Getting the Most Out of Wikidata

Semantic Technology Usage in Wikipedia's Knowledge Graph

Stas Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, Adrian Bielefeldt

Wikimedia Foundation

TU Dresden

Presentation of paper published at the
International Semantic Web Conference 2018
Download:
https://iccl.inf.tu-dresden.de/web/Inproceedings3044/en

WIKIMEDIA FOUNDATION

TECHNISCHE UNIVERSITÄT DRESDEN

Timeline of tropical cyclones

1600 1650 1700 1750 1800 1850 1900 1950 2000

1 August 1635
Great Colonial Hurricane of 1635

9 October 1780
Great Hurricane of 1780

1 January 1887
1887 Halloween tropical storm

5 October 1954
Hurricane Hazel

7 Octobe
Hurricane M

23 September 1848
Great Gale of 1848

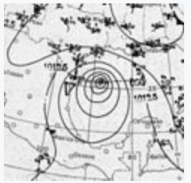
28 September 1837
1837 Racer's Storm

1 January 1831
1831 Barbados hurricane

1 September 1821
1821 Norfolk and Long Island hurricane

22 September 1815
1815 New England hurricane

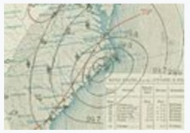
15 August 1812
1812 Louisiana hurricane



12 August 1916
1916 Texas hurricane

28 June 1916
1916 Gulf Coast hurricane

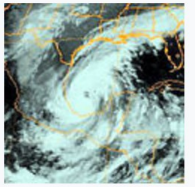
5 August 1915
1915 Galveston hurricane



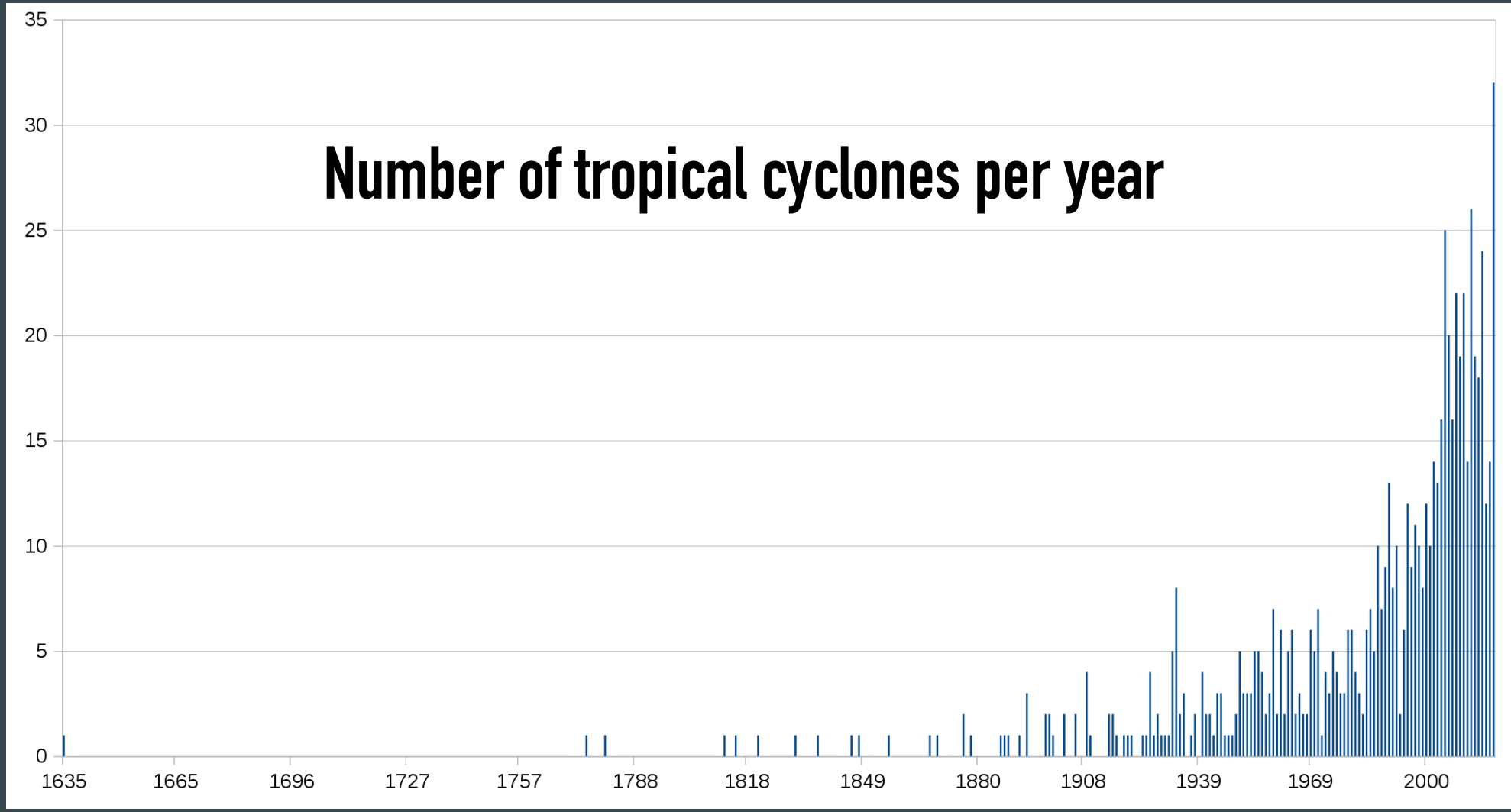
9 October 1910
1910 Cuba hurricane

7 November 1980
Hurricane Jeanne

31 July 1980
Hurricane Allen



14 September 1979


Number of tropical cyclones per year
35
30
25
20
15
10
5
0
1635
1665
1696
1727
1757
1788
1818
1849
1880
1908
1939
1969
2000

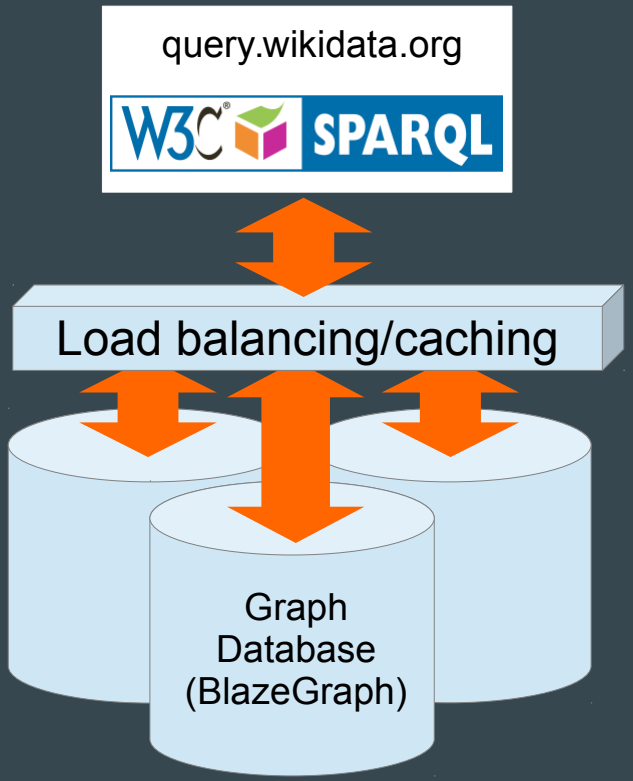

EXCLUSIVE: WARP DRIVE UNDERWATER • ARCTIC OIL VS. WILDLIFE
SCIENTIFIC AMERICAN
MAY 2001
WWW.SCIAM.COM
$4.95
Get the Idea?
(TOMORROW'S WEB WILL)
i know what you mean...
PLUS:
Antibiotics' Dim Future
Rorschach: A Waste of Ink
The Oldest Stars
Copyright 2001 Scientific American, Inc.

