Data Quality on Wikidata

how to measure and improve it

Data Quality Dimensions

Category	Dimension	Subdimension	Question
Intrinsic (evaluation does not depend on use case)	Accuracy	-	Is the data accepted as true and free of error?
	Objectivity	-	Is the data free of bias and impartial?
	Reputation	-	Are the references trustworthy?
	Consistency	-	Is the data consistently modelled?

Category	Dimension	Subdimension	Question
Contextual (evaluation does depend on use case)	Timeliness	-	Is the data up-to-date?
	Completeness	Schema	Are there enough classes and properties?
		Item	Does this Item have all relevant statements, labels, etc?
		Population	Are all instances of this class present?

Category	Dimension	Subdimension	Question
Representational	Interpretability	-	Can this data be interpreted by a machine without ambiguity?
	Ease of understanding	-	Can this data be interpreted by a human without ambiguity?
Accessibility	Interlinking	-	Is this data sufficiently interlinked to other resources?

Item Quality Evaluator

dog (Q144) •••

ORES predicted quality: A (4.63)

domestic animal

domestic dog | Canis lupus familiaris | Canis familiaris | dogs | 🐶 | 🐕

▸ In more languages

Statements

instance of	organisms known by a particular common name •••
	of: Canis familiaris •••
	of: Canis lupus familiaris •••
	▾ 0 references
subclass of	domesticated mammal •••
	▾ 0 references
	pet •••
	▾ 0 references

No label defined (Q35961733) •••

ORES predicted quality: E (1.07)

Wikimedia category

▾ In more languages

Configure

Language	Label	Description	Also known as
English	No label defined	Wikimedia category	
German	No label defined	Wikimedia-Kategorie	
French	Catégorie:Mots en russe issus d'un mot en catalan	page de catégorie de Wikimedia	
Bavarian	No label defined	Wikimedia-Kategorie	

All entered languages

Statements

instance of	Wikimedia category •••	edit
	▾ 0 references	+ add reference
		+ add value

+ add statement

DEMO

https://item-quality-evaluator.toolforge.org/

Sample items for demo -
Random subset of Programming Language - sample of items - 400 Items

1 demo - provide item id list and get
Sparl query - demo (cat)
https://docs.google.com/document/d/1RsG4cEltaUBDPozPFF2814w-b1_80B-QLVBa_SSeV5s/edit

https://item-quality-evaluator.toolforge.org/

How to improve the ORES scores?

- Adding labels in more languages
- Adding descriptions in more languages
- Adding references to statements that lack references
- Replace references to other Wikimedia projects with more appropriate references
- Adding more statements
- Adding images
- Adding external IDs

Constraints Violation Checker

https://github.com/wmde/wikidata-constraints-violation-checker

Ex: Random paintings

input file (.txt) and (.csv) format

```
item
Q20489864
Q20490034
Q20490152
Q20490240
Q20490399
Q20490501
Q20490510
Q20490535
Q20490718
Q20490869
Q20490984
Q20490997
Q20491014
Q20491064
Q20491445
Q20491466
Q20491504
Q20491513
Q20491565
Q20491978
Q20496930
Q20496942
Q20498926
Q20504591
Q20504792
Q20505137
Q20505968
```

	A	B	C
1	item		
2	Q20489864		
3	Q20490034		
4	Q20490152		
5	Q20490240		
6	Q20490399		
7	Q20490501		
8	Q20490510		
9	Q20490535		
10	Q20490718		
11	Q20490869		
12	Q20490984		
13	Q20490997		
14	Q20491014		
15	Q20491064		
16	Q20491445		
17	Q20491466		
18	Q20491504		
19	Q20491513		
20	Q20491565		
21	Q20491978		
22	Q20496930		
23	Q20496942		
24	Q20498926		
25	Q20504591		
26	Q20504792		
27	Q20505137		
28	Q20505968		
29	Q20506331		

