



Assessing the quality of sources in Wikidata across languages

Gabriel Maia Rocha Amaral



Cross-lingual Event-centric
Open Analytics Research Academy



Who am I?

- Gabriel Amaral
- BSc in Computer Sciences, Federal University of Ceará, Brazil
- Ph.D. candidate at King's College London
- Fellow in the Marie Curie European training network Cleopatra



**Why bother with Wikidata
reference quality?**



Wikidata

- Free knowledge base of structured data
- Built by a worldwide community of **volunteers**
- **Secondary** source of information: contents **should** be backed by high-quality references
- Does not ask claims to be true, only **verifiable**



+97M items	+1.38B statements
+10K properties	+23k active editors

Label — **Diego Velázquez** (Q297) — **Item identifier**

Description — Spanish painter (1599-1660)
Velázquez | Diego Rodríguez de Silva y Velázquez | Diego Rodriguez de Silva y Velázquez — **Aliases**
▶ [In more languages](#)

Statements

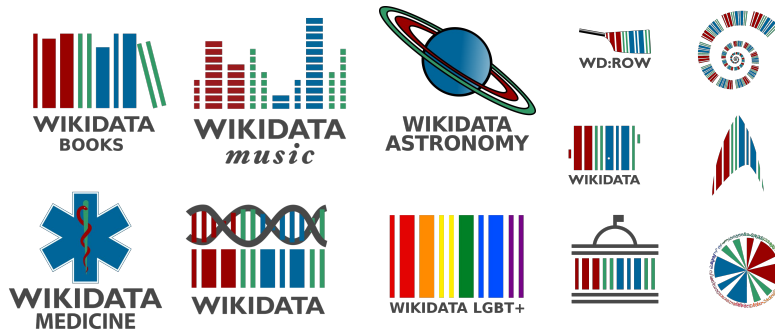
Property — **date of birth** — **6 June 1599** *Gregorian* — **Value**

Rank — **reference URL** — <http://engage.pcad.edu/blog/artist-spotlight-diego-vel%C3%A1zquez> — **Reference**

Statement group — **stated in** — **Spanish Biographical Dictionary** — **Reference**
Spanish Biographical Dictionary ID — [5114/diego-rodriguez-de-silva-y-velazquez](#)
named as — **Diego Rodríguez de Silva y Velázquez**
retrieved — **9 October 2017**

Why worry about its quality?

- Explosive growth in data
- Multiple agents using this data
- Voluntary editors
- Limited automated quality assessment
 - No reliable stats on references!
- Trustworthiness depends on references

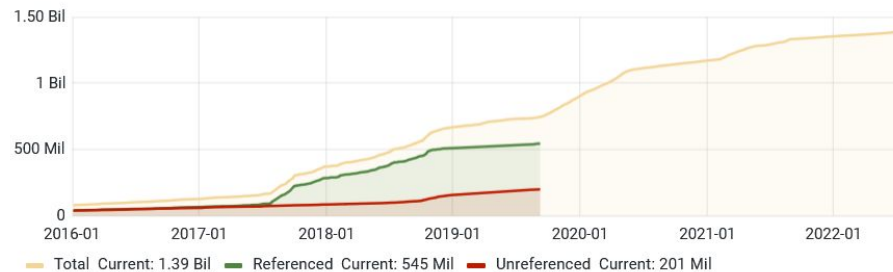


Why worry about its quality?

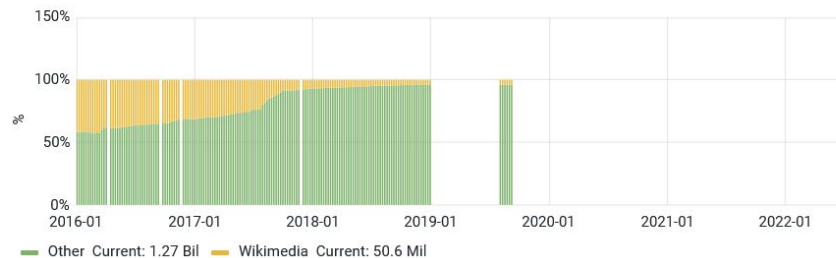
Item Statements



Statement Breakdown



Reference Main Snak Types



Defining and assessing quality in Wikidata



Wikidata's fit for use

- Focus on **references**, not on **truth**
- Taken from Wikidata's [verifiability](#) guidelines
- With some exceptions, statements should have references that:
 - Are **relevant**
 - Have **authoritative** sources
 - Are **easy to access** by some users
 - Note: Different from accessible



