Provenance Information in a Collaborative Knowledge Graph: an Evaluation of Wikidata External References

Alessandro Piscopo, Lucie-Aimée Kaffee,

Christopher Phethean, Elena Simperl

Web and Internet Science group
University of Southampton

A.Piscopo@soton.ac.uk

Why Wikidata?

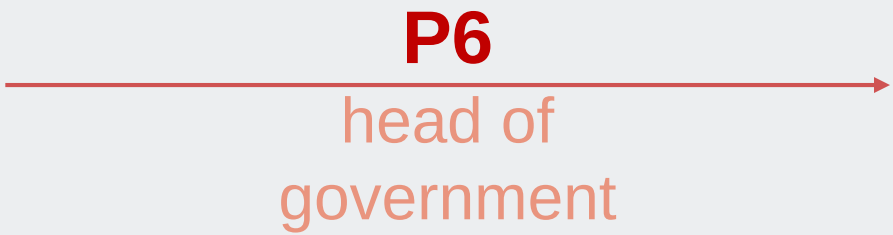
Collaborative knowledge graph

>100k registered users, >**30M** items

Open licence

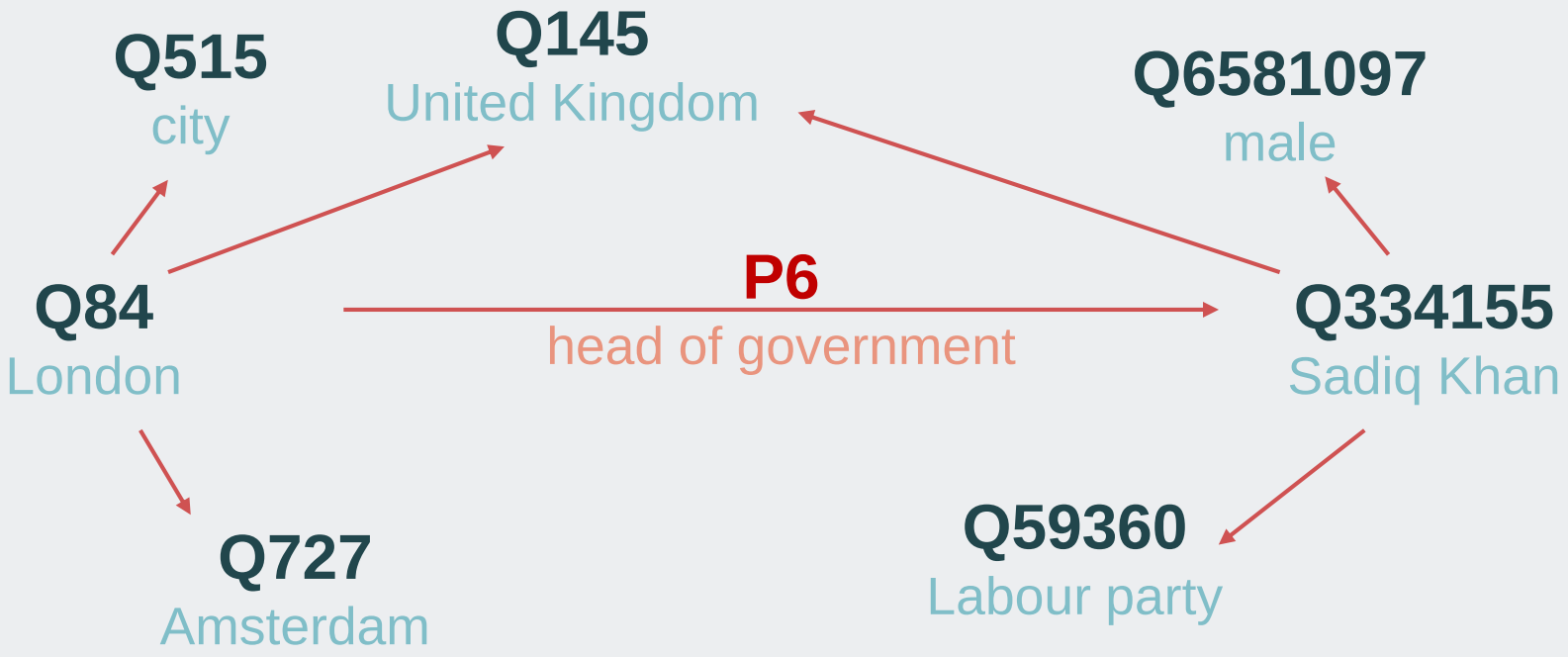
RDF exports, connected to LOD cloud

About Wikidata



- **Qxx** item identifiers
- **Pxx** property identifiers

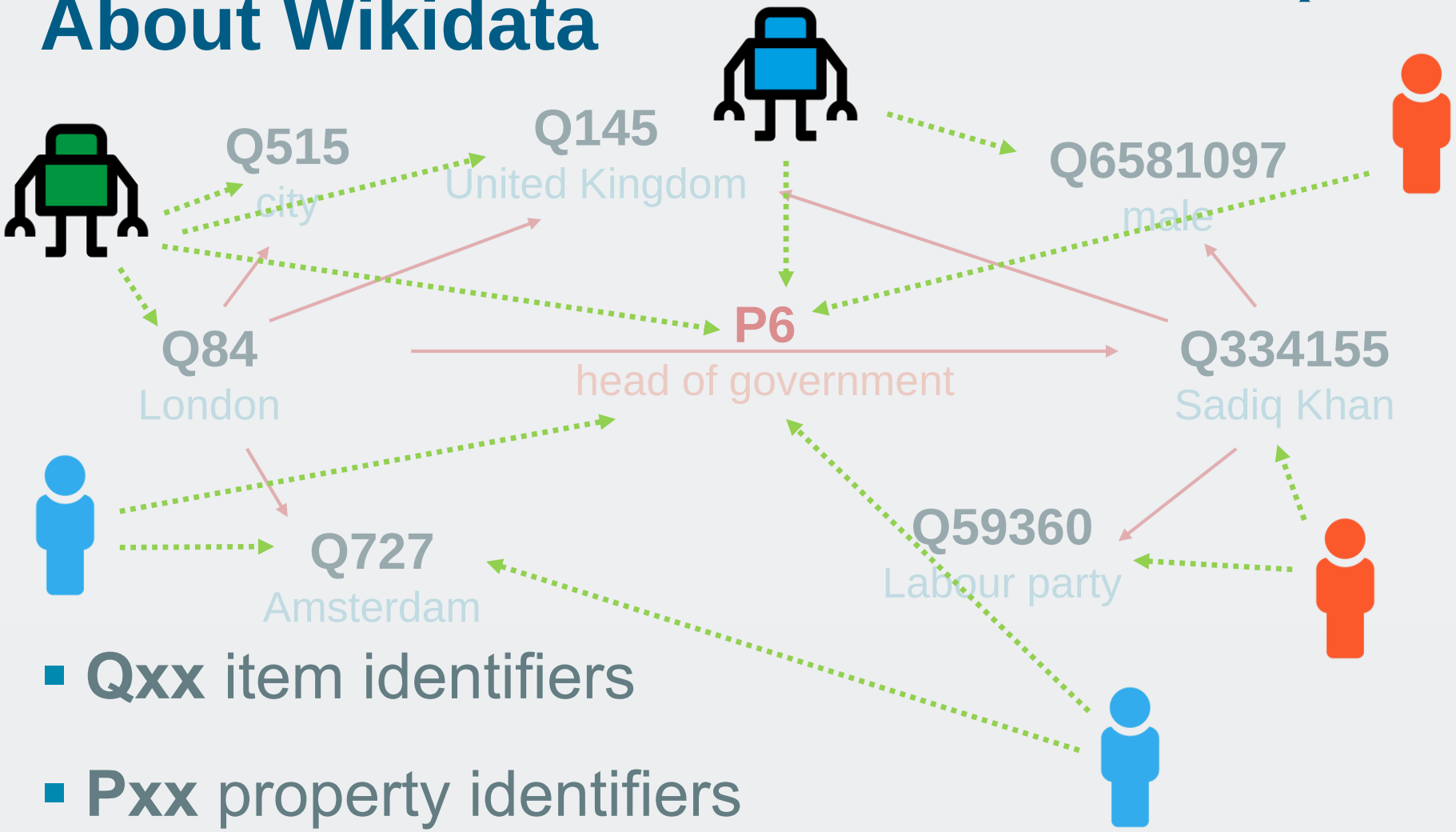
About Wikidata



- **Qxx** item identifiers
- **Pxx** property identifiers

UNIVERSITY OF Southampton

About Wikidata



- **Qxx** item identifiers
- **Pxx** property identifiers
- & have different editing patterns

About Wikidata

- Statements can be enriched by qualifiers and references

UNIVERSITY OF Southampton

Provenance in Wikidata

- All statements require **provenance**, i.e. a reference
- **Internal** references, linking to other item
- **External** references, linking to webpage

UNIVERSITY OF Southampton

