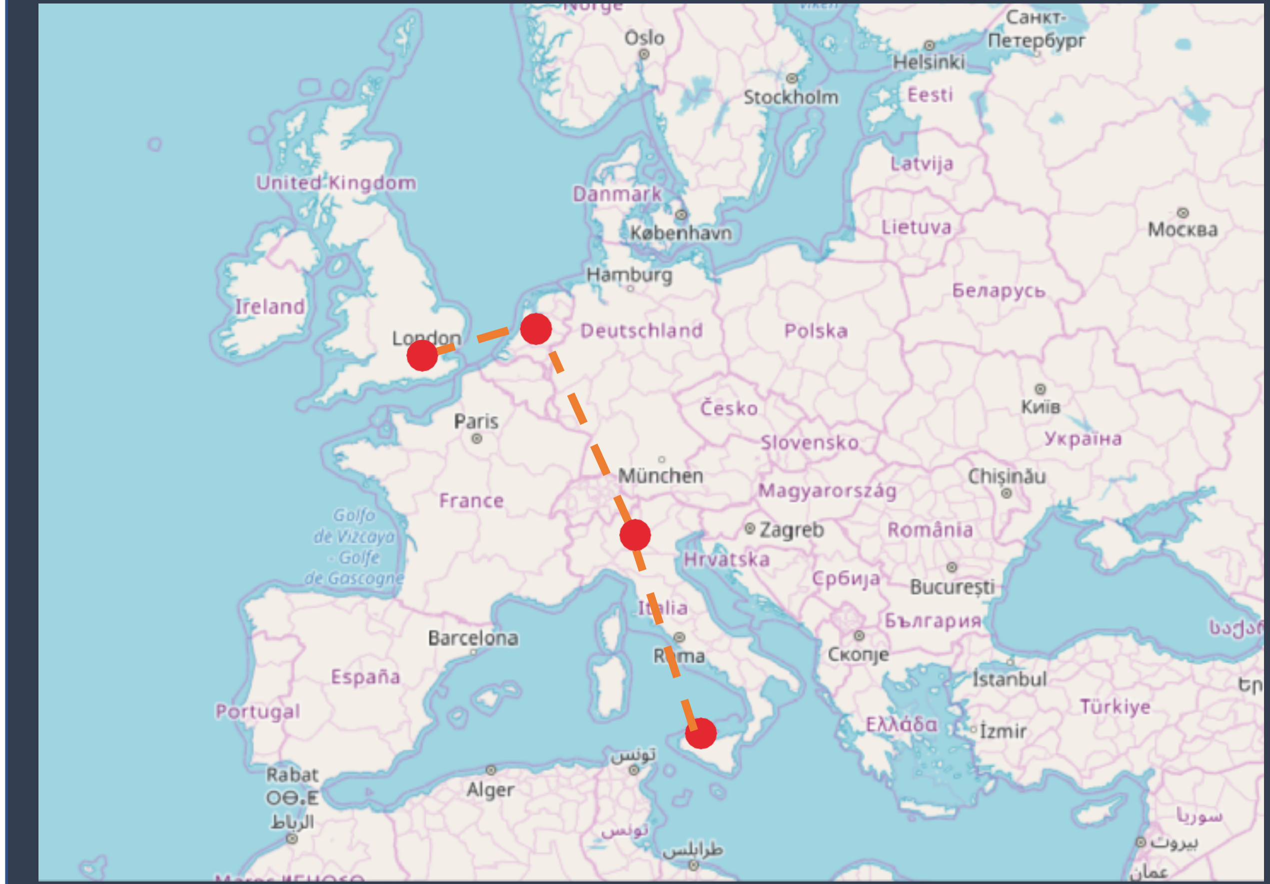
Structuring the world's knowledge: Socio-technical processes and data quality in Wikidata

Alessandro Piscopo

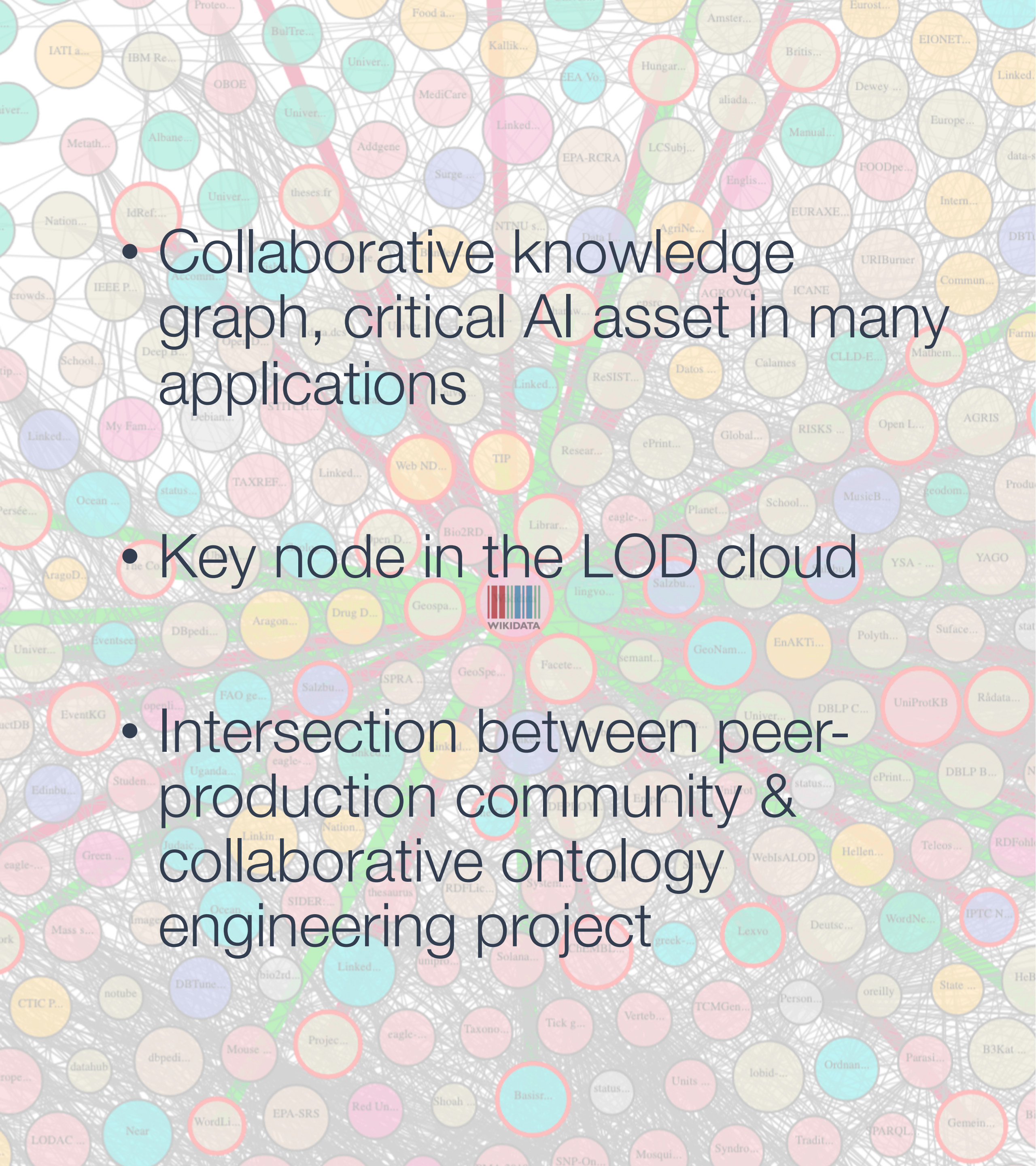
Principal Data Scientist @ Datalab, BBC

About myself

- MA in Classics, 2008
- Editor, Mondadori Education (educational publisher), 2010-2012
- MSc Information Studies – University of Amsterdam, 2013-2014
  - Piscopo, A., Siebes, R., & Hardman, L. (2017). Predicting sense of community and participation by applying machine learning to open government data. *Policy & Internet*, 9(1), 55-75.
- PhD in Computer Science, University of Southampton, 2015-2019
- Project Manager, The Open Data Institute
- Data Scientist, BBC, 2019
  - Data-driven public service recommendations @ Datalab (https://datalab.rocks/)



