Using Shape Expressions (ShEx) to Share RDF Data Models and to Guide Curation with Rigorous Validation

ESWC 2019

Katherine Thornton, Harold Solbrig, Gregory Stupp, Jose Labra, Daniel Mietchen, Eric Prud'Hommeaux and Andra Waagmeester

5 June, 2019

Link for these slides

- https://w.wiki/4dr

Our Collaboration

- 2 partially overlapping communities:
- **ShEx Community Group**
- **Wikidatans**

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

ShEx Implementations

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

ShEx Test Suite

- 1088 validation tests
- 99 negative syntax tests
- 14 negative structure tests
- 408 schema conversion tests between ShExC, ShExJ and ShExR

Online Validators for ShEx

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

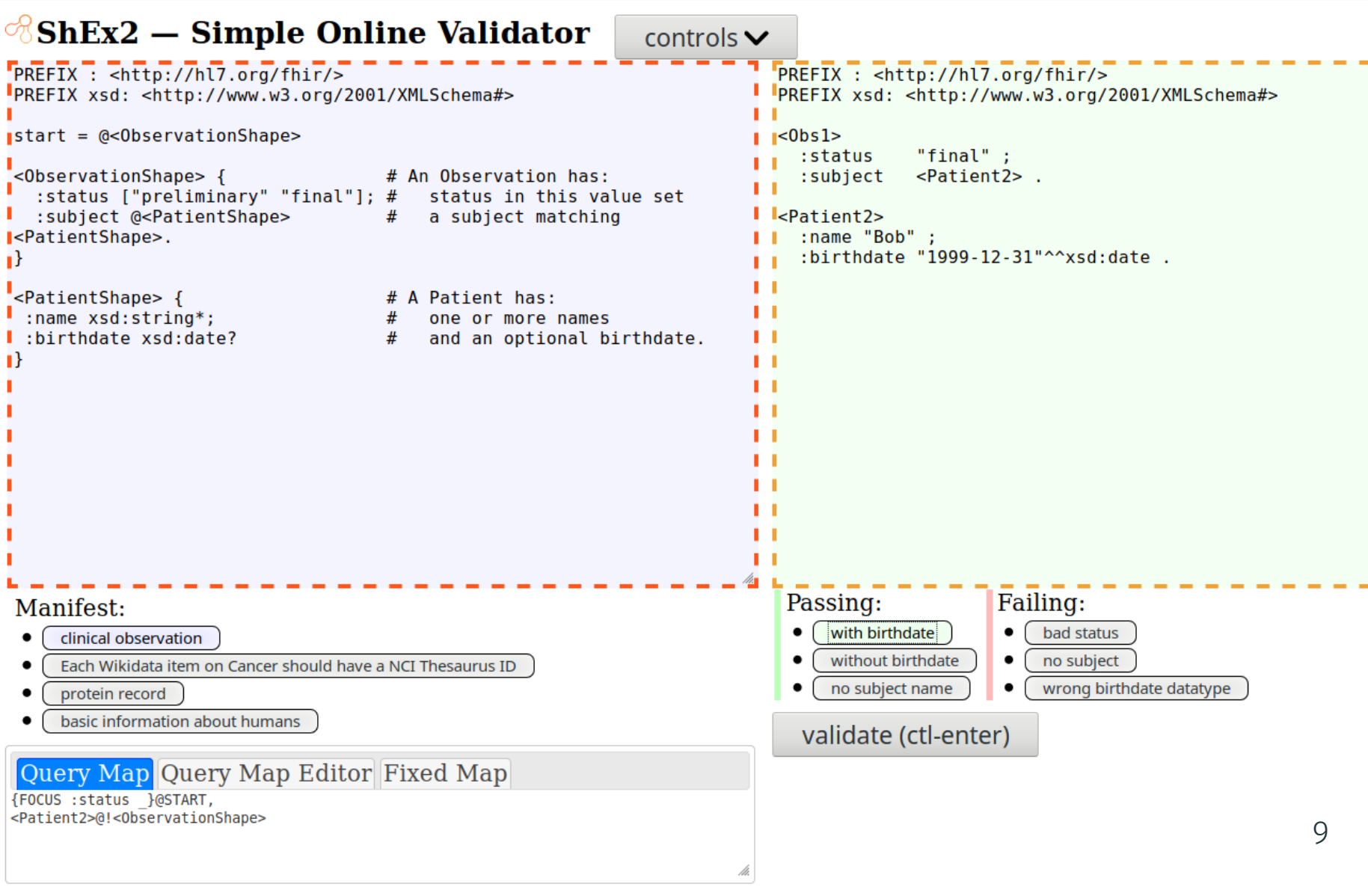
Three Use Cases

- **FHIR** (medical)
- **GeneWiki** (biomedical)
- **Digital Preservation** (computing)

