wikidata.org/entity>~ ] * ;
    # operator;
    wdt:P137 [ <http://www.wikidata.org/entity>~ ] * ;
```

Figure 2: E213 Schema for Spaceports

Schemas are data models

Schema Namespace

- Wikidata's E Namespace is dedicated to schemas.
- 300+ schemas have been contributed since the namespace became available in May, 2019.
- Wikidata supports multiple data models per domain.
- In some cases this is necessary, in other cases it will be possible to build consensus around one shared model.

ShEx

- ShEx is a formal modeling and validation language for RDF graphs
- Allows humans and machines communicate unambiguously about data assets
- Supports agile development of data models
- Learn more: http://shex.io

Humans and Machines Together

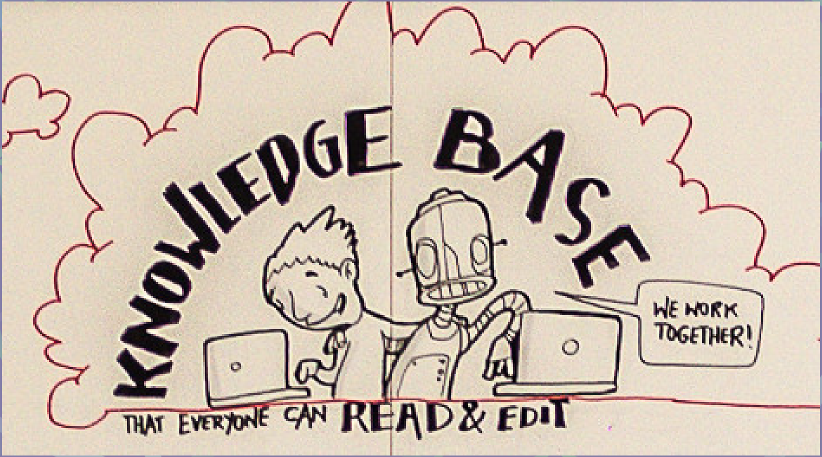


Figure 3: Detail of mural by Magdalena Wiegner

Members of National Academies of Science with academic degree PhD

```
SELECT ?item ?itemLabel ?id WHERE
{
?item wdt:P5380 ?id .
?item wdt:P512 wd:Q752297.
}
```

Members of National Academies of Science, educated at statement, academic degree as a qualifier

```
SELECT ?item ?id
WHERE
{
?item wdt:P5380 ?id .
?item p:P69 [
ps:P69 ?inst ;
pq:P512 wd:Q752297 ].
}
```

Building Consensus

- Frequency of usage of the 2 patterns?
- Propose a schema
- Discuss on talk page of schema

A Shared Data Model for Academic Degrees is Desirable

- People who would like to reuse Wikidata data could become frustrated by needing to identify multiple related modeling patterns.

Schemas: Internal or External

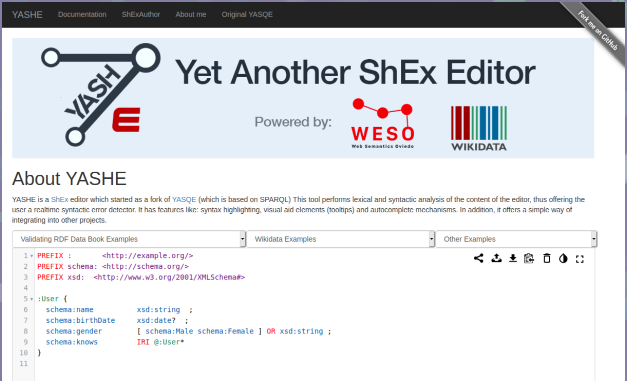
Bringing an External Schema into Wikidata

- Expressing a model from a database, ontology, system external to Wikidata
- PRONOM technical registry of file formats
- file format with PRONOM id (E79) and
- File format with identification pattern (E237)

Extracting an Internal Schema from Wikidata

- What if there is no schema yet?
- Then we can extract a schema
- Tools to try:
- sheXer
- Shape Designer

Extract Schemas from Wikibase



About YASHE

YASHE is a ShEx editor which started as a fork of YASQE (which is based on SPARQL) This tool performs lexical and syntactic analysis of the content of the editor, thus offering the user a realtime syntactic error detector. It has features like: syntax highlighting, visual aid elements (tooltips) and autocomplete mechanisms. In addition, it offers a simple way of integrating into other projects.

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

:User {
  schema:name          xsd:string  ;
  schema:birthDate     xsd:date?  ;
  schema:gender        [ schema:Male schema:Female ] OR xsd:string ;
  schema:knows         IRI @:User*
}

