



Database Validation: Interesting facet of counter-vandalism

Case Study of Wikidata

Houcemeddine Turki, *University of Sfax, Tunisia*





Database Validation

A process to ensure data quality of a given semantic resource by managing competency questions

- Accuracy: Verification of whether the definitions, classes, properties and individual entries in the assessed resource are correct or not
- Completeness: Coverage of a given knowledge domain in the evaluated resource
- Adaptability: Range of different anticipated uses of the evaluated resource
- Clarity: Effectiveness of communication of intended meanings of defined items by the assessed resource

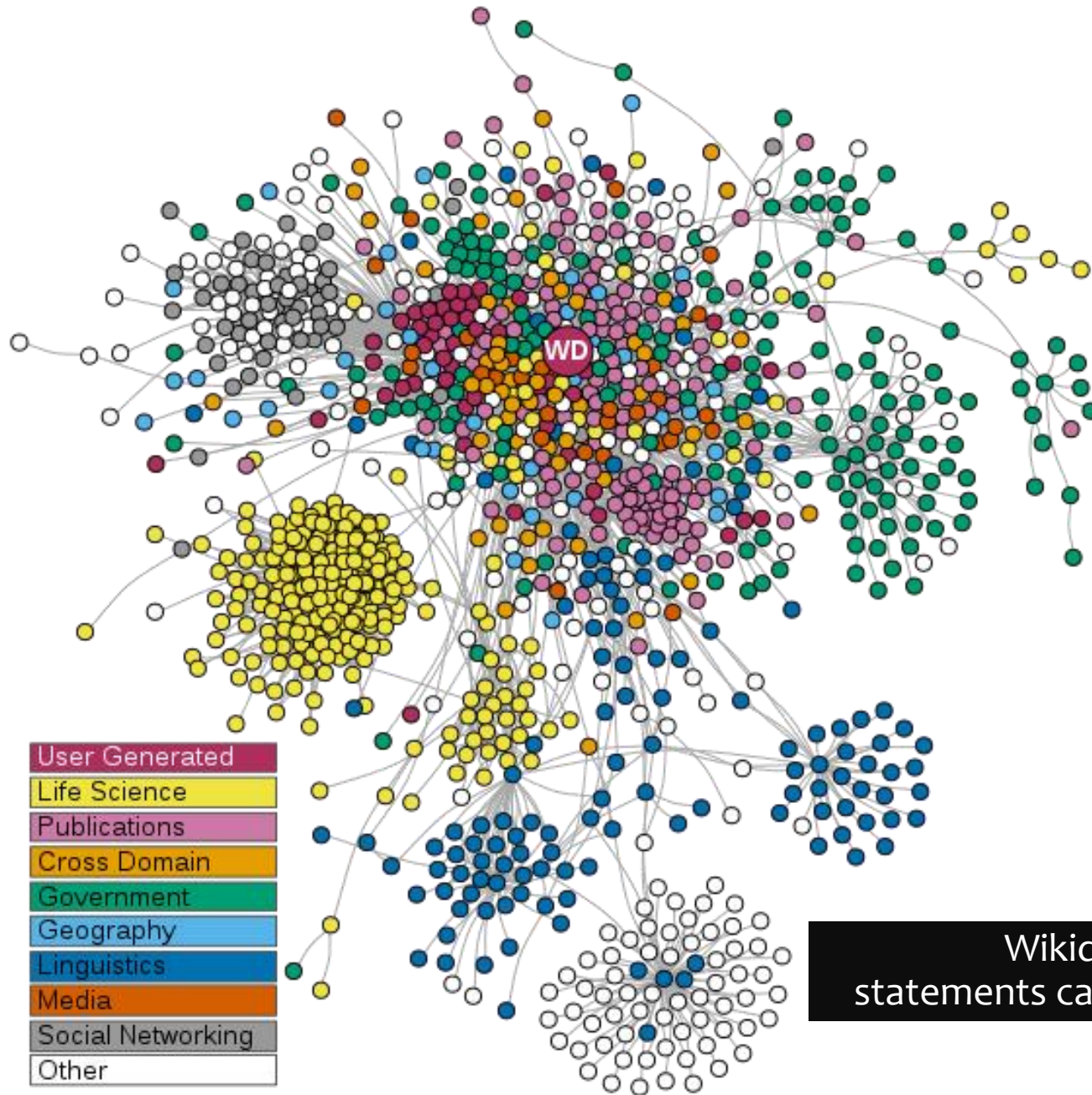
Main Purpose

Data Quality

- Ensuring an homogenous and exhaustive representation of structured information
- Eliminating redundancies and logical inconsistencies
- Verifying the accuracy of structured data
- Enhancing the reliability and trustworthiness of structured information

Counter-vandalism

- Ensuring that users do not change more accurate information by less accurate data
- Can also be evaluated by analyzing user behaviors and patrolling edits from non confirmed editors: <https://www.wikidata.org/wiki/Wikidata:Patrol>
- Statistics: <https://www.wikidata.org/wiki/User:Pyfisch/Counter-Vandalism>
- Tools for patrolling edits: <https://wdvd.toolforge.org/>, <https://pltools.toolforge.org/rech/>, and <https://speedpatrolling.toolforge.org/>
- Further information can be found at https://www.wikidata.org/wiki/Wikidata:WikiProject_Counter-Vandalism



[Toward a complete dataset of drug–drug interaction information from publicly available sources](#)

S Ayvaz, J Horn, O Hassanzadeh, Q Zhu, J Stan, NP Tatonetti, S Vilar, ...
Journal of biomedical informatics 55, 206-217

[Science Forum: Wikidata as a knowledge graph for the life sciences](#)

A Waagmeester, G Stupp, S Burgstaller-Muehlbacher, BM Good, M Griffith, ...
ELife 9, e52614

[Wikidata as a semantic framework for the Gene Wiki initiative](#)

S Burgstaller-Muehlbacher, A Waagmeester, E Mitraka, J Turner, ...
Database 2016

[WikiGenomes: an open web application for community consumption and curation of gene annotation data in Wikidata](#)

TE Putman, S Lelong, S Burgstaller-Muehlbacher, A Waagmeester, ...
Database 2017







[Wikidata: A large-scale collaborative ontological medical database](#)

H Turki, T Shafee, MA Hadj Taieb, M Ben Aouicha, D Vrandečić, D Das, ...
Journal of Biomedical Informatics 99, 103292