Motivations

- Hub of citations and references
- Provenance may increase trust in Wikidata
- Lack of provenance may hinder data reuse [Hartig et al., 2009]
- But quality of Wikidata references has not been studied yet!

Aim of the study

- Approach to **evaluate quality of provenance** in Wikidata
  - External references
  - Large-scale (the whole of Wikidata)
  - Bot– vs. Human-contributed references

Aim of the study

- Good references are[1]
  - **Relevant**: support the statement they are attached to
  - **Authoritative**: trustworthy, up-to-date, and free of bias for supporting a particular statement

1. According to Wikidata verifiability policy, https://www.wikidata.org/wiki/Wikidata:Verifiability

UNIVERSITY OF Southampton

Research questions

RQ1 To what extent are Wikidata external references relevant?

RQ2 To what extent are Wikidata external references authoritative?

- i.e. match author and publisher types from Wikidata policy

RQ3 To what extent can non-relevant and non-authoritative references be predicted in Wikidata?

UNIVERSITY OF Southampton

Methods

- Two-stage mixed approach
- **Microtask crowdsourcing**
  - Evaluate relevance & authoritativeness of a reference sample
  - Provide training set for Machine Learning model
- **Machine learning**
  - Large-scale reference quality prediction

RQ1

RQ2

RQ3

Stage1
Microtask crowdsourcing

UNIVERSITY OF Southampton

- 3 tasks on Crowdflower
- 5 workers/task, majority voting
- Test questions to select workers

RQ1

RQ2

Feature	Microtask	Description
Relevance	**T1**	Does the reference support the statement?
Authoritativeness	**T2**	Choose author type from list
	T3.A	Choose publisher type from list
	T3.B	Verify publisher type, then choose sub-type from list

Stage 2
Machine Learning

UNIVERSITY OF Southampton

RQ3

- Compared three algorithms
  - Naïve Bayes, Random Forest, SVM
  - Features based on [Lehmann et al., 2012 & Potthast et al. 2008]
- Baseline: item labels matching (relevance); deprecated domains list (authoritativeness)

Features	
URL reference uses	Subject parent class
Source HTTP code	Property parent class
Statement item vector	Object parent class
Statement object vector	Author type
Author activity	Author activity on refs.