Quality dimensions

- **Relevance:** Relevant references support their claims
- **Authoritative:** Considered by Wikipedia's guidelines to commonly provide reliable information
 - Type (URL)
 - Author
 - Publisher
- **Easy to access:** Users can access and understand relevant information with small perceived effort



A hybrid assessment workflow

- Study done at scale
 - 95% CI, 5% margin of error
- In different languages
 - English, Portuguese, Spanish, Japanese, Dutch, and Swedish
- With different types of references
 - Internal and External

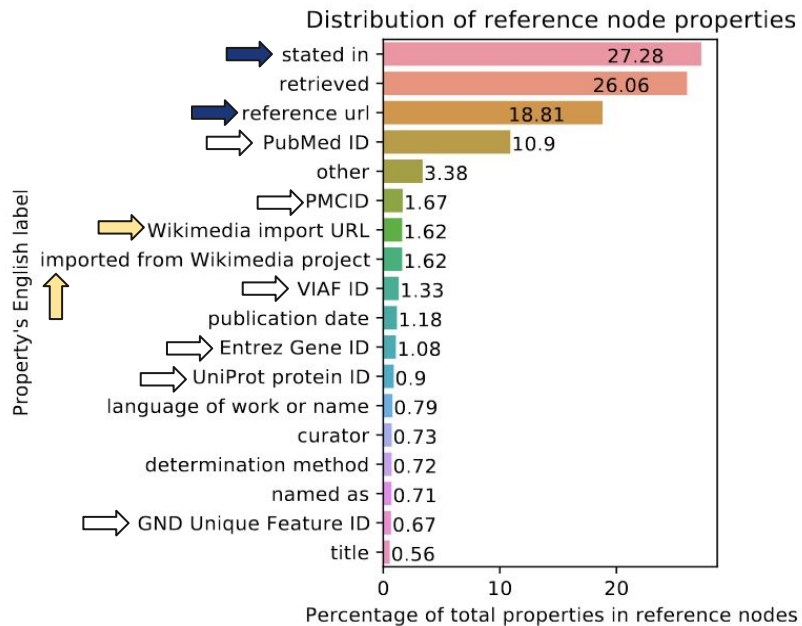
▼ 11 references

reference URL	http://engage.pcad.edu/blog/artist-spotlight-diego-vel%C3%A1zquez
---------------	---

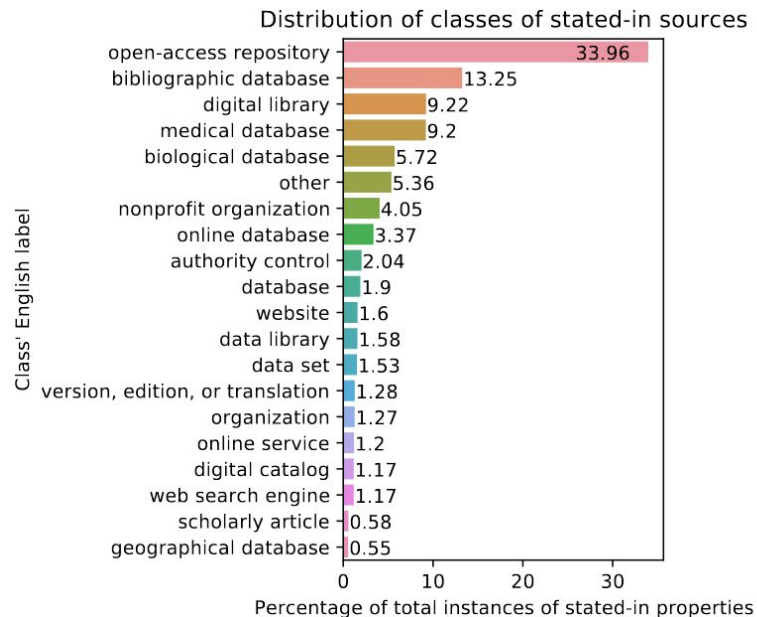
stated in	Spanish Biographical Dictionary
Spanish Biographical Dictionary ID	5114/diego-rodriguez-de-silva-y-velazquez
named as	Diego Rodríguez de Silva y Velázquez
retrieved	9 October 2017