Wikidata is linked to numerous external resources: Wikidata statements can be verified against information in external databases

Property statements and constraints

A set of semantic information defining the format of the statements using a given Wikidata property


subject item of this property	 affiliation ▾ 0 references
Wikidata property example	 Raoul Bott affiliation Institute for Advanced Study ▾ 0 references
subproperty of	 use ▾ 0 references
Wikidata property example	 dabrafenib medical condition treated skin melanoma ▾ 0 references
equivalent property	 http://purl.obolibrary.org/obo/RO_0002599 ▾ 0 references
inverse property	 drug used for treatment ▾ 0 references





Collaborative Approach to Developing a Multilingual Ontology: A Case Study of Wikidata

Authors

Authors and affiliations

John Samuel 

property constraint	 value type constraint					
	<table border="0"> <tr> <td>class</td> <td> <ul style="list-style-type: none"> organization fictional organization project fictional municipal police fictional gang institution institute Christian ministry church building place of worship fictional place of worship Christian movement religious denomination </td> </tr> <tr> <td>relation</td> <td>instance of</td> </tr> <tr> <td colspan="2">▾ 0 references</td> </tr> </table>	class	<ul style="list-style-type: none"> organization fictional organization project fictional municipal police fictional gang institution institute Christian ministry church building place of worship fictional place of worship Christian movement religious denomination 	relation	instance of	▾ 0 references
class	<ul style="list-style-type: none"> organization fictional organization project fictional municipal police fictional gang institution institute Christian ministry church building place of worship fictional place of worship Christian movement religious denomination 					
relation	instance of					
▾ 0 references						
	 type constraint					
	<table border="0"> <tr> <td>class</td> <td> <ul style="list-style-type: none"> human group of humans fictional organization fictional character organization </td> </tr> <tr> <td>relation</td> <td>instance of</td> </tr> <tr> <td colspan="2">▾ 0 references</td> </tr> </table>	class	<ul style="list-style-type: none"> human group of humans fictional organization fictional character organization 	relation	instance of	▾ 0 references
class	<ul style="list-style-type: none"> human group of humans fictional organization fictional character organization 					
relation	instance of					
▾ 0 references						