UNIVERSITY OF Southampton

Data

- Wikidata dumps to 1st October 2016
- 1.6M external references (6% of total)
  - 1.4M from two sources (protein KBs)
- 83,215 English-language references
  - Sample 2586 (99% conf., 2.5% m. of error)
  - 885 assessed automatically, e.g. links not working or csv files

Results Crowdsourcing

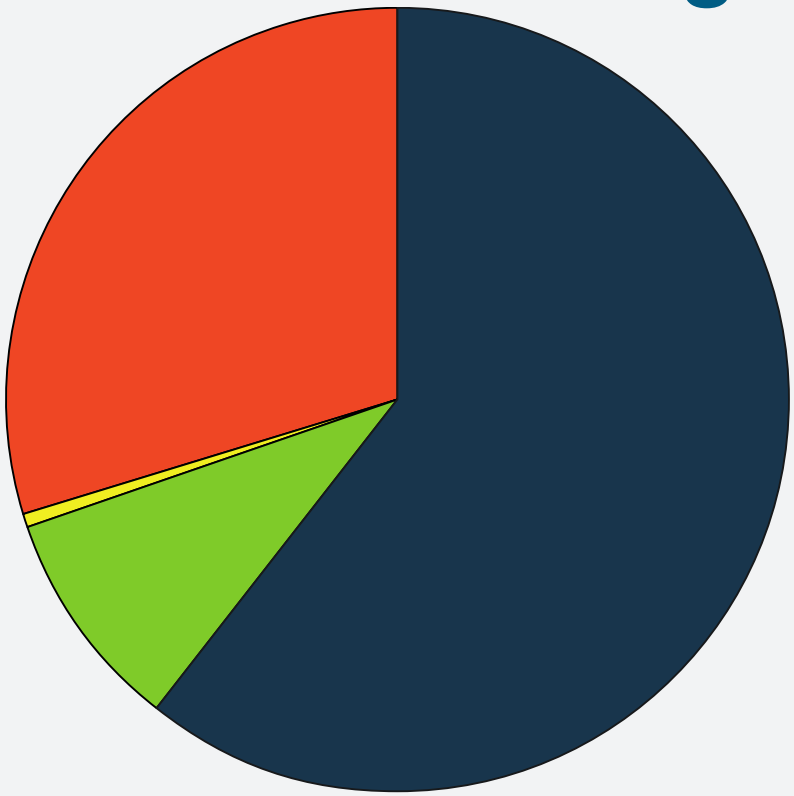
UNIVERSITY OF Southampton

Task	No. of microtasks	Total workers	Trusted workers	Workers' accuracy	Fleiss' *k*
T1	1701 references	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
- 95% of responses from T3.A confirmed in T3.B