Wikidata



- Collaborative knowledge graph, critical AI asset in many applications
- Key node in the LOD cloud
- Intersection between peer-production community & collaborative ontology engineering project

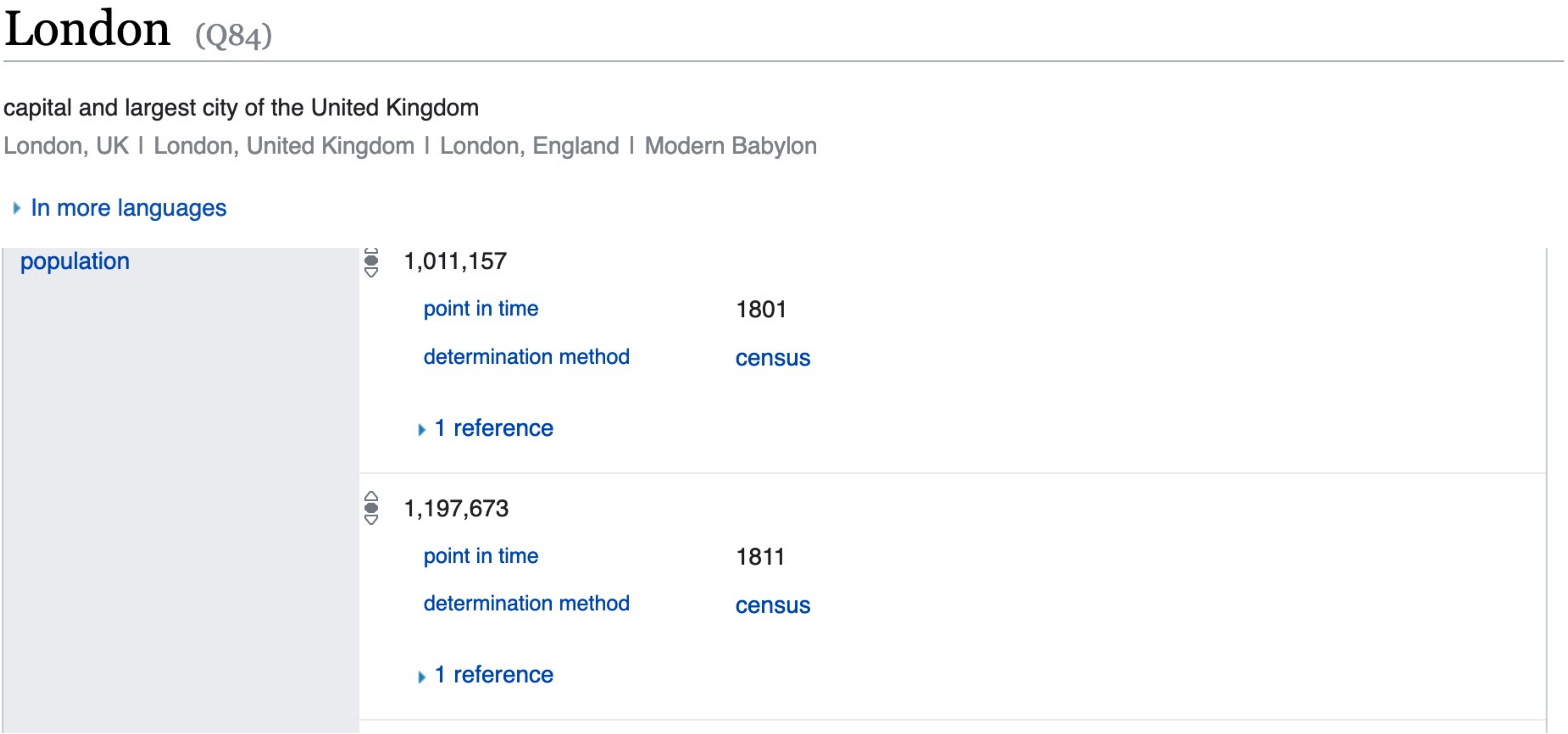
The Knowledge Graph

Q84
London

P17
country

Q145
United Kingdom

- The knowledge graph consists of all items and properties
- Qxx item identifiers, Pxx property identifiers
- Provenance (**references**) and context (**qualifiers**) can be added



London (Q84)

capital and largest city of the United Kingdom

London, UK | London, United Kingdom | London, England | Modern Babylon

▸ In more languages

population

1,011,157

point in time	1801
determination method	census

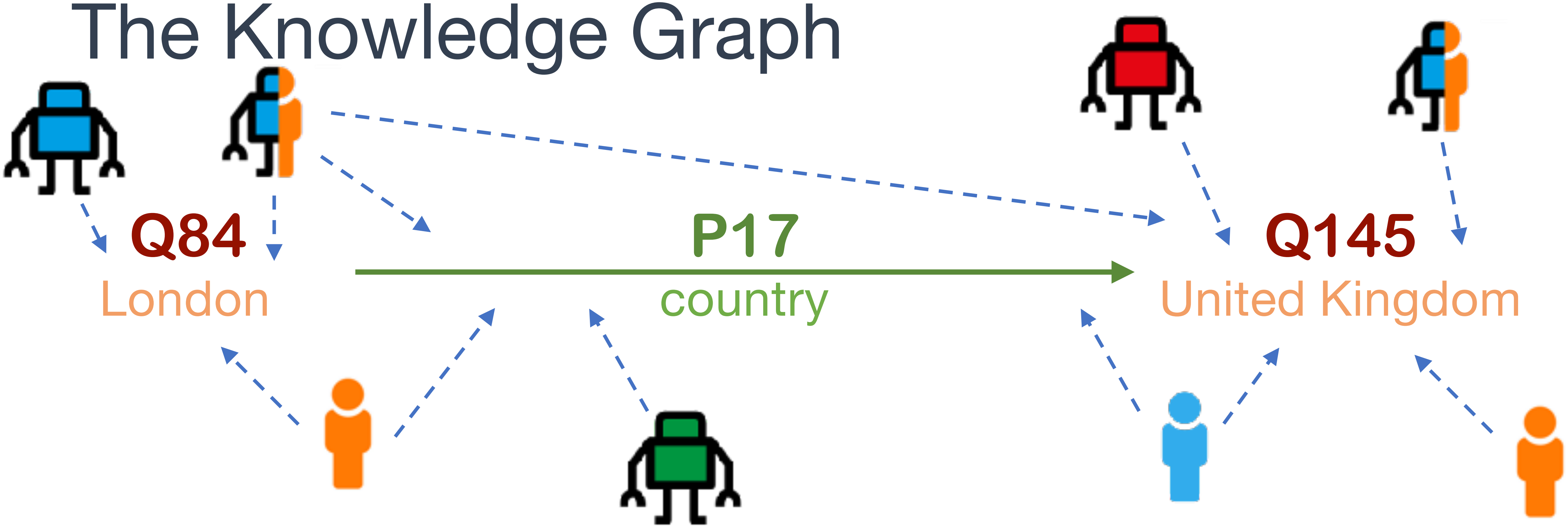
▸ 1 reference

1,197,673

point in time	1811
determination method	census

▸ 1 reference

The Knowledge Graph



- Human editors can register or work anonymously
- Bots created by community for routine task
- Editors can use semi-automated tools to perform revisions in batches