symptoms



temporary blindness

edit



Potential issues ✕

type constraint

[Help](#) [Discuss](#)

Entities using the **symptoms** property should be instances or subclasses of **physiological condition** or **fictional medical condition** (or of a subclass of them), but **Flash blindness** currently isn't.

A notification appears to Wikidata users when a given property constraint is violated

Recoin

A tool to identify missing statements for a given item by comparing it to its class members

2020 COVID-19 pandemic in Tunisia (Q87343682)...

viral outbreak in Tunisia

2020 coronavirus outbreak in Tunisia

Enter a schema to check against e.g. E10

▼ Recoin: Most relevant properties which are absent

Property ID	Label	Relative	Add Claim
P527	has part	7.21%	+
P8045	organized response related to outbreak	5.22%	+
P18	image	2.5%	
P5008	on focus list of Wikimedia project	2.44%	+
P1343	described by source	1.81%	+
P131	located in the administrative territorial entity	0.99%	+
P585	point in time	0.87%	
P582	end time	0.59%	
P8204	tabular case data	0.55%	
P1424	topic's main template	0.51%	+
P1001	applies to iurisdiction	0.49%	+

Recoin: Relative Completeness in Wikidata



Authors: Vevake Balaraman, Simon Razniewski, Werner Nutt [Authors Info & Affiliations](#)

Publication: WWW '18: Companion Proceedings of the The Web Conference 2018 • April 2018 • Pages 1787–1792 • <https://doi.org/10.1145/3184558.3191641>

- Identifies the Wikidata items that are not significantly described and that are consequently likely to be created due to vandalism.
- Can help identifying properties that are not commonly used for the members of a given class.

Shape Expressions (ShEx)

Structural schema language to define the format of the members of a given Wikidata class