Results Crowdsourcing

UNIVERSITY OF Southampton



Not relevant, not authoritative

Not relevant, authoritative

Relevant, not authoritative

Relevant, authoritative

RQ1

RQ2

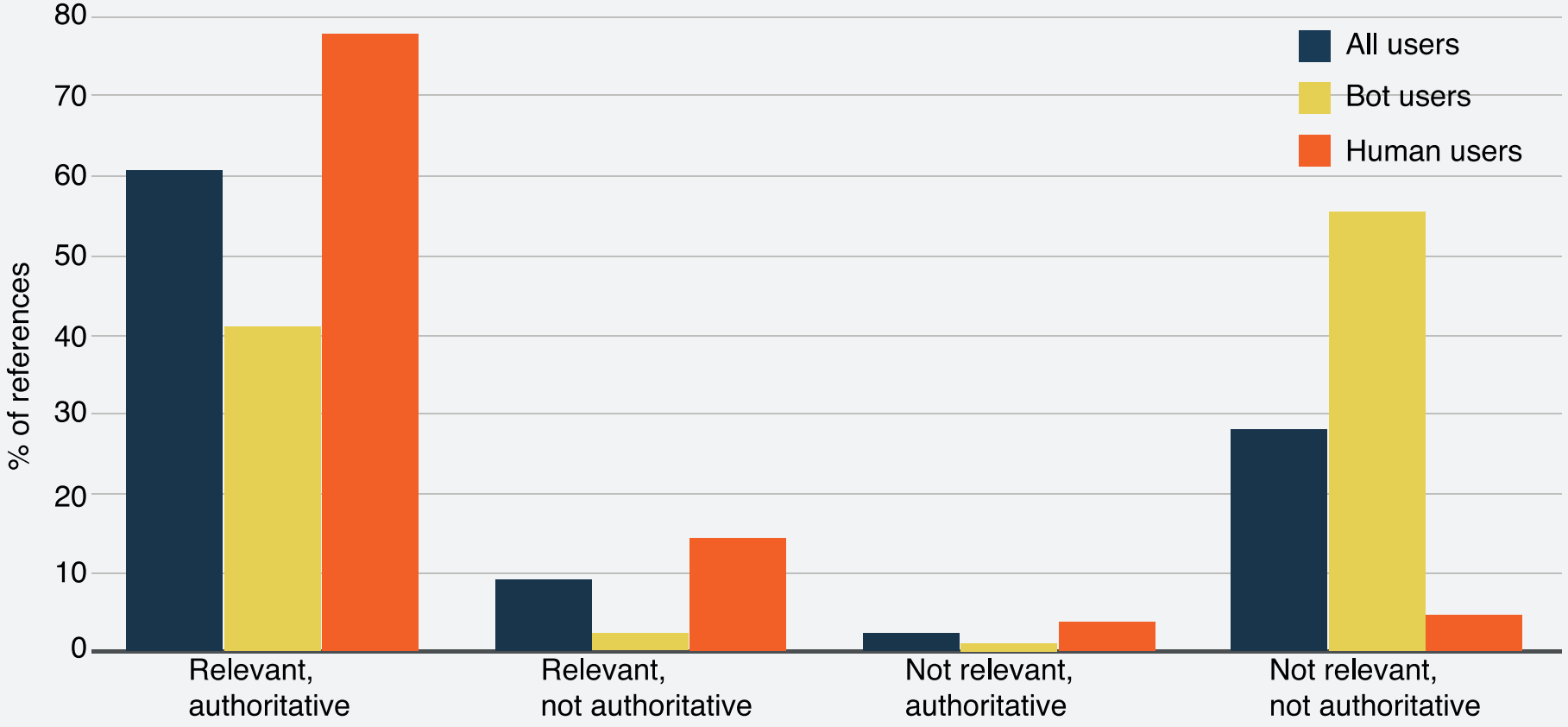
- 2586 references evaluated
  - Not-working URLs deemed not-relevant, nor authoritative

1674 URLs

345 domains

Results Crowdsourcing

UNIVERSITY OF Southampton



- Human-added references have higher quality

Results
Crowdsourcing

UNIVERSITY OF Southampton

RQ2

- Author type (T2)
  - 78% Organisation
  - 8% Individual
  - 3% Collective (…)
- Publisher type (T3)
  - 37% Governmental agencies
  - 14% Other companies & organisations
  - 12% Academic & research organisations
  - 11% Other academic organisations (…)

Results
Machine Learning

UNIVERSITY OF Southampton

		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
RQ3			
Authoritativeness	Baseline	0.53	0.16
	Naïve Bayes	0.86	0.78
	Random Forest	**0.89**	**0.83**
	SVM	0.89	0.79

- All our models outperformed the baseline
- Random Forest was the best performing

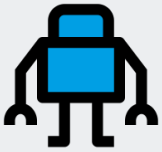
UNIVERSITY OF Southampton

Limitations

- Low number of web domains per property may be behind performance of ML models
- New studies to understand how subjective the task is
- Test of statistical significance of differences in workers' answers

UNIVERSITY OF Southampton

Conclusions

- Crowdsourcing+ML works!
- Most sources are high quality
- Bad references mainly non-working links, continuous control required
- Lack of diversity in bot-added sources
-  &  are good at different things

UNIVERSITY OF Southampton

Future work

- Non-English sources should be studied
- New approach needed for internal references
- Future implementation in Wikidata, in order to study changes in editors behaviour

UNIVERSITY OF Southampton

Future work

- Non-English sources should be studied
- New approach needed for internal references
- Future implementation in Wikidata, in order to study changes in editors behaviour

A.Piscopo@soton.ac.uk

WDAqua