Use Case One: FHIR

- Fast Healthcare Interoperability Resources (FHIR) specification defines 130+ healthcare resources and how they are represented in RDF, XML and JSON
- The FHIR documentation production framework calls the Shaclex implementation
- Exchange of electronic medical records generated in different software systems

FHIR schema example



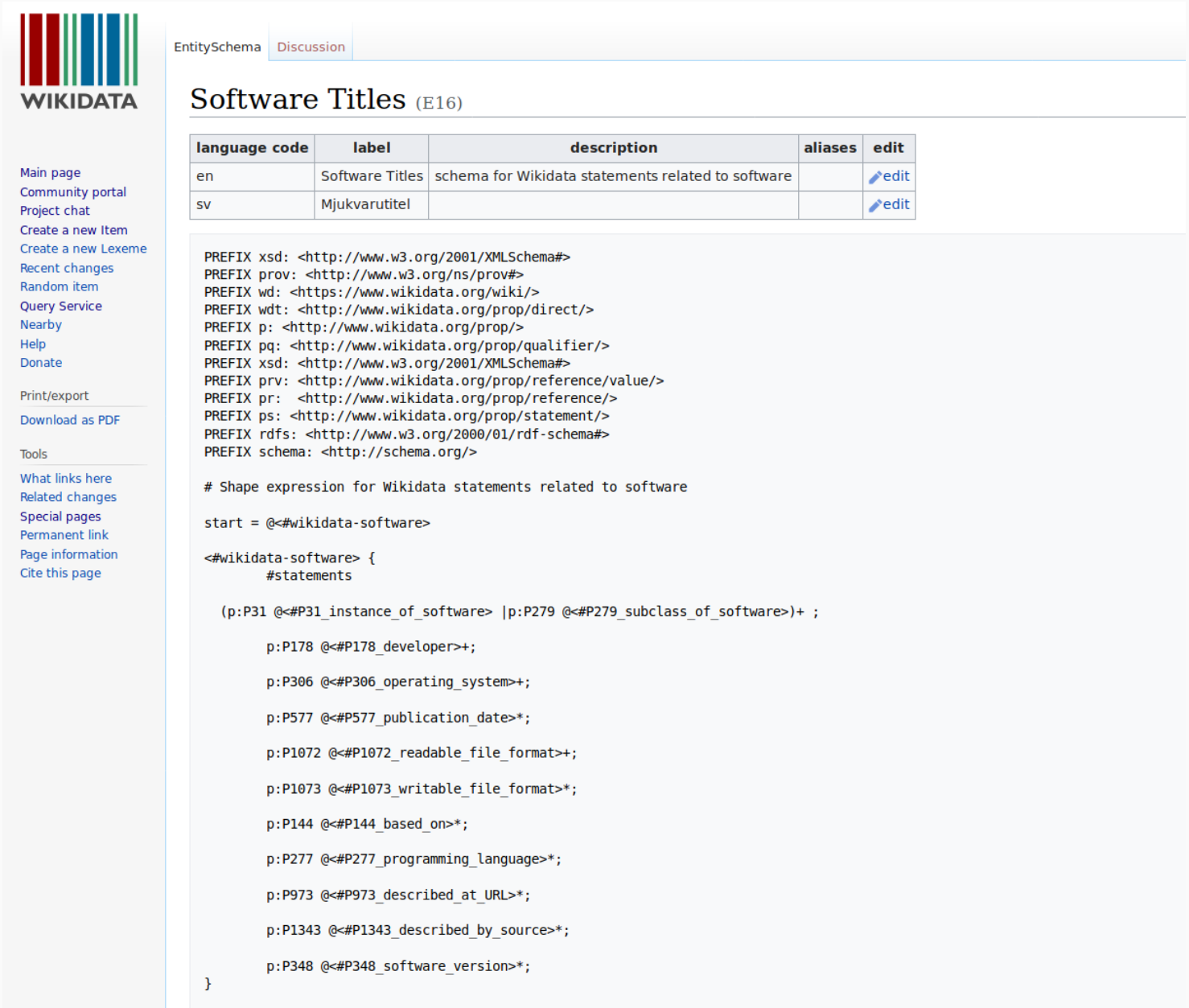
Use Cases Two and Three with Wikidata

- Wikidata does not have pre-defined data models
- ShEx schemas are used to share data models
- ShEx schemas are used to validate entity data, relevant for data quality