Motivations

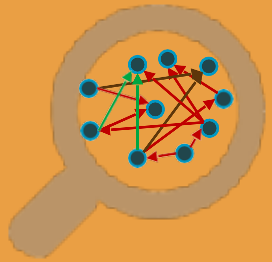
- **Key resource** in Semantic Web
- Several aspects of **quality** still not studied
- **Completely bottom-up effort** to build knowledge resource typically build by teams of trained experts (e.g. ontologies)

Why is this interesting?

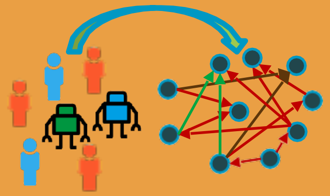
- Research may feed into the work done by practitioners, pointing out issues to be addressed.
- *On the field* practitioner experience is relevant for research.
- Let’s discuss about it!

How does the **socio-technical fabric** of Wikidata influence its **data quality**?

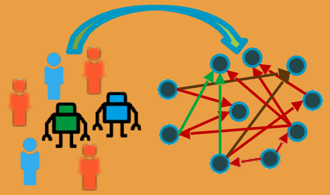
Research questions



Data quality (**provenance**) as function of contributions of different user types.

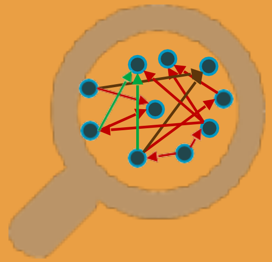


Influence of community make-up on outcomes: **group diversity & Item quality**.

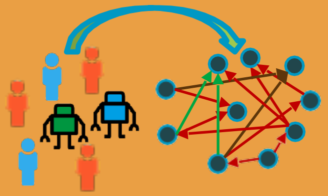


Influence of community make-up on outcomes: **user roles & ontology quality**.

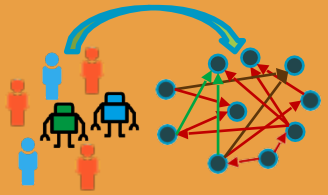
Research questions



How do registered **human and bot users** compare in terms of the quality of their edits?



To what extent does editor **group diversity** affect outcome quality in Wikidata?



What features of **editing roles** affect the quality of the Wikidata **ontology**?

Data

- Wikidata historical dumps – all revisions to every page logged
- October 2016 – October 2017 dumps
- Parsed (from XML+JSON) into PostgreSQL

Looking at the sources: bots and humans edits' quality

Piscopo, A., Kaffee, L. A., Phethean, C., & Simperl, E. (2017, October). Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References. In *International Semantic Web Conference* (pp. 542-558). Springer, Cham.

Background

- Bot– vs. Human-contributed references
- Quality is defined by the Wikidata verifiability policy
  - **Relevant:** support the statement they are attached to
  - **Authoritative:** trustworthy, up-to-date, and free of bias for supporting a particular statement
- Large-scale applicability (the whole of Wikidata)
- External references only (P854)

Our study

A. Are Wikidata external references relevant?

B. Are Wikidata external references authoritative?

- I.e. do they match Wikidata policy author and publisher types?

C. Can we automatically detect non-relevant and non-authoritative references in Wikidata?

Methods

A.

1. Microtask crowdsourcing
   - Evaluate relevance & authoritativeness of a reference sample B.
   - 3 tasks on Crowdflower, 5 workers/task, majority voting
     - T1: Relevance; T2: Author type; T3: Publisher type.
   - Provide training set for Machine Learning model
2. Machine learning C.
   - Large-scale reference quality prediction
   - Features from source URL, item/claim hierarchy, reference creator
   - Naïve Bayes, Random Forest, SVM

Data

- October 2016
- 83,215 English-language references (over 1.6M)
- Sample 2586 (99% conf., 2.5% m. of error)
    - 885 assessed automatically, e.g. links not working or csv files

Results: Crowdsourcing

A.

B.

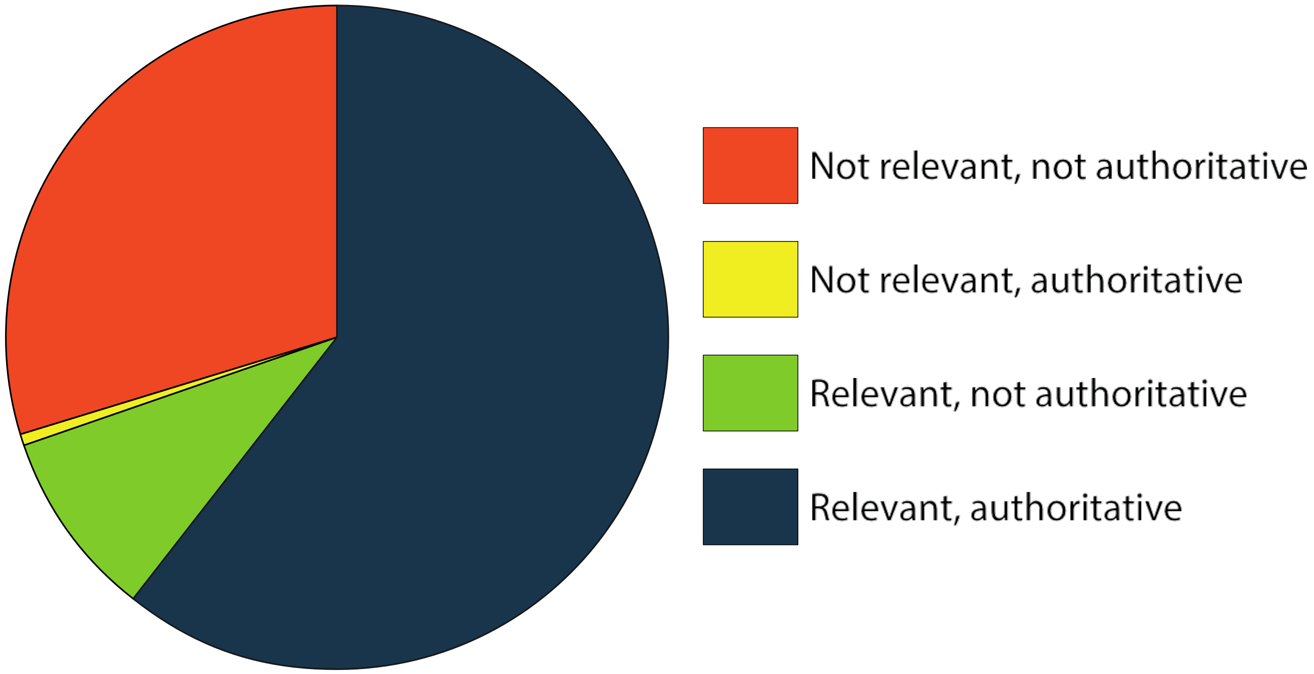
Task	# microtasks	Total workers	Trusted workers	Workers accuracy	Fleiss' k
T1	1701 refs.	457	218	75%	0.335
T2	1178 links	749	322	75%	0.534
T3.A	335 web domains	322	60	66%	0.435
T3.B	335 web domains	239	116	68%	0.391

- Trusted workers: >80% accuracy; moderate to fair agreement
- 95% responses in T3.A confirmed in T3.B

Results: Crowdsourcing

A.

