

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



Main page
Community portal
Project chat
Create a new item
Recent changes
Random item
Query Service
Nearby
Help
Donate
Lexicographical data
Create a new Lexeme
Recent changes
Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page
Check sitelink
Print/export
Download as PDF
Printable version

EntitySchema Discussion Read View history ★

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

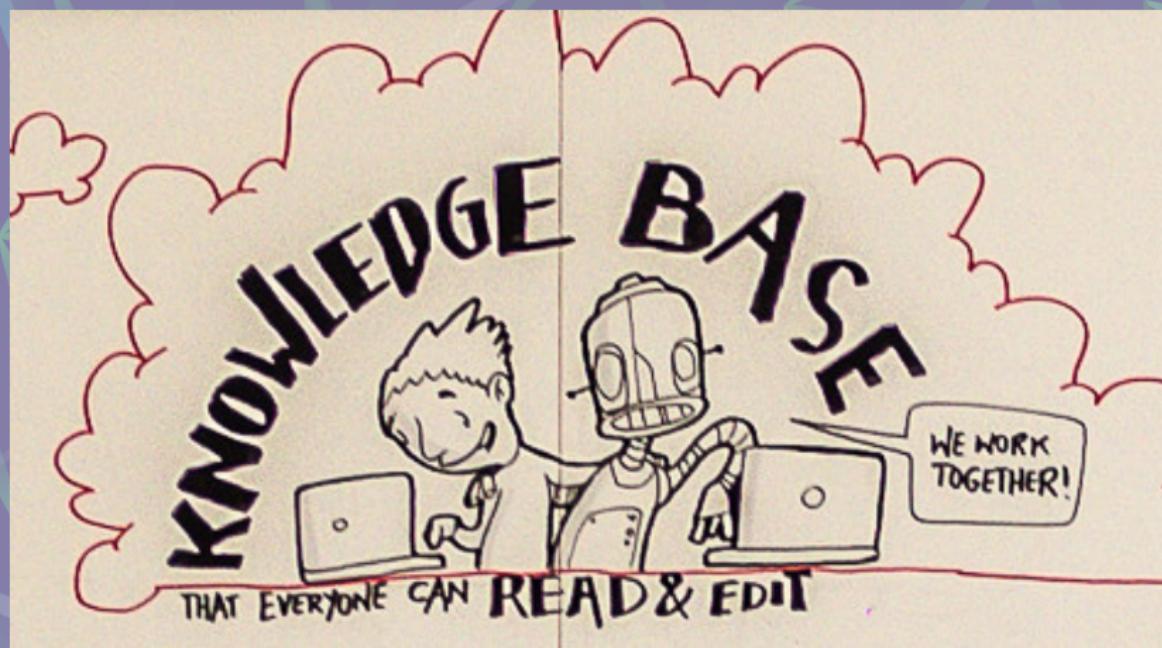


Figure 3: Detail of mural by Magdalena Wiegner

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 Wiki-data

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

Extract Schemas from Wikibase

The screenshot shows the YASHE (Yet Another ShEx Editor) web application. At the top, there's a navigation bar with links to 'YASHE', 'Documentation', 'ShExEditor', 'About me', and 'Original YASQE'. A 'Fork me on GitHub' button is also present. The main header features a logo with the letters 'YASH' and 'E' and the text 'Yet Another ShEx Editor'. Below the header, it says 'Powered by:' followed by logos for 'WESO Web Semantics Oaxaca' and 'WIKIDATA'. The central area is titled 'About YASHE' and contains a brief description: '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.' There are three tabs at the top of the code editor: 'Validating RDF Data Book Examples', 'Wikidata Examples', and 'Other Examples'. The code editor itself contains the following ShEx code:

```
1 * PREFIX : <http://example.org/>
2 * PREFIX schema: <http://schema.org/>
3 * PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
4 *
5 * :User {
6 *   schema:name xsd:string ;
7 *   schema:birthDate xsd:date? ;
8 *   schema:gender [ schema:Male schema:Female ] OR xsd:string ;
9 *   schema:knows IRI @:User*
10 }
11
```

Below the code editor are several small icons for file operations.