ShEx on Wikidata

- 28 May new namespace in Wikidata for schemas (E)
- Store schemas in Wikidata
- Use schemas for validation in a version of ShEx2 Simple Online Validator
- Discuss and extend schemas where the Wikidata community works
- Support for multilingual labels and descriptions
- Talk pages and revision history per schema

E namespace on Wikidata



WIKIDATA

Main page
Community portal
Project chat
Create a new Item
Create a new Lexeme
Recent changes
Random item
Query Service
Nearby
Help
Donate

Print/export
Download as PDF

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Cite this page

EntitySchema | Discussion

Software Titles (E16)

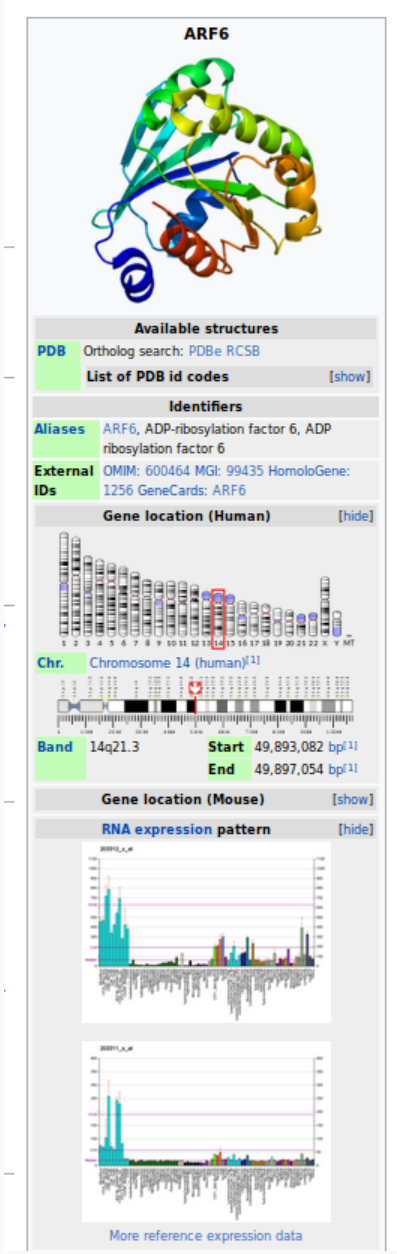
language code	label	description	aliases	edit
en	Software Titles	schema for Wikidata statements related to software		edit
sv	Mjukvarutitel			edit

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX wd: <https://www.wikidata.org/wiki/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prv: <http://www.wikidata.org/prop/reference/value/>
PREFIX pr:  <http://www.wikidata.org/prop/reference/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>

# Shape expression for Wikidata statements related to software

start = @<#wikidata-software>

<#wikidata-software> {
        #statements

  (p:P31 @<#P31_instance_of_software> |p:P279 @<#P279_subclass_of_software>)+ ;

        p:P178 @<#P178_developer>+;

        p:P306 @<#P306_operating_system>+;

        p:P577 @<#P577_publication_date>*;

        p:P1072 @<#P1072_readable_file_format>+;

        p:P1073 @<#P1073_writable_file_format>*;

        p:P144 @<#P144_based_on>*;

        p:P277 @<#P277_programming_language>*;

        p:P973 @<#P973_described_at_URL>*;

        p:P1343 @<#P1343_described_by_source>*;

        p:P348 @<#P348_software_version>*;
}
```

Use Case Two: ShEx is used in GeneWiki

- Started in 2008, goal to create and maintain **infoboxes** in English-language Wikipedia articles about **human genes**
- In 2012, shifted from curating infoboxes on Wikipedia pages to curating the corresponding items on Wikidata
- **24k** human genes and **20k** mouse genes
- **8,700** disease concepts from the **Disease Ontology**
- **2,700** FDA-approved drugs

Infobox populated by structured data from Wikidata



Validation example for GeneWiki

- Schema E37 on Wikidata

Validation in Wikidata's Toolforge implementation

ShEx2 — Simple Online Validator

```
    # that expresses the requirement that both properties are not mandatory:
        (
                p:P644 @<#P644_genomic_start> ; # Its genomic start location
                p:P645 @<#P645_genomic_end> ; # Its genomic end location
        )* ; # Zero or more start and end locations.

        p:P684 @<#P684_ortholog>* ; # Zero or more known orthologs.
        p:P688 @<#P688_encodes>* ; # Zero or more known geneproducts.
        p:P703 @<#P703_found_in_taxon_human> ; # In which taxonomy and where in that taxonomy this gene is found
        p:P1057 @<#P1057_chromosome>* ;  # Zero or more known chromosomes the gene is located on.
        p:P2888 @<#P2888_exact_match>+ ; # One or more external Internationalized Resource Identifiers of a node on the
                                         # semantic web, describing the same concept as linked data.
        p:P2548 @<#P2548_strand_orientation> ; # Its strand orrientation