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

start = @<app>


<app> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q90790055 # instance of COVID-19 dashboard or
           wd:Q91136116 # search engine or
           wd:Q91137337 # dataset
         ] ;
  wdt:P1476 LITERAL * ; #title
  wdt:P366 . * ; #use
  wdt:P123 . * ; #publisher
  wdt:P178 . * ; #developers
  wdt:P495 . * ; #country of origin
  wdt:P306 . * ; #operating system
  wdt:P856 . * ; #official website
  wdt:P921 . * ; #main subject
  wdt:P144 . * ; #based on
  wdt:P577 . ? ; #publication date
  wdt:P7103 . ? ; #start of covered period
  wdt:P275 . * ; #copyright license
  wdt:P5008 . * ; #on focus list of Wikimedia project
}
```



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

Authors

[Authors and affiliations](#)

Katherine Thornton , Harold Solbrig, Gregory S. Stupp, Jose Emilio Labra Gayo, Daniel Mietchen, Eric Prud'hommeaux,

Andra Waagmeester

Key	Meaning
wdt:<PropertyID>	Defined property
wd:<ItemID>	Defined item
.	Object
*	Zero or more
?	Zero or one
+	One or more
LITERAL	Monolingual text
EXTRA	Object is one value from of a defined list

<https://www.wikidata.org/wiki/EntitySchema:E205>

Shape Expressions (ShEx)


Structural schema language to define the format of the members of a given Wikidata class



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

Authors

[Authors and affiliations](#)

Katherine Thornton , Harold Solbrig, Gregory S. Stupp, Jose Emilio Labra Gayo, Daniel Mietchen, Eric Prud'hommeaux,

Andra Waagmeester

Shape Expression for class [\[edit\]](#)

Originally proposed at [Wikidata:Property proposal/Generic](#)

	On hold
Description	Shape Expression that members of a class should conform to
Represents	Shape Expressions (Q29377880)
Data type	<datatype-type-EntitySchema> (not available yet)
Domain	class
Example 1	human (Q5) → E10
Example 2	film festival (Q220505) → E11
Example 3	film festival edition (Q27787439) → E12
Example 4	natural number (Q21199) → E13

Motivation [\[edit\]](#)

Property to link a class to the Shape Expression that members of it should conform to.

This will make it easier to query for Shape Expressions that exist, and quickly see what has been defined for a particular class. [Jheald](#) (talk) 16:56, 28 May 2019 (UTC)

Tracked in [Phabricator](#)
[Task T214884](#)

Note: Implementation will require EntitySchema to be added to the set of data-types that can be values for Wikidata statements. There is a ticket for this on Phabricator, which Léa hopes should be resolved in the coming weeks.[\[1\]](#)[↗](#). [Jheald](#) (talk) 07:54, 29 May 2019 (UTC)

- A property to link EntitySchemas to Wikidata classes is currently on hold
- There is no need for this property if a script can be built to automatically validate Wikidata items against concerned EntitySchemas

Shape Expressions (ShEx)

Structural schema language to define the format of the members of a given Wikidata class

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>

start = @<app>


<app> EXTRA wdt:P31 {
  wdt:P31 [ wd:Q90790055 # instance of COVID-19 dashboard or
           wd:Q91136116 # search engine or
           wd:Q91137337 # dataset
         ] ;
  wdt:P1476 LITERAL * ; #title
  wdt:P366 . * ; #use
  wdt:P123 . * ; #publisher
  wdt:P178 . * ; #developers
  wdt:P495 . * ; #country of origin
  wdt:P306 . * ; #operating system
  wdt:P856 . * ; #official website
  wdt:P921 . * ; #main subject
  wdt:P144 . * ; #based on
  wdt:P577 . ? ; #publication date
  wdt:P7103 . ? ; #start of covered period
  wdt:P275 . * ; #copyright license
  wdt:P5008 . * ; #on focus list of Wikimedia project
}
```



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

Authors

[Authors and affiliations](#)

Katherine Thornton , Harold Solbrig, Gregory S. Stupp, Jose Emilio Labra Gayo, Daniel Mietchen, Eric Prud'hommeaux,

Andra Waagmeester

Interesting hint: ShEx statements where the object is a defined Wikidata item (Not ?, * or +) can be used to define the condition of the application of the EntitySchema

<https://www.wikidata.org/wiki/EntitySchema:E205>

Consistency rules

Conditions allowing the identification of data inconsistencies through the comparison of Wikidata statements



Using logical constraints to validate information in collaborative knowledge graphs: a study of COVID-19 on Wikidata

Houcemeddine Turki; Dariusz Jemielniak; Mohamed Ali Hadj Taieb; Jose Emilio Labra Gayo; Mohamed Ben Aouicha; Mus'ab Banat; Thomas Shafee; Eric Prud'Hommeaux; Tiago Lubiana; Diptanshu Das; Daniel Mietchen

Example 1: For a given disease, the number of cases (P1603) in day Z should be inferior or equal to the one in day Z+1.