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

EntitySchema Discussion Read View history ⭐ Search

hospital (Malayalam) (E210)

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

PREFIX എൽജീബിഡി: <http://www.w3.org/2001/XMLSchema#>
PREFIX ഓക്സിഡറ്റ്: <http://www.wikidata.org/prop/direct/>
PREFIX വികിഡാറ്റ്: <http://www.wikidata.org/entity/>

#PREFIX ഓട്ടുഡി: <http://www.wikidata.org/prop/direct/>
#PREFIX വികിഡാറ്റ്: <http://www.wikidata.org/entity/>

#SELECT ?കൗൺസിൽ {
?കൗൺസിൽ ഓട്ടുഡി:P31 വികിഡാറ്റ്:Q16917
#}
#LIMIT 10

start = @<മൂലസ്തുതി>

<മൂലസ്തുതി> EXTRA ഓട്ടുഡി:P31 {
ഓട്ടുഡി:P31 [വികിഡാറ്റ്:Q16917 വികിഡാറ്റ്:07257872] ;# കൗൺസിൽ
ഓട്ടുഡി:P17 .{1} ;#ഒരും
ഓട്ടുഡി:P131 .{1} ;#ഒരുവിശദ്ധന ഉള്ളായിലും
ഓട്ടുഡി:P625 .+ ;# അനാർഥിക്യവാക്യങ്ങൾ
ഓട്ടുഡി:P6801 എൽജീബിഡി:decimal ? ;#കൗൺസിൽ കിട്ടുകളുടെ ഏറ്റവും
ഓട്ടുഡി:P856 IRI* ;# ശാഖാവിക ലോഗിന്റെ
ഓട്ടുഡി:P1705 .+ ;# മാറ്റൊന്നു
ഓട്ടുഡി:P6375 .* ;# സ്കീഞ്ചർ വിവരങ്ങൾ }

Importing One Schema into Another



The image shows a screenshot of a Wikidata EntitySchema page for the Danish verb (E166). The page has a header with tabs for EntitySchema and Discussion, and buttons for Read and View history. Below the header is a table with columns for language code, label, and description. The table contains the following data:

language code	label	description
en	Danish verb	basic schema for Danish verbs
da	dansk udsagnsord	
eo	dana verbo	
ja	デーマーク語の動詞	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 (danes)	