        # IDENTIFIER STATEMENTS
        p:P351 @<#P351_ncbi_gene_id> ; # Exactly one ncbi gene identifier
        p:P353 @<#P353_hgnc_gene_symbol> ; # Exactly one hgnc gene symbol
        p:P354 @<#P354_hgnc_gene_id> ; # Exactly one hgnc gene identifier
        p:P594 @<#P594_ensembl_gene_id>* ; # Zero or more Ensembl gene identifier
        p:P639 @<#P639_refseq_rna_id>* ; # Zero or more RefSeq RNA identifiers
        p:P704 @<#P704_ensembl_transcript_id>* ; # Zero or more Ensembl Transcript identifiers.
        p:P593 @<#P593_homologene_id> ; # Exactly one homologene identifier

        # Negative shapes
        p:P352 @<#P352_uniprot_id_wor>{0} ; # A gene can't have a uniprot identifier.
}
```

validate (ctl-enter)

Query Map | Entities to check

```
SELECT ?item WHERE { ?item wdt:P351 ?geneid ; wdt:P703 wd:Q15978631 .} LIMIT 10
```

✓wd:Q40108@START

✗wd:Q227339@!START

```
validating http://www.wikidata.org/entity/Q227339 as //www.wikidata.org/wiki/Special:EntitySchemaText/E37#wikidata-human-gene:
    validating http://www.wikidata.org/entity/statement/Q227339-b0607ff1-4570-23cc-85b4-c2fa23121060:
      validating http://www.wikidata.org/entity/Q422445:
        NodeConstraintError: expected to match [<http://www.wikidata.org/entity/Q7187> <http://www.wikidata.org/entity/Q20747295> <http://www.wikidata.org/entity/Q427087> <http://www.wikidata.org/entity/Q284578> <http://www.wiki
```

✓wd:Q238509@START

✓wd:Q248215@START

✓wd:Q282418@START

✓wd:Q286987@START

✓wd:Q289013@START

✗wd:Q305481@!START

```
validating http://www.wikidata.org/entity/Q305481 as //www.wikidata.org/wiki/Special:EntitySchemaText/E37#wikidata-human-gene:
    validating http://www.wikidata.org/entity/statement/Q305481-22FAA446-1278-45CA-AD9C-6480517DA189:
      validating http://www.wikidata.org/reference/503d708cd6c6cc9d3acfc20457905a4eb5c8961e: exceeds cardinality
  OR
  validating http://www.wikidata.org/entity/statement/Q305481-B4865C3A-9CA2-47D5-A87C-BE92AC525F44:
    validating http://www.wikidata.org/reference/4e41b08d50ba1b8f572908a35e4a88a28e3feade:
      Missing property: http://www.wikidata.org/prop/reference/P854
      validating http://www.wikidata.org/entity/Q16335166:
        Error validating http://www.wikidata.org/entity/Q16335166 as {"type":"NodeConstraint","values":["http://www.wikidata.org/entity/Q20641742"]}: value http://www.wikidata.org/entity/Q16335166 not found in set ["http://www.w
```

✓wd:Q410688@START

✓wd:Q414043@START

Wikidata Integrator (WDI)

- WDI was created by the GeneWiki team within the Su Lab
- Used to write the bots that populate data in the biomedical domain
- PyShEx has now been combined with WDI
- Power tool for working with ShEx in the Wikidata ecosystem

Use Case Three: ShEx is used in Digital Preservation

- Many cultural heritage organizations have **legacy software** in their collections
- The software stored on CD-ROMs and floppy disks at risk of **deterioration**
- If people would like to use the software, the organization might not have **relevant computing environments**

Describing Software and File formats in Wikidata

- Creating items in Wikidata for resources in the domain of computing
- Creating machine-actionable metadata to describe configured emulated environments
- Proposing data models to be discussed with the community
- Testing entity data for conformance to schemas

Screenshot of Validation

ShEx2 — Simple Online Validator controls

```
#Shape Expression for file formats on Wikidata

#BASE <http://base.example/#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX p: <http://www.wikidata.org/prop/>
PREFIX pq: <http://www.wikidata.org/prop/qualifier/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX prv: <http://www.wikidata.org/prop/reference/value/>
PREFIX pr:  <http://www.wikidata.org/prop/reference/>
PREFIX ps: <http://www.wikidata.org/prop/statement/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX schema: <http://schema.org/>

# Shape expression for Wikidata statements related to file formats

start = @<#wikidata-file_format> # Indicates which shape to use to start iterating over the graph if none is provided.