B.



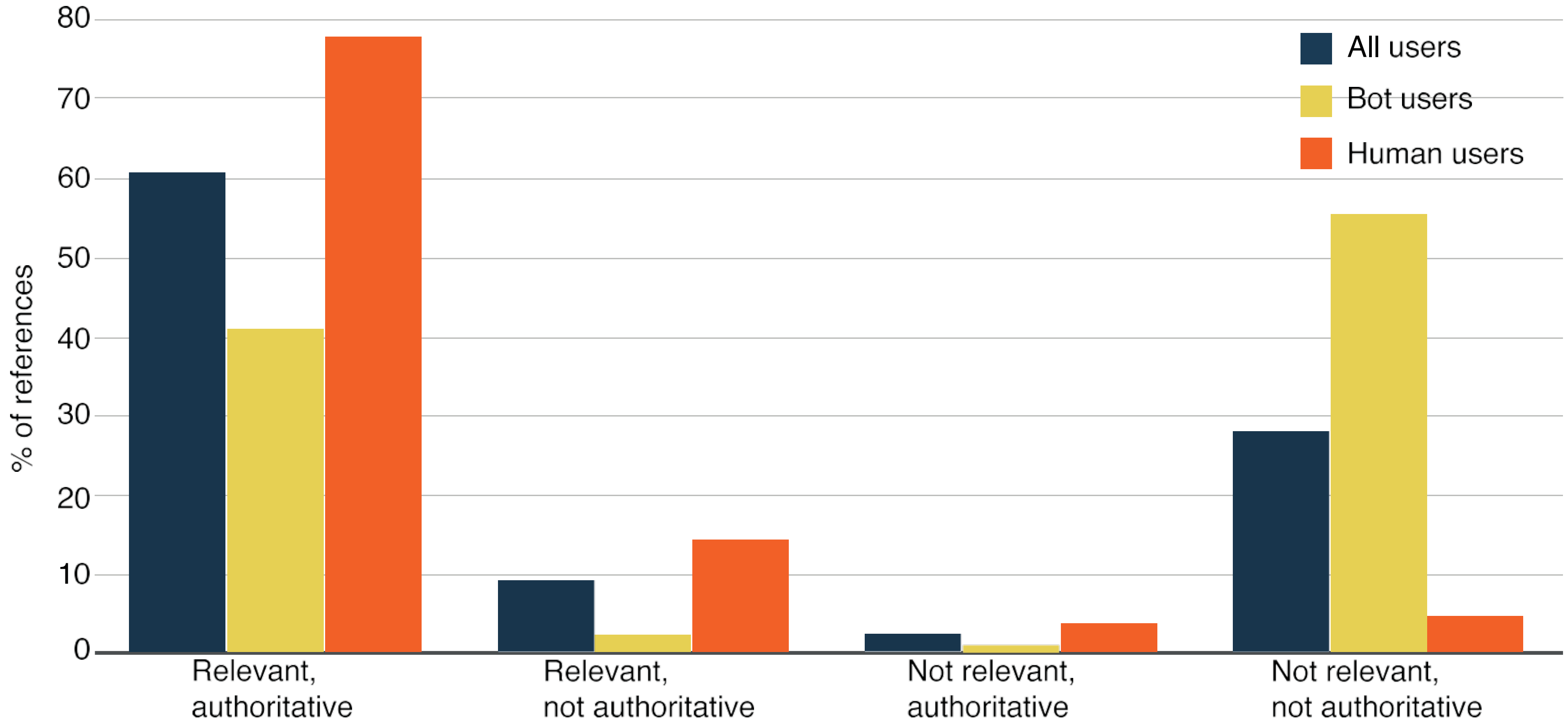
- Trusted workers: >80% accuracy; moderate to fair agreement
- Not-working URLs deemed not-relevant, nor authoritative

Results: Crowdsourcing

- Humans are better at adding references

A.

B.



Results: Machine Learning

C.

- Random Forest is the best performing model

		F_1	MCC
Relevance	Baseline	0.84	0.68
	Naïve Bayes	0.90	0.86
	Random Forest	0.92	0.89
	SVM	0.91	0.87
Authoritativeness	Baseline	0.53	0.16
	Naïve Bayes	0.86	0.78
	Random Forest	0.89	0.83
	SVM	0.89	0.79

Lessons learnt

🤖 & 🧍 are good at different things

A.

Bad references mainly non-working links

Continuous control needed

Many sources are high quality

B.

Lack of diversity in bot-added sources

Crowdsourcing + ML works!

C.

The right mix of users: Group Composition & quality

Piscopo, A., Phethean, C., & Simperl, E. (2017, September). What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *International Conference on Social Informatics* (pp. 305-322). Springer, Cham.

Background

- Wikidata human editors have varied tenure & interests
- Group composition impacts outcomes
    - Diversity has multiple effects
    - Moderate tenure diversity increases outcome quality
    - Interest diversity positively leads to increased group productivity

Our study

- Analysed edit history of Items
    - 5k manually assessed Items corpus (5 levels) (April 2017)
    - Edit history focused on community make-up
- Group is defined as set of editors of an item
- Considered features from group diversity literature and Wikidata-specific aspects