Command Line

```
QID;statements;violations_mandatory_level;violations_normal_level;violations_suggestion_level;violated_statements;total_sitelinks;wikipedia_sitelinks;ores_score
Q20489864;11;0;1;0;1;0;0;D
Q20490034;11;0;1;0;1;0;0;D
Q20490152;11;0;1;0;1;0;0;D
Q20490240;11;0;1;0;1;0;0;D
Q20490399;11;0;1;0;1;0;0;D
Q20490501;11;0;1;0;1;0;0;D
Q20490510;11;0;1;0;1;0;0;D
Q20490535;11;0;1;0;1;0;0;D
Q20490718;13;0;1;0;1;0;0;D
Q20490869;17;0;1;0;1;0;0;D
Q20490984;19;0;1;0;1;0;0;D
Q20490997;11;0;1;0;1;0;0;D
Q20491014;13;0;1;0;1;0;0;D
Q20491064;13;0;1;0;1;0;0;D
Q20491445;11;0;1;0;1;0;0;D
Q20491466;12;0;1;0;1;0;0;D
Q20491504;11;0;1;0;1;0;0;D
Q20491513;11;0;1;0;1;0;0;D
Q20491565;8;0;1;0;1;0;0;D
```

	A	B	C	D	E	F	G	H	I
1	QID	statements	violations_mandatory_level	violations_normal_level	violations_suggestion_level	violated_statements	total_sitelinks	wikipedia_sitelinks	ores_score
2	Q20489864	11	0	1	0	1	0	0	D
3	Q20490034	11	0	1	0	1	0	0	D
4	Q20490152	11	0	1	0	1	0	0	D
5	Q20490240	11	0	1	0	1	0	0	D
6	Q20490399	11	0	1	0	1	0	0	D
7	Q20490501	11	0	1	0	1	0	0	D
8	Q20490510	11	0	1	0	1	0	0	D
9	Q20490535	11	0	1	0	1	0	0	D
10	Q20490718	13	0	1	0	1	0	0	D
11	Q20490869	17	0	1	0	1	0	0	D
12	Q20490984	19	0	1	0	1	0	0	D
13	Q20490997	11	0	1	0	1	0	0	D
14	Q20491014	13	0	1	0	1	0	0	D
15	Q20491064	13	0	1	0	1	0	0	D
16	Q20491445	11	0	1	0	1	0	0	D
17	Q20491466	12	0	1	0	1	0	0	D
18	Q20491504	11	0	1	0	1	0	0	D
19	Q20491513	11	0	1	0	1	0	0	D
20	Q20491565	8	0	1	0	1	0	0	D
21	Q20491978	14	0	1	0	1	0	0	D
22	Q20496930	10	0	1	0	1	0	0	D
23	Q20496942	13	0	0	3	3	0	0	C
24	Q20498926	16	0	1	0	1	0	0	D

output file (.txt) and (.csv) format

What’s coming next?

Checker Against External Databases

Performs checks on mismatching data between Wikidata and external databases. How does this work?

For which Items should the checks be perfomed?

Item list | SPARQL

Q184746

One Item-Identifier (for example: Q1234) on each line

Check items

Checker Against External Databases

Performs checks on mismatching data between Wikidata and external databases. How does this work?

For which Items should the checks be perfomed?

Item list | SPARQL

Q184746
Q45901

One Item-Identifier (for example: Q1234) on each line

Check items

These are the results for the item Jane Goodall (Q184746).

#	Property	Value on Wikidata	Value on External Database	Sources
1	Date of birth	4 September 1990	3 April 1914	Integrated Authority File
2	Place of birth	Hampstead, UK	New York, USA	VIAF
3	Sex or gender	female	male	BnF Authorities
4	Spouse	Hugo van Lawick	Sean Penn	VIAF
5	Language Spoken	English	Japanese	BnF Authorities

The checks on this item were performed last time: 2 weeks ago. For more information on the frequency of the check, please visit here.

Thank you!

See you on Wikidata

Wikidata.org

@wikidata, @nightrose

lydia.pintscher@wikimedia.de

amrutha.chandra@wikimedia.de

Appendix

How to enable ORES score for your Wikidata account ?

- Replace your username in this link https://www.wikidata.org/wiki/User:Lydia_Pintscher_(WMDE)/common.js
- Go to the link and add this line to the page as shown - ***importScript("User:EpochFail/ArticleQuality.js")***
- Go to any random Item , and now you can see the ORES predicted quality

Appendix

- https://www.mediawiki.org/wiki/ORES
- https://blog.wikimedia.de/2016/01/02/teaching-machines-to-make-your-life-easier-quality-work-on-wikidata/
- https://item-quality-evaluator.toolforge.org/