# wikidata-file_format> is the shape for a file format data model in Wikidata. Each line between the brackets
# represents the structure that can be enforced to validate data from file format items.
#
# We distinguish between value statements, and identifier statements.
# Value statements contain either actual values, or pointers to other Wikidata items. Identifier statements capture
# external identifiers
```

```
Endpoint: https://query.wikidata.org/sparql
```

Manifest:

- Fileformats in Wikidata

Passing:

- Get 10 file formats from Wikidata
- Get 100 file formats from Wikidata
- Get all file formats from Wikidata

validate (ctl-enter)

Query Map | Query Map Editor | Fixed Map

```
SPARQL '''PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

SELECT ?item WHERE { ?item wdt:P31 wd:Q235557 .} LIMIT 10'''@START
```

✓wd:Q684554@START

✓wd:Q691652@START

✓wd:Q719519@START

✓wd:Q722609@START

✓wd:Q723030@START

✓wd:Q726218@START

✓wd:Q731135@START

✓wd:Q737207@START

✓wd:Q741654@START

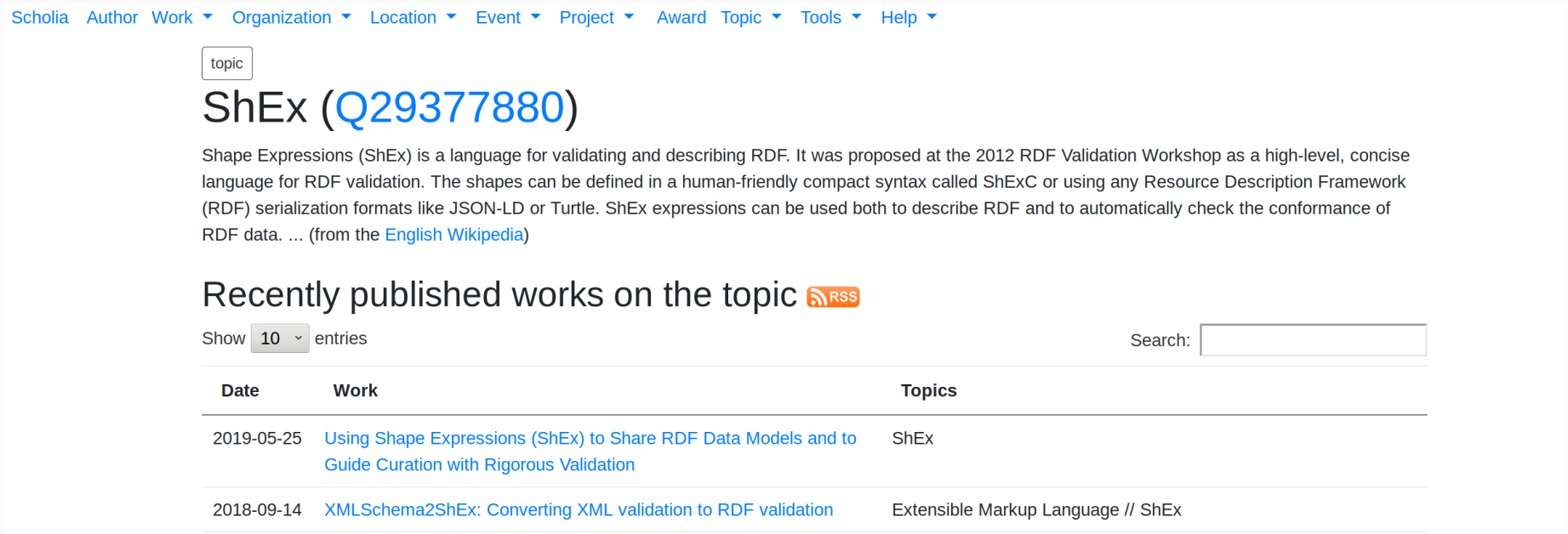
✓wd:Q743275@START

Why ShEx?

- Implementations in multiple languages under open source licenses
- Concise, human-readable syntax
- Comprehensive test suite

Find ShEx Papers

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



Quick Start

- Primer: http://shex.io/shex-primer/

Getting started with ShEx on Wikidata

- Wikidata Wikiproject ShEx
- Schema namespace

Join the conversation

- ShEx Gitter

ShEx Community Group

- W3C Community Group
- You are welcome to join if ShEx is of interest to you

Questions?