Hypotheses

H1
- ▲ Bot edits
- ▲ Item quality

H2
- ▲ Human-bot interaction
- ▲ Item quality

H3
- ▲ Anonymous edits
- ▼ Item quality

H4
- ▲ Tenure diversity
- ▲ Item quality

H5
- ▲ Interest diversity
- ▲ Item quality

Methods

- Ordinal Logistic Regression analysis, 4 models

Dependent variable Wikidata Items quality labels

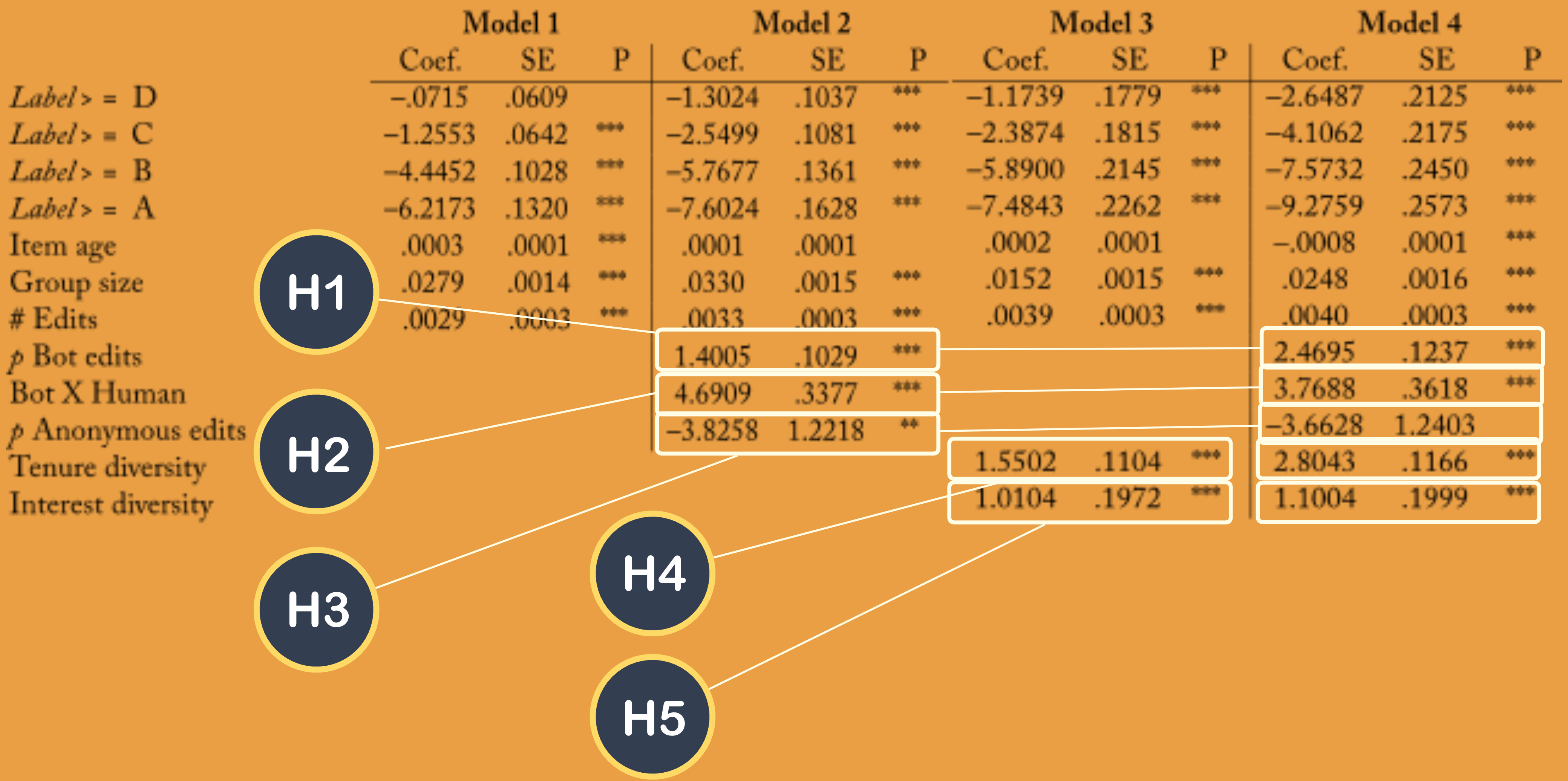
Independent variables

- Proportion of bot edits
- Bot X Human edit proportion
- Proportion of anonymous edits
- Tenure diversity: Coefficient of variation
- Interest diversity: User editing matrix

Control variables

- Group size
- Item age
- No. of edits

Results



	Model 1			Model 2			Model 3			Model 4		
	Coef.	SE	P	Coef.	SE	P	Coef.	SE	P	Coef.	SE	P
Label > = D	–.0715	.0609		–1.3024	.1037	***	–1.1739	.1779	***	–2.6487	.2125	***
Label > = C	–1.2553	.0642	***	–2.5499	.1081	***	–2.3874	.1815	***	–4.1062	.2175	***
Label > = B	–4.4452	.1028	***	–5.7677	.1361	***	–5.8900	.2145	***	–7.5732	.2450	***
Label > = A	–6.2173	.1320	***	–7.6024	.1628	***	–7.4843	.2262	***	–9.2759	.2573	***
Item age	.0003	.0001	***	.0001	.0001		.0002	.0001		–.0008	.0001	***
Group size	.0279	.0014	***	.0330	.0015	***	.0152	.0015	***	.0248	.0016	***
# Edits	.0029	.0003	***	.0033	.0003	***	.0039	.0003	***	.0040	.0003	***
p Bot edits				1.4005	.1029	***				2.4695	.1237	***
Bot X Human				4.6909	.3377	***				3.7688	.3618	***
p Anonymous edits				–3.8258	1.2218	**				–3.6628	1.2403	
Tenure diversity							1.5502	.1104	***	2.8043	.1166	***
Interest diversity							1.0104	.1972	***	1.1004	.1999	***

Lessons learnt

Diversity matters!

🤖 edits are key for quality, but 🤖 + 🧍 are better.

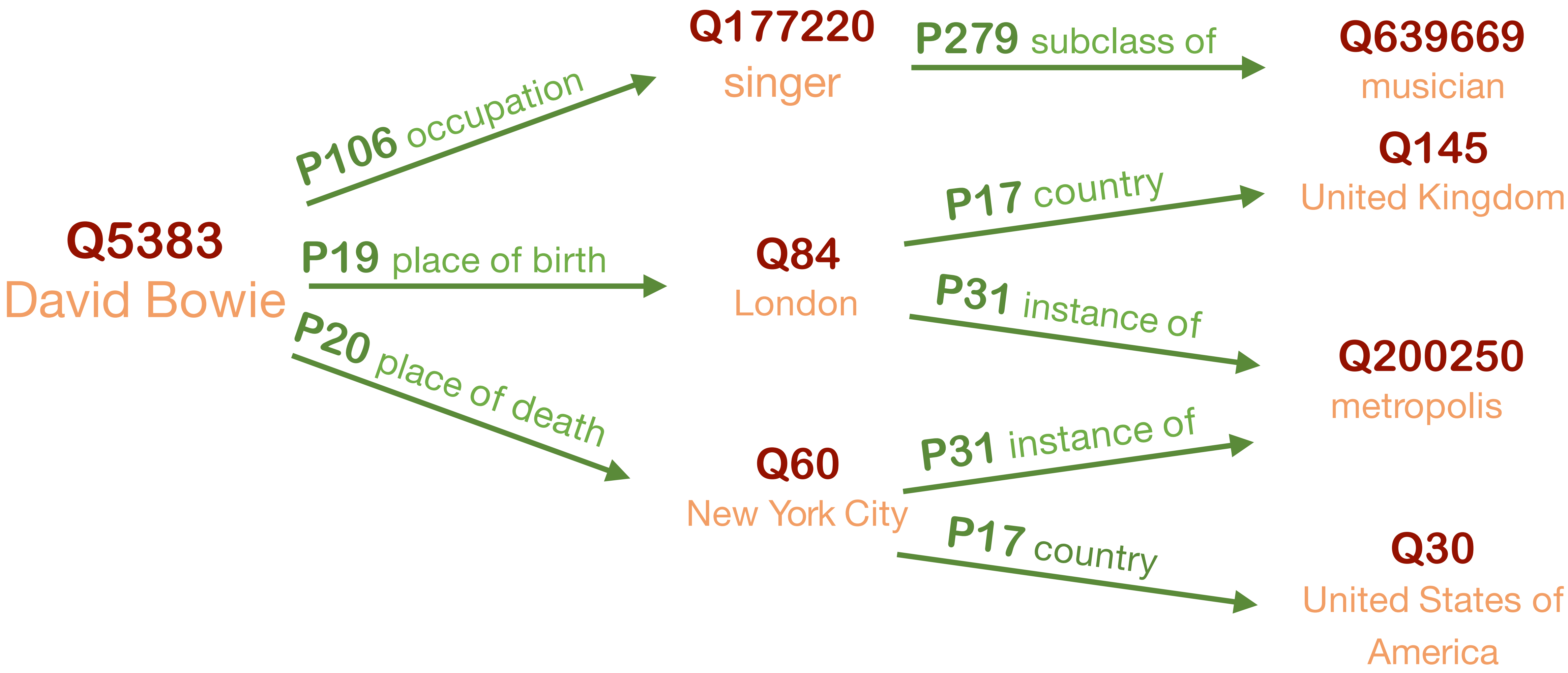
Anonymous users have small influence on quality.