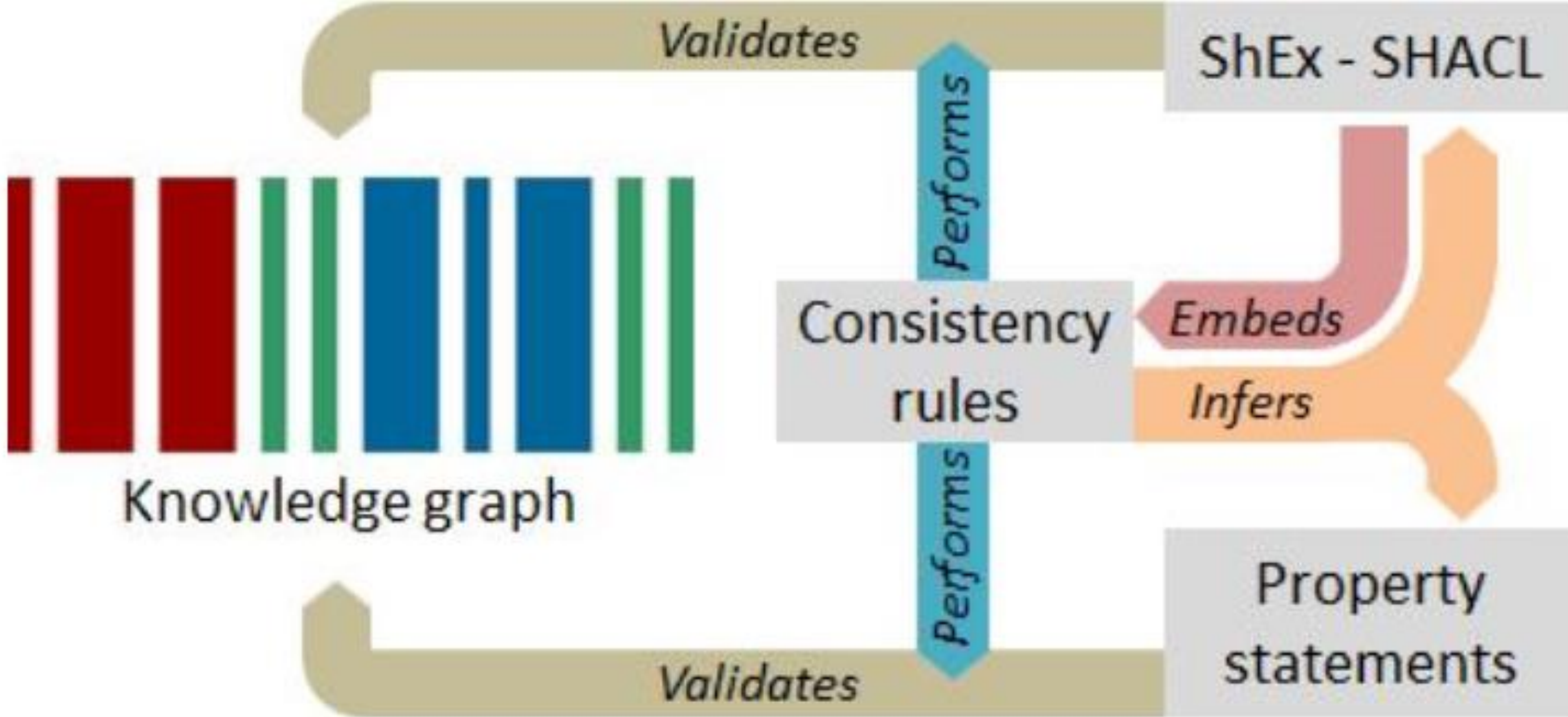
Example 2: For a given disease, the number of cases (P1603) in a given continent for day Z should be equal to the sum of the number of cases in the sovereign states in that continent in the same day Z.

Example 3: For a given disease, the number of deaths (P1120) and the number of recoveries (P8010) in a given location in day Z should be inferior or equal to the number of cases (P1603) in that location in the same day Z.

Example 4: If X is an instance of disease (P31 Q12136) and Y is the drug used for treatment (P2176) of X, Y should be an instance of a drug (P31 Q12140).

Example 5: If X is a drug that has dyspnea (Q188008) as a side effect (P1909), X cannot be the drug used for treatment (P2176) of asthma (Q35869) as well as of COPD (Q199804).