Descriptive Statistics

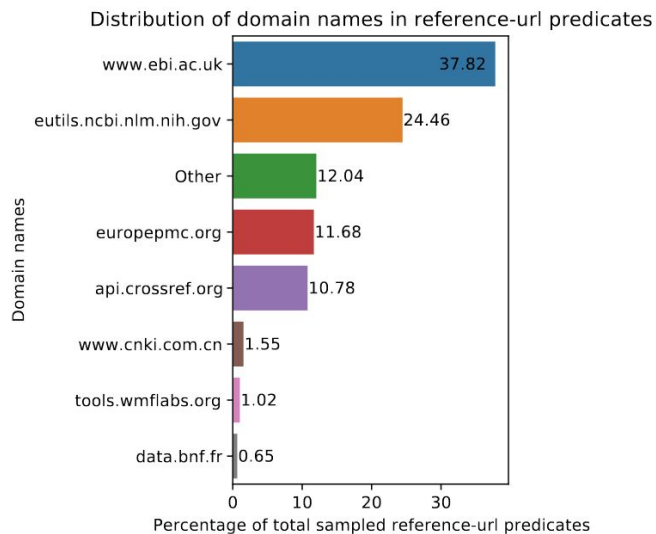
Reference encoding



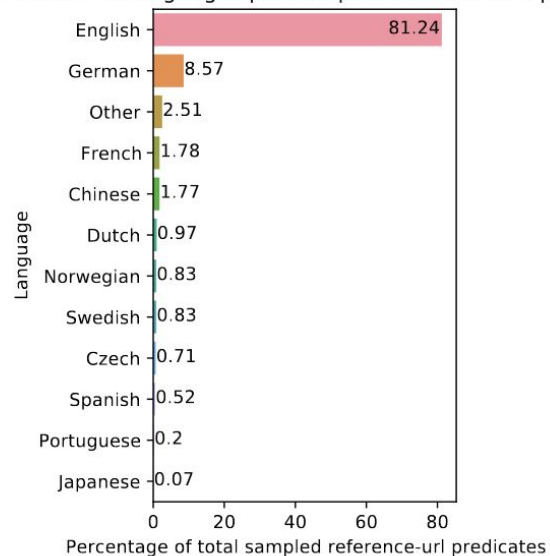
“Stated in” predicate



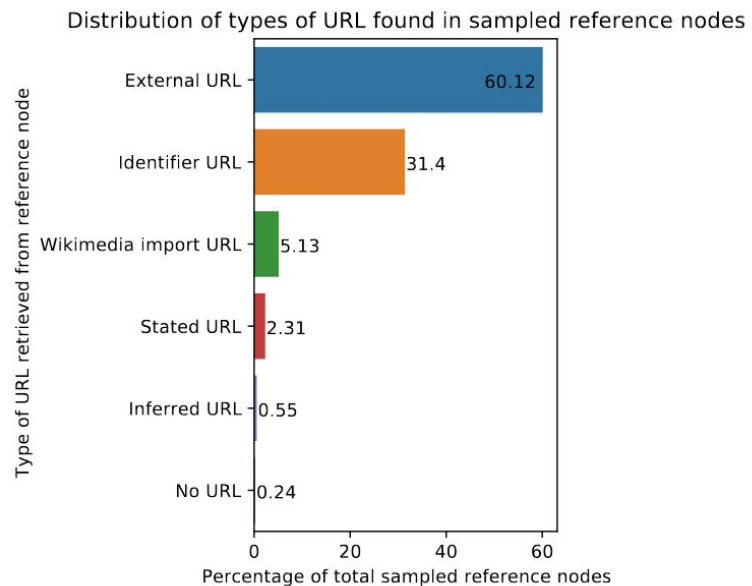
Reference URL



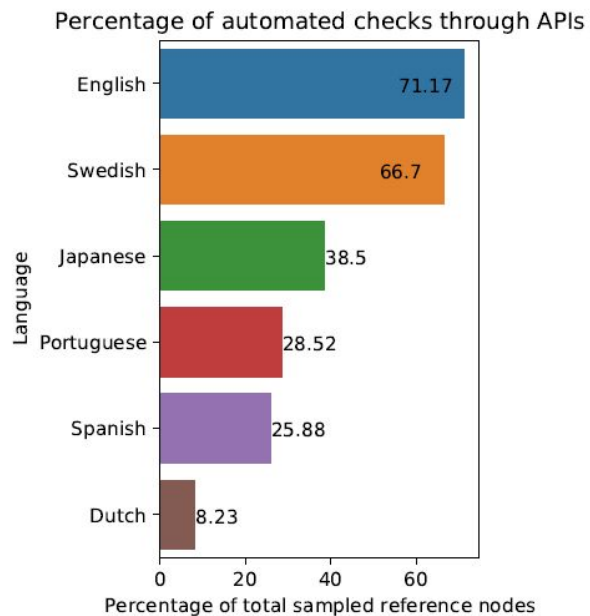
Distribution of languages per sampled reference-url predicate