Look at how users focus on different tasks.

Who models the world? User roles and ontology quality

Piscopo, A., & Simperl, E. (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 141.

Give me all musicians who were
born in London and died
in a big US city



Q5383
David Bowie

P106 occupation

P19 place of birth

P20 place of death

Q177220
singer

Q84
London

Q60
New York City

P279 subclass of

P17 country

P31 instance of

P31 instance of

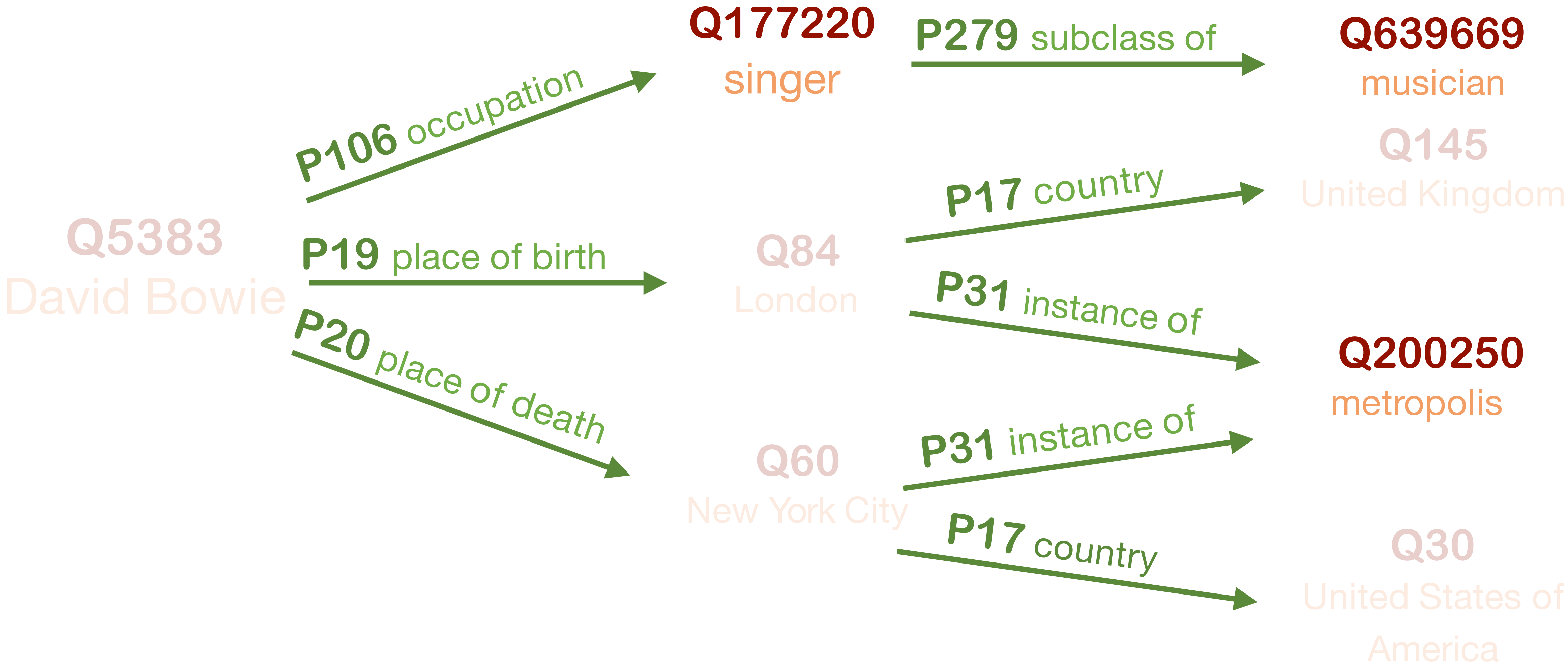
P17 country

Q639669
musician

Q145
United Kingdom

Q200250
metropolis

Q30
United States of America



Q5383
David Bowie

P106 occupation

P19 place of birth

P20 place of death

Q177220
singer

Q84
London

Q60
New York City

P279 subclass of

P17 country

P31 instance of

P31 instance of

P17 country

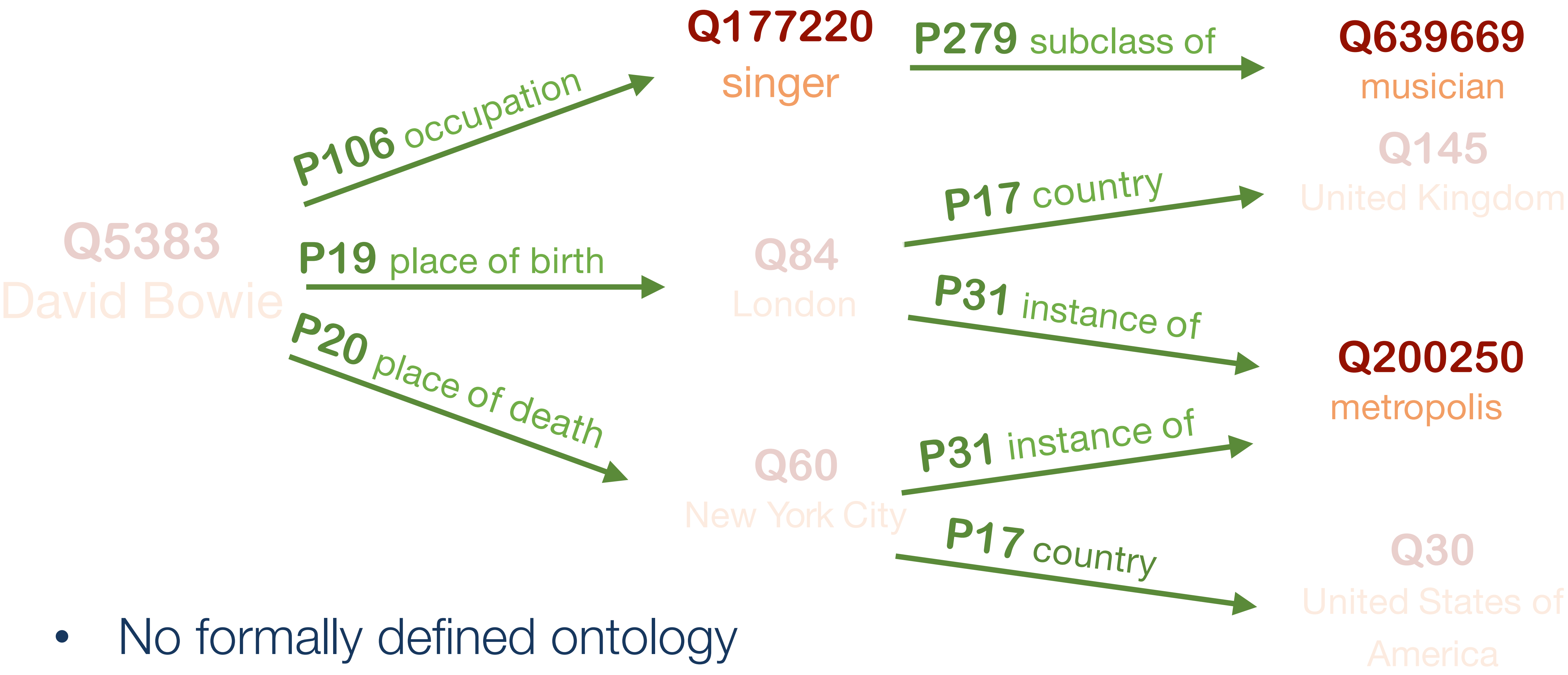
Q639669
musician

Q145
United Kingdom

Q200250
metropolis

Q30
United States of America

- No formally defined ontology



Q5383
David Bowie

P106 occupation

P19 place of birth

P20 place of death

Q177220
singer

Q84
London

Q60
New York City

P279 subclass of

P17 country

P31 instance of

P31 instance of

P17 country

Q639669
musician

Q145
United Kingdom

Q200250
metropolis

Q30
United States of America

- No formally defined ontology
- Items can be entities or classes
- Classes are subject/object of P279 (subclass of) or object of P31 (instance of)

Background

Reports about contributors failing to use correctly the taxonomic relations P31 and P279

Wikidata editors present various behaviours

- Novices perform simpler tasks
- Established users focus on quality control, are more active within the community, and maintain the ontology