The Wikidata Query Service

Wiki Website

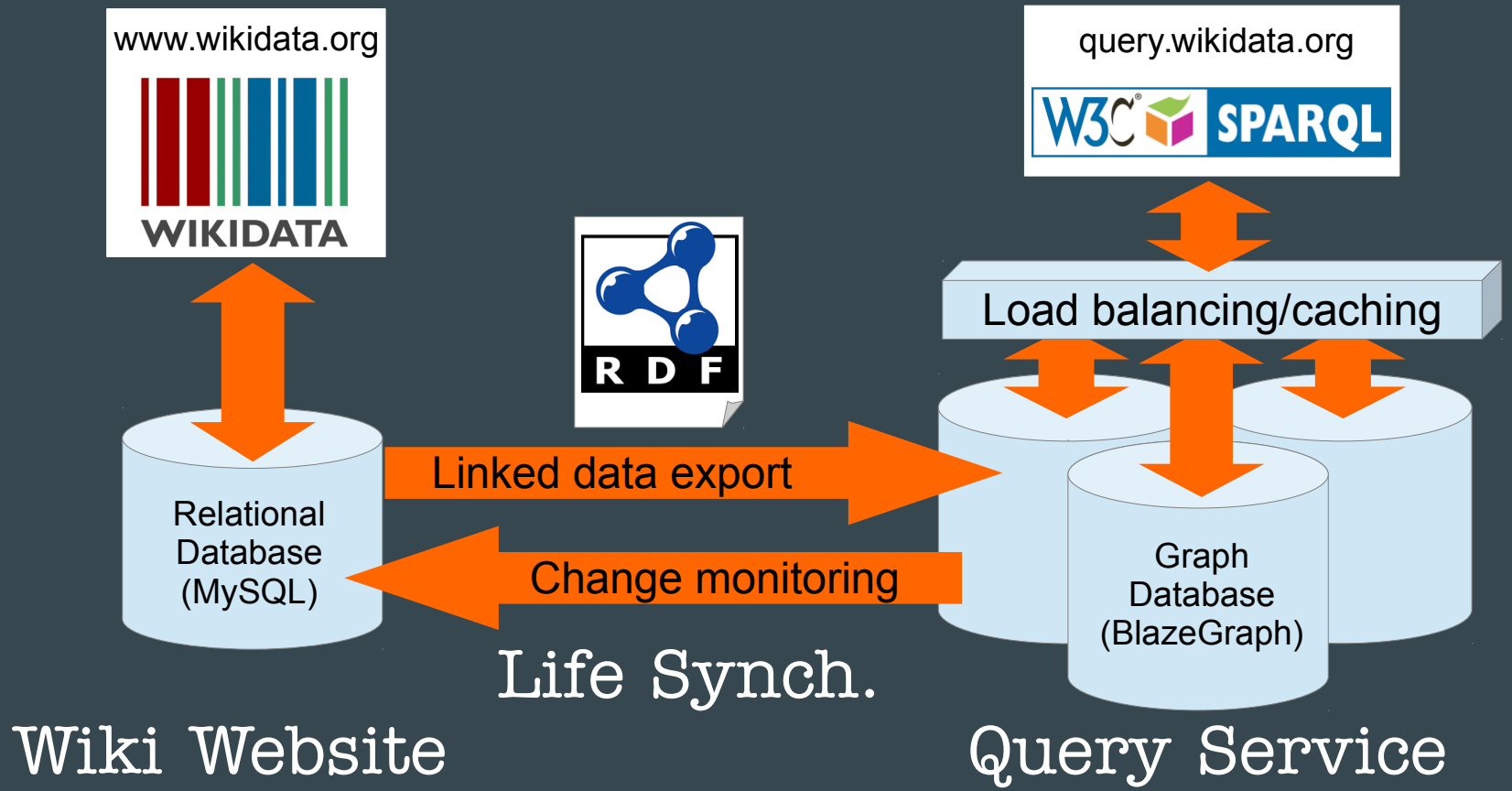
The Wikidata Query Service

Wiki Website



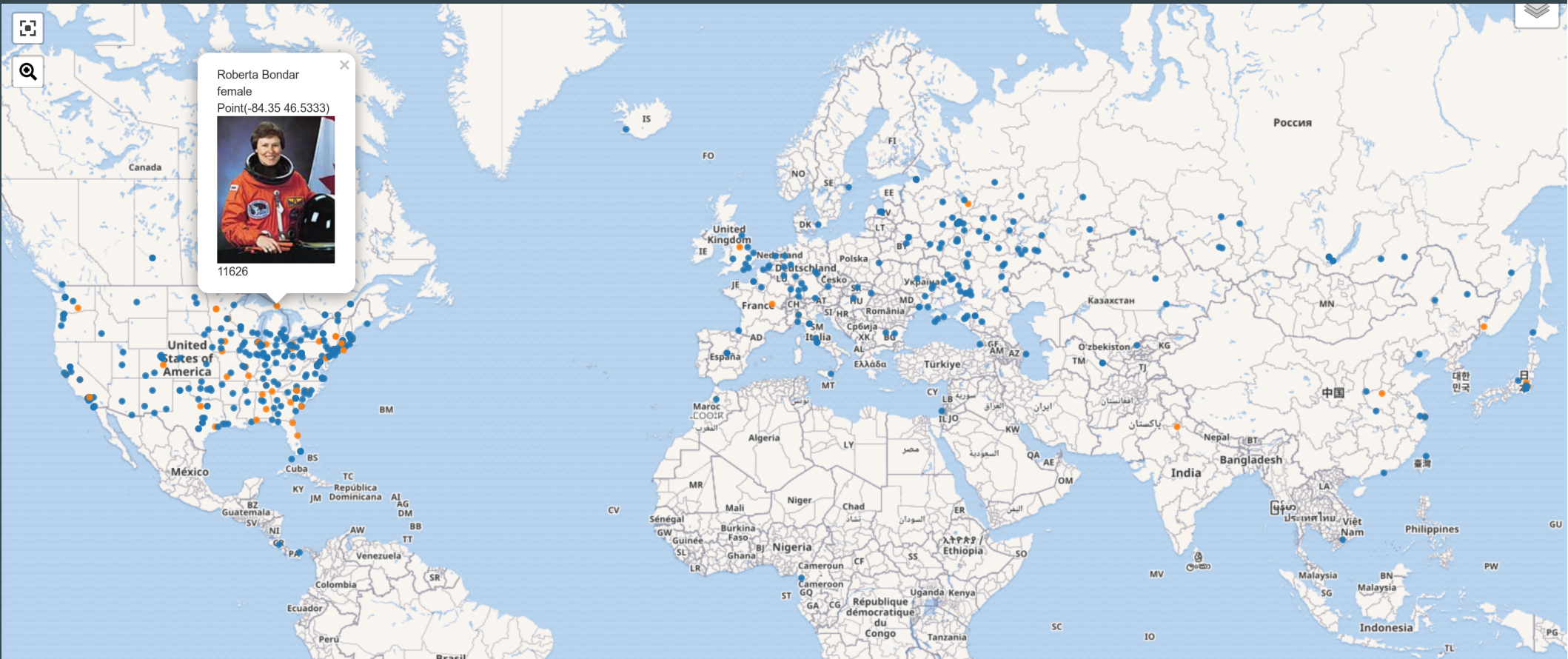
Query Service

The Wikidata Query Service



“Where are people born who travel to space?”

(Colour-coded by gender)



“Which 19th century paintings show the moon?”


commons:Xi Bannister, Moon P.jpg
Moonlight Marine
Moon


commons:Maufraleverdeluneaucrepusculei1899.jpg
Moonrise at Twilight
Moon


commons:Johan Christian Clausen Dahl - Kysten ved Lauvvig i Norge i måneskin - Thorvalds...
The coast at Lauvvig, Norway
Moon


commons:Johan Christian Claussen Dahl - Bugten ved Napoli i måneskin (1821).jpg
The Bay of Naples by moonlight
Moon


commons:'Dresden in Moonlight' by Johan Christian Dahl, Bergen Kunstmuseum.JPG
Dresden by Moonlight
Moon

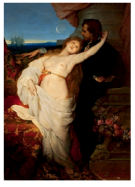

commons:Johan Christian Dahl - Den Esrum sjøen i måneskinn.jpg
Esrom Lake in Moonlight
Moon


commons:Dahl, Der Kopenhagener Hafen im Mondschein, 1831.jpg
Q24704472
Moon


commons:Dupré Landscape by moonlight.jpg
Q24699126