Below the table is a large block of SPARQL query code. The code starts with an import statement and defines several prefixes: dct, v3, prop, statement, entity, wd, and wikibase. It then defines a variable start with the value @:danish-verb#. The query uses this variable in a SELECT clause to find all lexemes where the language is the Danish language (wd:09835) and the lexical category is a verb (wd:024985). It also includes a condition that the verb must have a word stem (wd:PS187). The query then lists statements for the Danish verb, including its auxiliaries (P5481), and concludes with a statement about the verb being either a value or no value (P56140).

```
IMPORT <https://www.wikidata.org/wiki/Special:EntitySchemaText/E16>
PREFIX E15: <https://www.wikidata.org/wiki/Special:EntitySchemaText/E15#>

PREFIX dct: <http://purl.org/dc/terms/>
PREFIX v3: <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 wdrc: <http://www.wikidata.org/prop/rect/>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>

a SELECT ?lexeme { ?lexeme dct:language wd:09835 ;
wikibase:lexicalCategory wd:024985 }

start = @:danish-verb#

<danish-verb> {

dct:language [ wd:09835 ]
// rdf:label "Danish language"
// rdf:comment "must be Danish language"

wikibase:lexicalCategory [ wd:024985 ]
// rdf:label "verb"
// rdf:comment "lexeme must have verb as the lexical category" ;

wd:PS187 [ @ea ]
// rdf:label "word stem"
// rdf:comment "the word stem must be Danish" ;

wd:P5481 [ wd:L3828 wd:L5925 ] (1,2)
// rdf:label "auxiliary verb"
// rdf:comment "Danish verb must have one or two auxiliary verbs" ;

P56140 @E15:danner-statement
// rdf:label "DanNet, either a value or novalue" ;

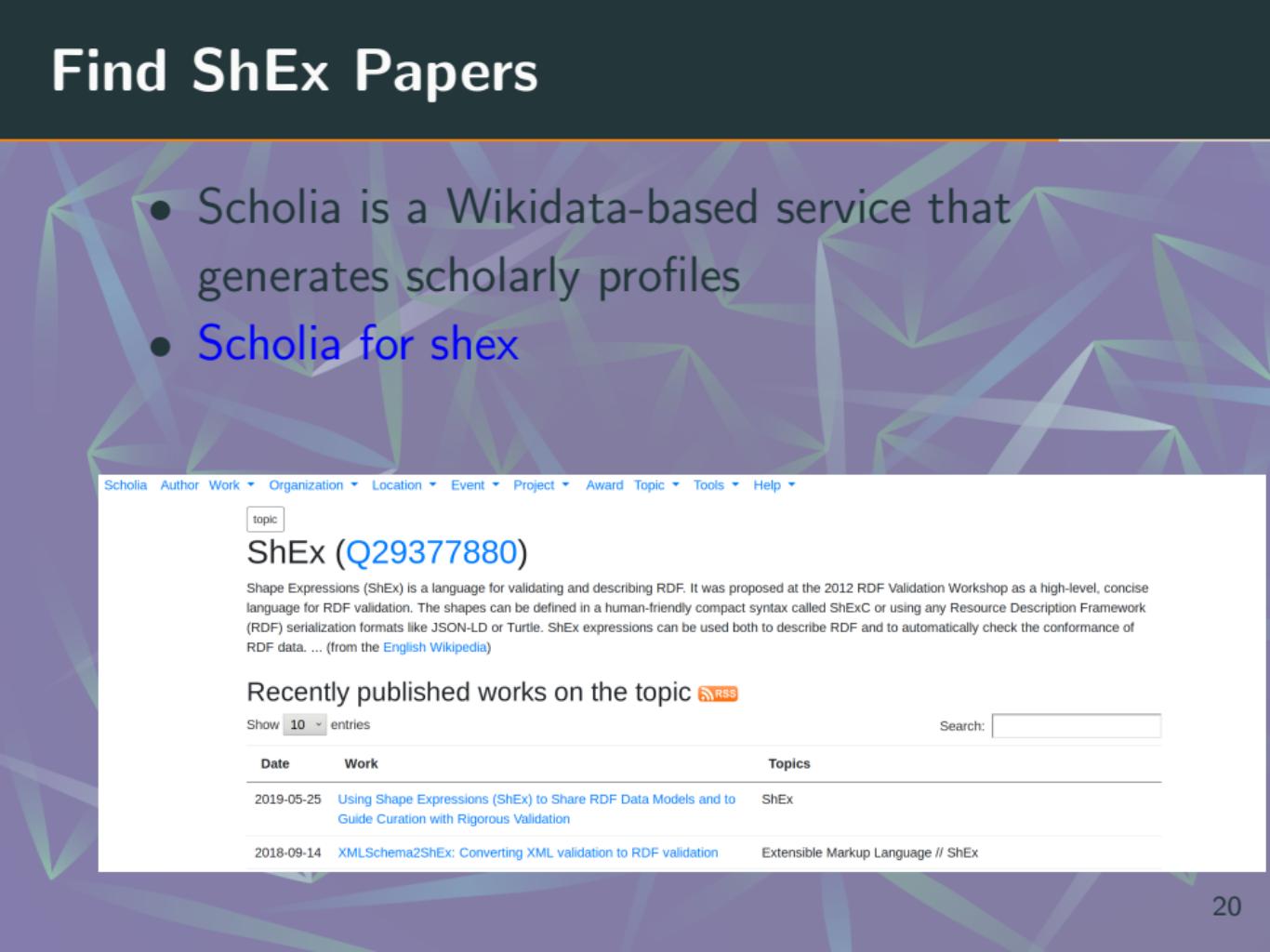
P5912 @E15:ogpassiveriffrifik-statement +
// rdf:label "Ogpassiveriffrifik ID"
// rdf:comment "Must have novalue or one or more" ;
```

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

Show entries

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

Search:

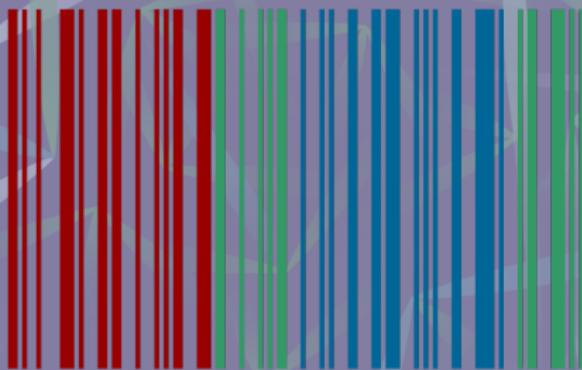
Quick Start

- 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!



WIKIDATA