- Property editors, property engineers

Our work

S1. Set of ontology quality metrics for Wikidata

- Evaluated the quality of the ontology over time

S2. User roles based on emerging activity patterns

S3. Influence of user roles on ontology quality

Ontology quality: Requirements

Review of ontology metrics research

R1 Consider factors influenceable by users

R2 Able to assess ontology over time

R3 Use only the ontology for the assessment

R4 Indicators that could be implemented unambiguously and be computed automatically

Ontology quality: Metrics

S1.

- 7 ontology frameworks analysed
- Structural metrics, able to show changes

Indicator	Description	Indicator	Description
noi	Number of instances	*ap; mp*	Average and median population
noc	Number of classes	*rr*	Relationship richness
norc	Number of root classes	*ir, mr*	Inheritance and median richness
nolc	Number of leaf classes	*cr*	Class richness
nop	Number of properties	*ad, md, maxd*	Average, median, and max explicit depth

Ontology quality: Results

S1.

Large ontology (>1.5M classes, ~4000 properties) Uneven quality.

- Number of classes increases at same rate than total items

Ontology quality: Results

S1.

Large ontology (>1.5M classes, ~4000 properties). Uneven quality.

- Number of classes increases at same rate than total items
- *ap* and *cr* decrease over time

Ontology quality: Results

S1.

Large ontology (>1.5M classes, ~4000 properties). Uneven quality.

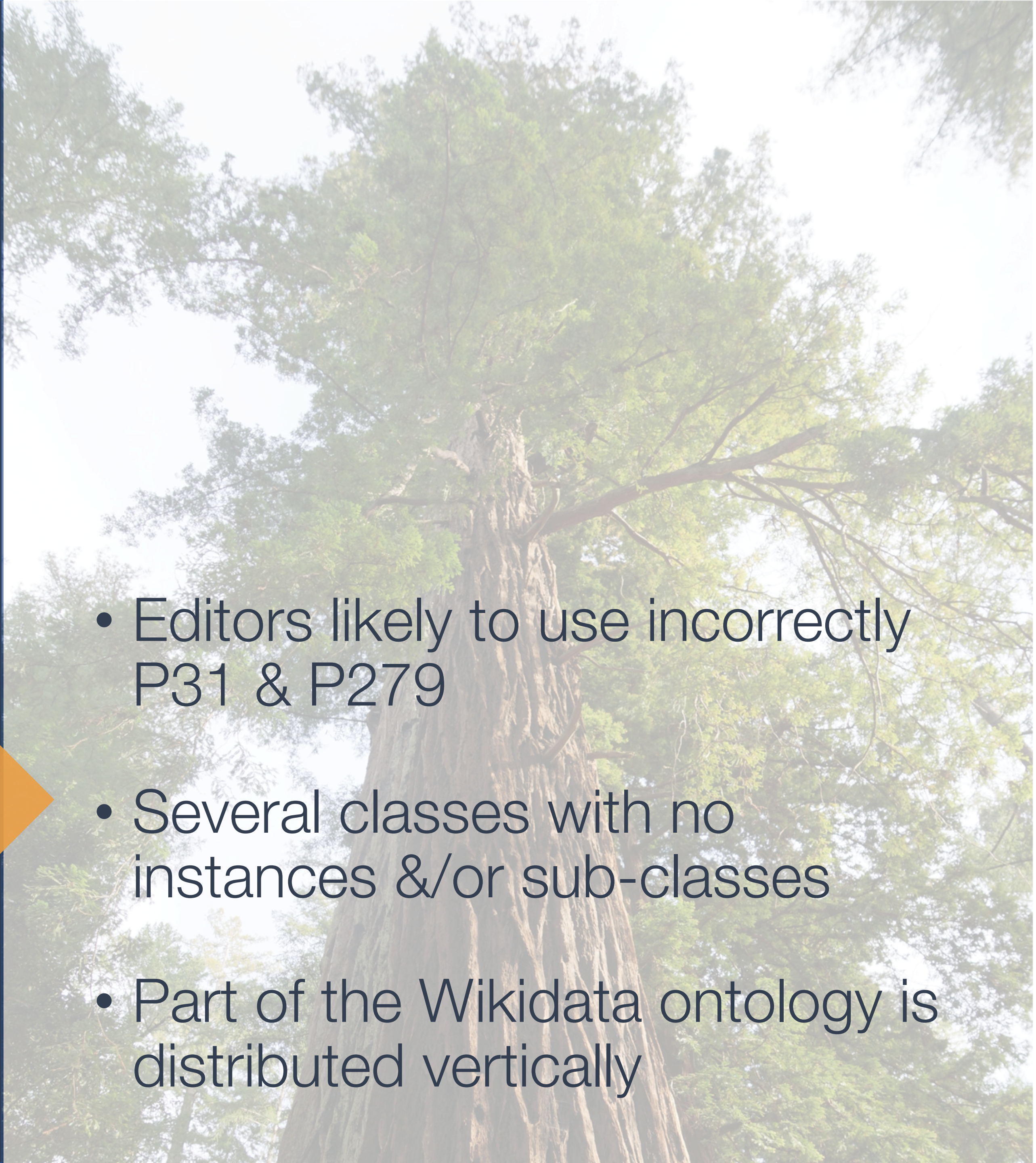
- Number of classes increases at same rate than total items
- *ap* and *cr* decrease over time
- *ir* & *maxd* increase

Ontology quality: Results

S1.

Large ontology (>1.5M classes, ~4000 properties). Uneven quality.