Moon


commons:Szermentowski Morning star.jpg
Morning star
Moon


commons:Lampi Carthusian monastery.jpg
Q23730383
Moon


commons:Max Tannhäuser.jpg
Q23009000
Moon


commons:Louis Rémy Mignot Marsh in Ecuador.jpg
Moonlight over a Marsh in Ecuador
Moon


commons:Johan Christian Dahl - View of Dresden by Moonlight - Google Art Project (NwHK-NsdlnFIMQ).jpg
View of Dresden by Moonlight
Moon


commons:Johan Christian Dahl - Dresden by Moonlight - Google Art Project.jpg
Dresden by Moonlight
Moon


commons:Elin Danielson-Gambogi Kuutamo n 1890.jpg
Moonlight
Moon


commons:Ladislav Medňanszk...
Night Travellers at a Cross
Moon


commons:Night Travellers at a ...
Night Travellers at a Cross
Moon


commons:Martin Johnson Heade - Two Owls at Sunset.jpg
Two Owls at Sunset
Moon


commons:Martin Johnson Heade -Sailing by Moonlight.jpg
Sailing by Moonlight
Moon


commons:'Point Judith, Rhode Island' by Martin Johnson Heade, 1867-68.JPG
Point Judith, Rhode Island
Moon


commons:Midwinter Moonlight by Regis-Francois Gignoux, before 1869, oil on board ...
Mid-Winter Moonlight
Moon