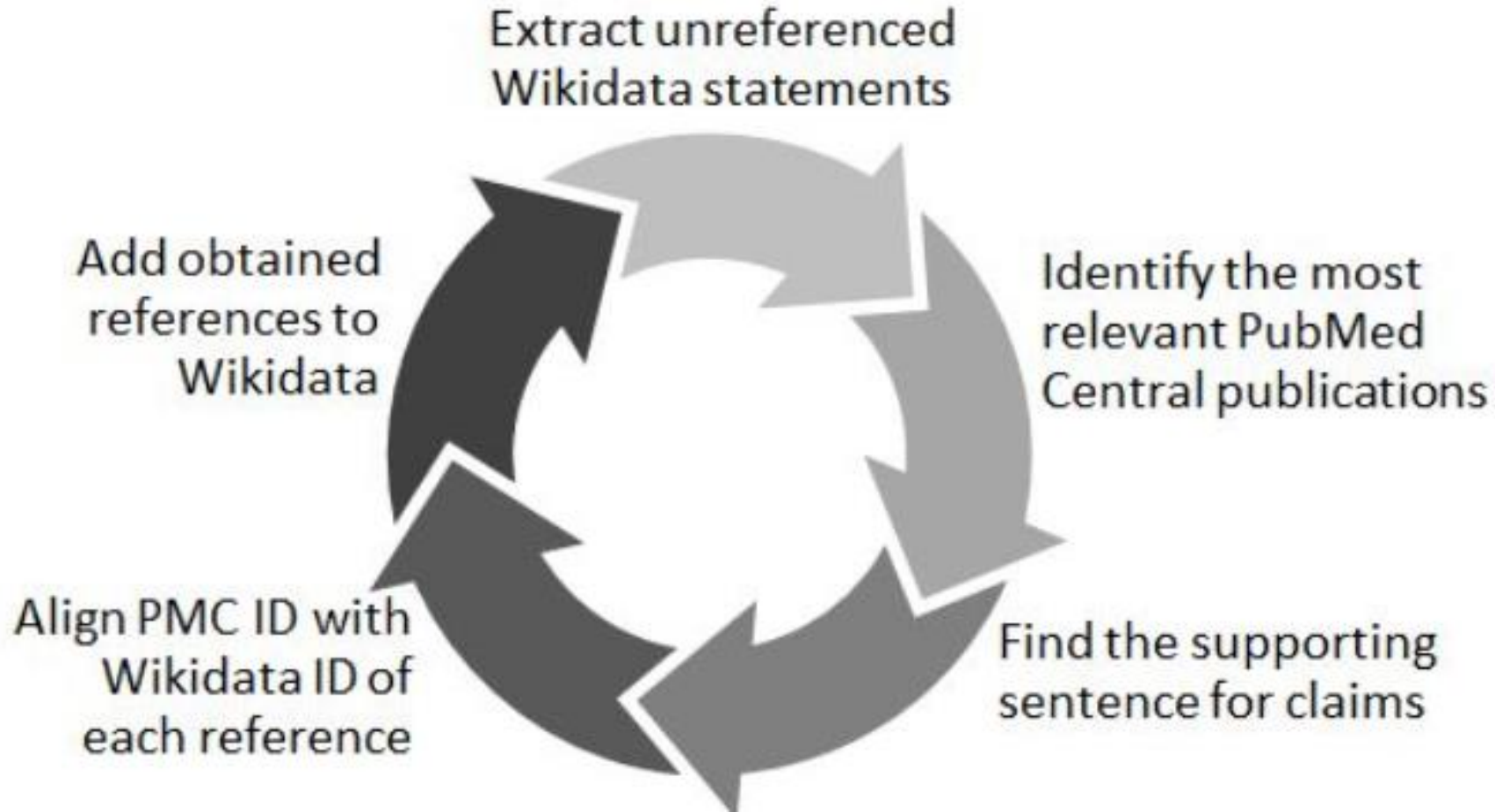
- Cannot be verified using Property Constraints or EntitySchemas
- Can be significantly implemented using SPARQL