- Number of classes increases at same rate than total items
- *ap* and *cr* decrease over time
- *ir* & *maxd* increase



- Editors likely to use incorrectly P31 & P279
- Several classes with no instances &/or sub-classes
- Part of the Wikidata ontology is distributed vertically

User roles: Methods

S2.

- **K-means**, features based on previous studies

Feature	Description	Feature	Description
# edits	Total number of edits in a month.	*# property edits*	Total number of edits on Properties in a month.
# ontology edits	Number of edits on classes.	*# taxonomy edits*	Number of edits on P31 and P279 statements.
# discussion edits	Number of edits on talk pages.	*p batch edits*	Number of edits done through automated tools.
# modifying edits	Number of revisions on previously existing statements.	*item diversity*	Proportion between number of edits and number of items edited.
admin	True if user in an admin user group, false otherwise.	*lower admin*	True if user in a user group with enhanced user rights, false otherwise.

User roles: Results

S2.

- 190,765 unique users working over 55 months (783k total)
- 18k users active across >10 months
- 2 clusters, obtained using gap statistic (tested 2≥k≥8)

Leaders: more active minority (~1%), higher amount of contributions to ontology, engaged within the community

Contributors: less active, lower amount of contributions to ontology and lower proportion of batch edits

User roles & Ontology: Methods

Lagged multiple regression

Dependent variables

Ontology metrics’ change across timeframes

Independent variables

Amount of contributions of leaders and contributors

Control variables

Bot and anonymous contributions

S3.

User roles & Ontology: Hypotheses

H1

▲ Leader activity

▼ # classes (*noc*)
root classes (*norc*)
leaf classes (*nolc*)

H2

▲ Leader activity

▲ Inheritance richness (*ir*)
Average population (*ap*)
Average depth (*ad*)

User roles & Ontology: Results

S3.

H1 ✗ H2 ✓

- Only inheritance richness (*ir*) and average depth (*ad*) interact significantly with leader edits ($p<0.01$)
- Bot edits significantly and positively affect the number of subclasses and instances per class (*ir* & *ap*) ($p<0.05$)

Lessons learnt

Curated core next to large number of empty classes.

Suitable framework to monitor changes to ontology quality.

Formal roles do not correspond to activity patterns.

S1.

S2.

Creation of conceptual knowledge still a hard task.

User roles less articulated than in other platforms.

Possible decline of motivation after first months.

S3.

Lessons learnt

Curated core next to large number of empty classes.	Suitable framework to monitor changes to ontology quality.	Formal roles do not correspond to activity patterns. **S2.**
S1. Creation of conceptual knowledge still a hard task.	User roles less articulated than in other platforms.	**S3.** Possible decline of motivation after first months.

Bonus content

The cost of freedom

- Property constraints are not enforced.
  - Contrasting points of view can be expressed.
  - No restrictions are imposed on content, except for data type
- Analysed property constraints in relation to outcome quality, knowledge diversity and edit wars.

H1 – Property constraints increase the perspicuity of properties.

H2 – The introduction of constraints affects negatively knowledge diversity.

H3 – The introduction of constraints raises the chances of edit wars.

The cost of freedom

H1 – Supported, but limited to some constraints.

H2 – Not supported, mixed results.

H3 – Not supported, edit wars decrease as constraints are introduced.

- Constraints extracted from Talk pages
- Metrics not really clear
  - Hard to measure diversity
  - Hardly any edit war in structured part of data

DOLCE+Wikidata

- Foundational ontologies are “axiomatic accounts of high-level domain-independent categories about the real world”.
- Paulheim & Gangemi (2015) use DOLCE to identify inconsistencies in DBPedia.

Extract Wikidata ontology

Align Wikidata ontology with DOLCE

Run reasoner on whole of Wikidata

DOLCE+Wikidata

- Wikidata ontology is huge & very messy!
- Aligned DOLCE with Wikidata, using DBPedia and YAGO as intermediate KBs
- Never managed to make the ontology consistent
  - Pruned, removed instanceless classes
  - Cleaned, manually removing inconsistencies
- Alignment dataset is still available!

Conclusions
& lessons learnt

Conclusions & lessons learnt

- Collaboration between 🤖 & 👤 is important, ensuring large scale growth and quality.
- Community still young, important to analyse it in a few years to understand how it articulates itself.
- 🤖 add large amount of data, i.e. damage they can cause is larger than human editors.
- Anonymous edits seem detrimental but influence is small.
- Diversity is beneficial - the community needs contributions from both newcomers and seasoned users

Conclusions & lessons learnt

- Roles may not be fully developed yet/approach used is not sensitive enough to differences in users activity patterns.
- Roles reflect to some extent what seen in previous peer production project, but not fully. The combination of auto/semi-auto/manual edits + scope of the project are new.
- Ontology has uneven quality, external references are generally good.
- Some automated checks are easy to implement and would bring immediate gain to the Wikidata community.

Future work

- Wikidata still in early development stage at time of study
- Hard to say something definitive: new approaches may be required for Wikidata
- Semi-automated edits still not studied
- Talk pages: how are they used?
    - How do they support collaborative knowledge engineering processes?
- Semantic evaluation of ontology, i.e. modelling primitives
- User activity across different semantic areas, e.g. do user edit along branches of trees jumping from a branch to another? Do they go wide or deep?

- How accurate, comprehensive, etc. is Wikidata?

- SPARQL query logs

References

1. Piscopo, A., Kaffee, L. A., Phethean, C., & Simperl, E. (2017, October). Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References. In *International Semantic Web Conference* (pp. 542-558). Springer, Cham.
2. Piscopo, A., Phethean, C., & Simperl, E. (2017, September). What Makes a Good Collaborative Knowledge Graph: Group Composition and Quality in Wikidata. In *International Conference on Social Informatics* (pp. 305-322). Springer, Cham.
3. Piscopo, A., & Simperl, E. (2018). Who Models the World?: Collaborative Ontology Creation and User Roles in Wikidata. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW), 141.
4. Piscopo, A., Phethean, C., & Simperl, E. (2017, January). Wikidatians are born: paths to full participation in a collaborative structured knowledge base. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.
5. Piscopo, A., & Simperl, E. (2019, August). What we talk about when we talk about Wikidata quality: a literature survey. In *Proceedings of the 15th International Symposium on Open Collaboration* (pp. 1-11).
6. Piscopo, A. (2019). *Structuring the world's knowledge: socio-technical processes and data quality in Wikidata* (Doctoral dissertation, University of Southampton).
7. Müller-Birn, C., Karran, B., Lehmann, J., & Luczak-Rösch, M. (2015, August). Peer-production system or collaborative ontology engineering effort: What is Wikidata?. In *Proceedings of the 11th International Symposium on Open Collaboration* (p. 20). ACM.
8. Paulheim, H., & Gangemi, A. (2015, October). Serving DBpedia with DOLCE–more than just adding a cherry on top. In *International Semantic Web Conference* (pp. 180-196). Springer, Cham.
9. Schneider L. (2003) How to Build a Foundational Ontology. In: Günter A., Kruse R., Neumann B. (eds) KI 2003: Advances in Artificial Intelligence. KI 2003. Lecture Notes in Computer Science, vol 2821. Springer, Berlin, Heidelberg.
10. Sicilia, M. A., Rodríguez, D., García-Barriocanal, E., & Sánchez-Alonso, S. (2012). Empirical findings on ontology metrics. Expert Systems with Applications, 39(8), 6706-6711.