```

Figure 4: Try YASHE out!

ShEx Implementations

- shex.js (runs on n3.js)
- SHACLex (Scala)
- ShexJava
- PyShEx
- Ruby ShEx

ShEx Test Suite

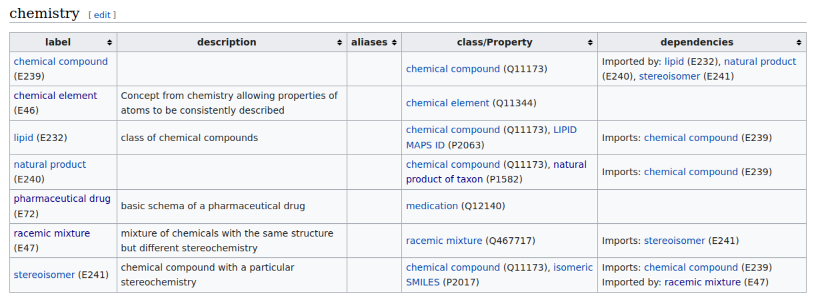
- 1103 validation tests
- 99 negative syntax tests
- 14 negative structure tests
- 431 schema conversion tests between ShExC, ShExJ and ShExR

Online Validators for ShEx

- ShEx2 Simple Online Validator (JavaScript)
- RDF Shape (Scala)

Interlinked ecosystem of schemas

Schema Gallery

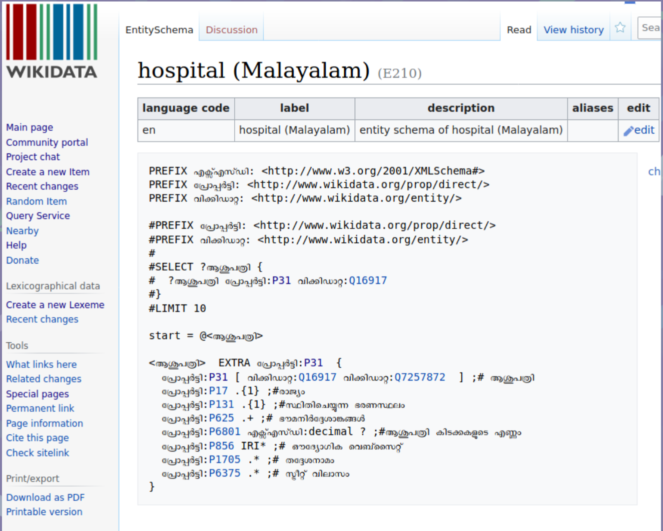


chemistry [edit]

label	description	aliases	class/Property	dependencies
chemical compound (E239)			chemical compound (Q11173)	Imported by: lipid (E232), natural product (E240), stereoisomer (E241)
chemical element (E46)	Concept from chemistry allowing properties of atoms to be consistently described		chemical element (Q11344)	
lipid (E232)	class of chemical compounds		chemical compound (Q11173), LIPID MAPS ID (P2063)	Imports: chemical compound (E239)
natural product (E240)			chemical compound (Q11173), natural product of taxon (P1582)	Imports: chemical compound (E239)
pharmaceutical drug (E72)	basic schema of a pharmaceutical drug		medication (Q12140)	
racemic mixture (E47)	mixture of chemicals with the same structure but different stereochemistry		racemic mixture (Q467717)	Imports: stereoisomer (E241)
stereoisomer (E241)	chemical compound with a particular stereochemistry		chemical compound (Q11173), isomeric SMILES (P2017)	Imports: chemical compound (E239) Imported by: racemic mixture (E47)