Reference type



API automated checks




Results



Annotations

- Relevance (Yes/No)
- Ease of access
 - Ease of navigation (0 to 4)
 - Barriers (6)
- Authoritativeness
 - Author type
 - Publisher type
 - Authoritativeness
(Yes/No/Inaccessible)

Aggregated results



	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

T1.3 Barriers:

- 0. Page not available
- 1. Security issues
- 2. Credentials needed
- 3. Paywall
- 4. Domain Knowledge
- 5. Subject or predicate not mentioned
- 6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

References are around 90% relevant

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

Most languages are fairly (3/4) to very (4/4) easy to access

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

Except for English

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

Japanese references had a huge problem with
Yahoo Japan being out of service

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

Most barriers do not occur anywhere but in English

- Government website with security issues
- Lots of bio/science websites

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T1.1	T1.2					T1.3						
Language	R	0	1	2	3	4	0	1	2	3	4	5	6
All	91.73%	3.78%	2.88%	14.91%	34.26%	44.17%	19.37%	2.09%	0.52%	0.52%	1.05%	51.31%	25.13%
Dutch	87.27%	0.30%	0.60%	6.85%	14.88%	77.38%	6.12%	0.00%	0.00%	0.00%	0.00%	61.22%	32.65%
English	93.25%	19.37%	8.26%	2.28%	59.26%	10.83%	5.88%	11.76%	0.00%	2.94%	5.88%	47.06%	26.47%
Japanese	92.73%	0.00%	0.84%	44.82%	11.20%	43.14%	64.29%	0.00%	0.00%	0.00%	0.00%	25.00%	10.71%
Portuguese	91.17%	1.14%	5.70%	17.38%	30.20%	45.58%	11.76%	0.00%	2.94%	0.00%	0.00%	73.53%	11.76%
Spanish	93.25%	0.28%	0.84%	12.26%	36.21%	50.42%	15.38%	0.00%	0.00%	0.00%	0.00%	50.00%	34.62%
Swedish	94.81%	1.64%	1.10%	5.48%	52.60%	39.18%	30.00%	0.00%	0.00%	0.00%	0.00%	35.00%	35.00%

Most barriers were legitimate lack of information

T1.3 Barriers:

0. Page not available
1. Security issues
2. Credentials needed
3. Paywall
4. Domain Knowledge
5. Subject or predicate not mentioned
6. Subject and predicate mentioned, but not object

Aggregated results

	T2.1				T2.2.						Auth.		
Language	I.	O.	C.	N.	A.	C.	G.	Ne.	S.	N.	Yes	No	Ina.
All	0.69%	67.66%	28.87%	2.77%	22.16%	37.75%	6.71%	1.77%	28.79%	2.81%	66.97%	30.22%	2.81%
Dutch	0.26%	94.81%	4.68%	0.26%	8.75%	71.43%	14.81%	0.52%	4.42%	0.26%	94.81%	4.94%	0.26%
English	0.52%	91.43%	5.19%	2.86%	83.12%	5.19%	3.64%	0.26%	4.94%	2.86%	90.91%	6.23%	2.86%
Japanese	0.00%	53.51%	42.08%	4.42%	1.30%	48.57%	2.60%	0.78%	42.34%	4.42%	52.47%	43.12%	4.42%
Portuguese	2.60%	67.53%	29.87%	0.00%	17.92%	31.95%	13.51%	6.75%	29.61%	0.26%	65.71%	34.03%	0.26%
Spanish	0.78%	55.84%	34.29%	9.09%	21.56%	27.79%	5.19%	2.08%	34.29%	9.09%	55.32%	35.58%	9.09%
Swedish	0.00%	42.86%	57.14%	0.00%	0.52%	41.56%	0.52%	0.26%	57.14%	0.00%	57.40%	42.60%	0.00%

T2.1:

- I: Individual
- O: Organisation
- C: Collective
- N: No access

T2.2:

- A: Academic/Scientific
- C: Company/Organisation
- G: Government
- Ne: News
- S: Self-published
- N: No access

Aggregated results

	T2.1				T2.2.						Auth.		
Language	I.	O.	C.	N.	A.	C.	G.	Ne.	S.	N.	Yes	No	Ina.
All	0.69%	67.66%	28.87%	2.77%	22.16%	37.75%	6.71%	1.77%	28.79%	2.81%	66.97%	30.22%	2.81%
Dutch	0.26%	94.81%	4.68%	0.26%	8.75%	71.43%	14.81%	0.52%	4.42%	0.26%	94.81%	4.94%	0.26%
English	0.52%	91.43%	5.19%	2.86%	83.12%	5.19%	3.64%	0.26%	4.94%	2.86%	90.91%	6.23%	2.86%
Japanese	0.00%	53.51%	42.08%	4.42%	1.30%	48.57%	2.60%	0.78%	42.34%	4.42%	52.47%	43.12%	4.42%
Portuguese	2.60%	67.53%	29.87%	0.00%	17.92%	31.95%	13.51%	6.75%	29.61%	0.26%	65.71%	34.03%	0.26%
Spanish	0.78%	55.84%	34.29%	9.09%	21.56%	27.79%	5.19%	2.08%	34.29%	9.09%	55.32%	35.58%	9.09%
Swedish	0.00%	42.86%	57.14%	0.00%	0.52%	41.56%	0.52%	0.26%	57.14%	0.00%	57.40%	42.60%	0.00%

References pointing to Wikipedia

T2.1:

- I: Individual
- O: Organisation
- C: Collective
- N: No access

T2.2:

- A: Academic/Scientific
- C: Company/Organisation
- G: Government
- Ne: News
- S: Self-published
- N: No access

Aggregated results

	T2.1				T2.2.						Auth.		
Language	I.	O.	C.	N.	A.	C.	G.	Ne.	S.	N.	Yes	No	Ina.
All	0.69%	67.66%	28.87%	2.77%	22.16%	37.75%	6.71%	1.77%	28.79%	2.81%	66.97%	30.22%	2.81%
Dutch	0.26%	94.81%	4.68%	0.26%	8.75%	71.43%	14.81%	0.52%	4.42%	0.26%	94.81%	4.94%	0.26%
English	0.52%	91.43%	5.19%	2.86%	83.12%	5.19%	3.64%	0.26%	4.94%	2.86%	90.91%	6.23%	2.86%
Japanese	0.00%	53.51%	42.08%	4.42%	1.30%	48.57%	2.60%	0.78%	42.34%	4.42%	52.47%	43.12%	4.42%
Portuguese	2.60%	67.53%	29.87%	0.00%	17.92%	31.95%	13.51%	6.75%	29.61%	0.26%	65.71%	34.03%	0.26%
Spanish	0.78%	55.84%	34.29%	9.09%	21.56%	27.79%	5.19%	2.08%	34.29%	9.09%	55.32%	35.58%	9.09%
Swedish	0.00%	42.86%	57.14%	0.00%	0.52%	41.56%	0.52%	0.26%	57.14%	0.00%	57.40%	42.60%	0.00%

English is mainly Academic

T2.1:

- I: Individual
- O: Organisation
- C: Collective
- N: No access

T2.2:

- A: Academic/Scientific
- C: Company/Organisation
- G: Government
- Ne: News
- S: Self-published
- N: No access

Takeaways



Takeaways

- References are, in general, moderately to highly **accessible** (levels 3 and 4)
- 2/3 of all references were deemed **authoritative**, varying from 52% to 90% based on the presence of Wikipedia references
- Around 90% of all references were deemed **relevant**
- Inter-worker agreement indicates **external references** were much clearer to workers
- Quality variation between languages was tied to **domain variability**



Discussion points

- Improving provenance representation
 - Wikipedia references are the most easy to access (77% on 4/4) and most relevant (98%)
 - Direct URLs combined with external identifiers are the best option
 - A considerable portion can be automatically verified!
 - Link rot is very present and might also be automatically verified!
 - **Harder cases of relevance need a sophisticated NLP pipelined approach**
- Content across languages
 - Tied to trends in content
 - Communities should both focus on the **overall graph** or ensure **domain-specific quality**
- Limitations
 - Bias on the crowd
 - Volatility of Wikidata

Thank you!