A framework where consistency rules, property constraints and RDF validation schemas interact to validate semantic information will enhance Wikidata data quality

Reference bots

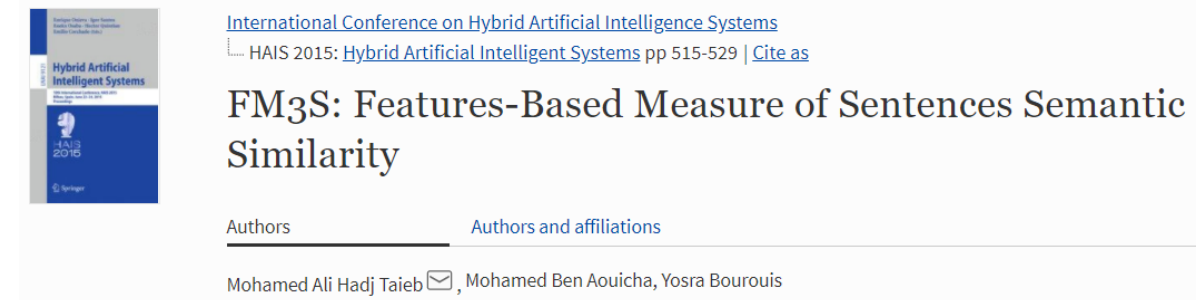
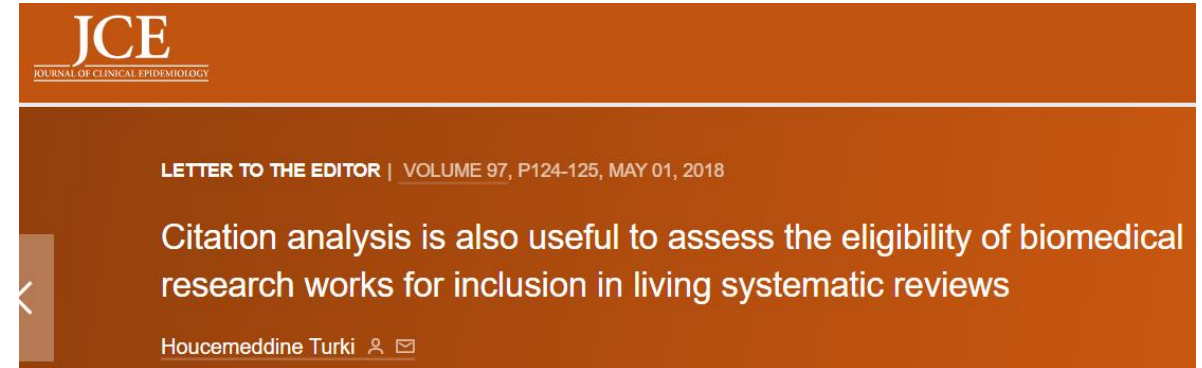
A bot that can add references to Wikidata statements from scholarly databases



- When such algorithms cannot find a reference for a given statement, this statement is likely to be wrong
- When such algorithms find a reference for a given statement, this statement is likely to be correct

Bibliometric-Enhanced Validation of Wikidata statements

- Citation-Based Validation of main subject (P921) statements:
 - When a paper does not cite (P2860) a work linked to the same topic, the statement is likely to be wrong
 - When a paper is not co-cited with another paper linked to the same topic, the statement is likely to be wrong
 - When a paper is not cited by a work linked to the same topic, the statement is likely to be wrong
- Semantic-Based Validation of main subject (P921) statements:
 - Sentence-level semantic similarity measures compute the level of similarity between sentences based on the characteristics of an is-a taxonomy.
 - ‘Subclass of (P279)’, ‘Part of (P361)’ or ‘Instance of (P31)’ Wikidata taxonomy can be used as a reference resource for this computation
 - If the semantic similarity between the titles of two research papers represented in Wikidata is limited, the two scholarly publications are not likely to evocate the same research topic.





Thank You

Houcemeddine Turki
turkiabdelwaheb@hotmail.fr
+21629499418