Figure 5: Chemistry schemas available in Wikidata

Multilingual Schemas

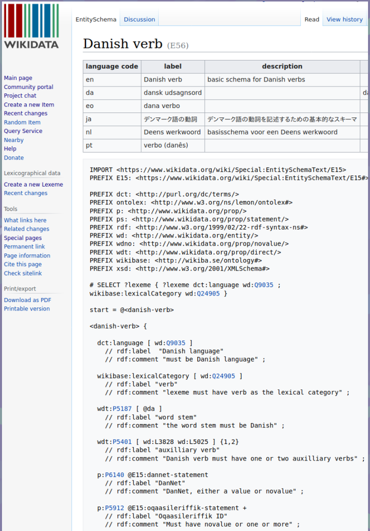


hospital (Malayalam) (E210)

language code	label	description	aliases	edit
en	hospital (Malayalam)	entity schema of hospital (Malayalam)		edit

```
PREFIX എക്സ്എസ്ഡി: <http://www.w3.org/2001/XMLSchema#>
PREFIX പ്രോപ്പർട്ടി: <http://www.wikidata.org/prop/direct/>
PREFIX വിക്കിഡാറ്റ: <http://www.wikidata.org/entity/>

#PREFIX പ്രോപ്പർട്ടി: <http://www.wikidata.org/prop/direct/>
#PREFIX വിക്കിഡാറ്റ: <http://www.wikidata.org/entity/>
#
#SELECT ?ആശുപത്രി {
#  ?ആശുപത്രി പ്രോപ്പർട്ടി:P31 വിക്കിഡാറ്റ:Q16917
#}
#LIMIT 10

start = @<ആശുപത്രി>

<ആശുപത്രി>  EXTRA പ്രോപ്പർട്ടി:P31  {
  പ്രോപ്പർട്ടി:P31 [ വിക്കിഡാറ്റ:Q16917 വിക്കിഡാറ്റ:Q7257872  ] ;# ആശുപത്രി
  പ്രോപ്പർട്ടി:P17 .{1} ;#രാജ്യം
  പ്രോപ്പർട്ടി:P131 .{1} ;#സ്ഥിതിചെയ്യുന്ന ഭരണസ്ഥലം
  പ്രോപ്പർട്ടി:P625 .+ ;# ഭൗമനിർദ്ദേശാങ്കങ്ങൾ
  പ്രോപ്പർട്ടി:P6801 എക്സ്എസ്ഡി:decimal ? ;#ആശുപത്രി കിടക്കകളുടെ എണ്ണം
  പ്രോപ്പർട്ടി:P856 IRI* ;# ഔദ്യോഗിക വെബ്സൈറ്റ്
  പ്രോപ്പർട്ടി:P1705 .* ;# തദ്ദേശനാമം
  പ്രോപ്പർട്ടി:P6375 .* ;# സ്ട്രീറ്റ് വിലാസം
}
```

Importing One Schema into Another



Danish verb (E56)

language code	label	description	
en	Danish verb	basic schema for Danish verbs	
da	dansk udsagnsord		da
eo	dana verbo		
ja	デンマーク語の動詞	デンマーク語の動詞を記述するための基本的なスキーマ	
nl	Deens werkwoord	basisschema voor een Deens werkwoord	
pt	verbo (danês)		

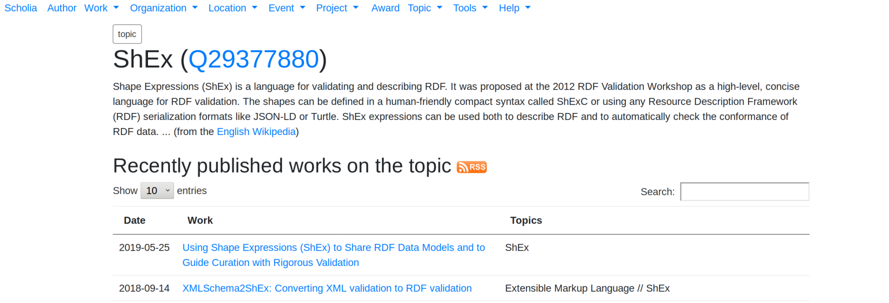
```
IMPORT <https://www.wikidata.org/wiki/Special:EntitySchemaText/E15>
PREFIX E15: <https://www.wikidata.org/wiki/Special:EntitySchemaText/E15#>

PREFIX dct: <http://purl.org/dc/terms/>
PREFIX ontolex: <http://www.w3.org/ns/lemon/ontolex#>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdno: <http://www.wikidata.org/prop/novalue/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

# SELECT ?lexeme { ?lexeme dct:language wd:Q9035 ;
wikibase:lexicalCategory wd:Q24905 }

start = @<danish-verb>

<danish-verb> {

  dct:language [ wd:Q9035 ]
    // rdf:label  "Danish language"
    // rdf:comment "must be Danish language" ;

  wikibase:lexicalCategory [ wd:Q24905 ]
    // rdf:label "verb"
    // rdf:comment "lexeme must have verb as the lexical category" ;

  wdt:P5187 [ @da ]
    // rdf:label "word stem"
    // rdf:comment "the word stem must be Danish" ;

  wdt:P5401 [ wd:L3828 wd:L5025 ] {1,2}
    // rdf:label "auxilliary verb"
    // rdf:comment "Danish verb must have one or two auxilliary verbs" ;

  p:P6140 @E15:dannet-statement
    // rdf:label "DanNet"
    // rdf:comment "DanNet, either a value or novalue" ;

  p:P5912 @E15:oqaasileriffik-statement +
    // rdf:label "Oqaasileriffik ID"
    // rdf:comment "Must have novalue or one or more" ;
```

What’s Next?

- More editors will contribute schemas
- More schemas will be interlinked via IMPORT
- This ecosystem of interlinking schemas will support the work of more editors
- Wikidatans will create more tooling that leverages schemas

Find ShEx Papers

- Scholia is a Wikidata-based service that generates scholarly profiles
- Scholia for shex


Scholia Author Work Organization Location Event Project Award Topic Tools Help
topic

ShEx (Q29377880)

Shape Expressions (ShEx) is a language for validating and describing RDF. It was proposed at the 2012 RDF Validation Workshop as a high-level, concise language for RDF validation. The shapes can be defined in a human-friendly compact syntax called ShExC or using any Resource Description Framework (RDF) serialization formats like JSON-LD or Turtle. ShEx expressions can be used both to describe RDF and to automatically check the conformance of RDF data. ... (from the English Wikipedia)

Recently published works on the topic RSS

Show 10 entries Search:

Date	Work	Topics
2019-05-25	Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation	ShEx
2018-09-14	XMLSchema2ShEx: Converting XML validation to RDF validation	Extensible Markup Language // ShEx

Quick Start

- Jose Labra's Wikipedia Weekly Entity Schemas and Shape Expressions session
- Primer: http://shex.io/shex-primer/
- ISWC 2020 Tutorial Shapes applications and tools
- Read the book Validating RDF Data

Getting started with ShEx on Wikidata

- Wikidata Wikiproject ShEx
- Schema namespace
- Browse a gallery of schemas

Wikidata's E Namespace: Where the Ecosystem of Schemas Thrives

Thank you!

9 YEARS

WIKIDATA