commons:Sanford Robinson Gifford - Crépuscule sur le mont Hunter.jpg
Hunter Mountain, Twilight
Moon


commons:Samuel Colman - The Rock of Salvation - Google Art Pr...
The Rock of Salvation
Moon


commons:John William Casilear-Mo...
Moonrise on the Coast
Moon


commons:Albert Pinkham Ryder - The Lovers' Boat (c.188...
The Lovers' Boat
Moon


commons:Johan Christian Clausen Dahl - Ved den napolitanske golf. Måneskin - Staten,...
The Gulf of Naples. Moonlight
Moon


commons:Johan Christian Clausen Dahl - Ved den napolitanske golf. Måneskin - Staten...
The Gulf of Naples. Moonlight
Moon


commons:Albert Bierstadt - Cloud ...
Cloud Study, Moonlight
Moon


commons:John Martin - The Eve of the Deluge - WGA14146.jpg
The Eve of the Deluge
Moon


commons:Thomas Chambers - Storm-Tossed Frigate.jpg
Storm-Tossed Frigate
Moon


commons:Richards William Trost Moonlight On Mount Lafayette New hampshire.jpg
Moonlight on Mount Lafayette, New Hampshire
Moon


commons:Moonrise by George Inness 1887.jpeg
Moonrise
Moon


commons:'Moonlight' by George Inness, 1893.JPG
Moonlight
Moon

“Which days of the week do disasters occur on?”

Date	Mon	Tue	Wed	Thu	Fri	Sat	Sun
1	25	33	22	18	26	28	23
2	24	26	23	23	22	32	12
3	24	27	21	31	23	28	38
4	24	25	33	25	26	26	24
5	37	23	32	18	19	17	29
6	25	28	32	20	24	33	22
7	18	22	25	16	22	18	17
8	32	28	19	25	22	23	19
9	20	25	29	29	27	21	23
10	20	20	19	14	25	25	29
11	30	34	28	23	22	20	20
12	41	33	27	30	20	20	23
13	35	26	29	21	25	24	25
14	24	23	27	23	22	28	17

“The free knowledge base that anyone can edit”

Entities: 50M

Statements: 570M

Labels: 260M

Descriptions: 1.5B

Links to Wikis: 65M

“The free knowledge base that anyone can edit”

Editors: >230K

Tim Berners-Lee (Q80)

British computer scientist, inventor of the World Wide Web

edit

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | T.J. Berners-Lee

Tim Berners-Lee (Q80)

British computer scientist, inventor of the World Wide Web

edit

TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | T.J. Berners-Lee

place of birth	London	edit
	1 reference	
		+ add value

Tim Berners-Lee (Q80)

British computer scientist, inventor of the World Wide Web

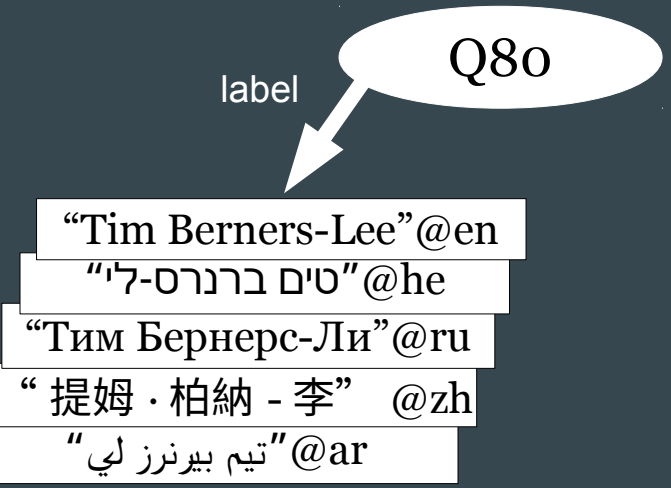
edit

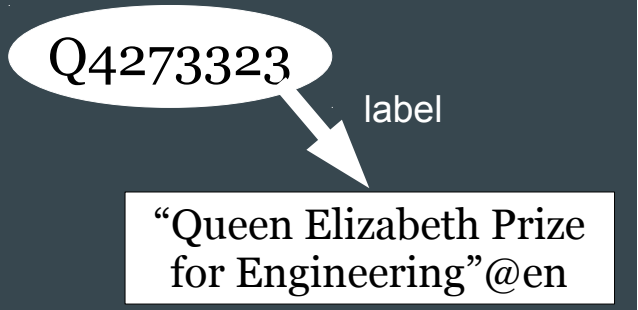
TimBL | Sir Tim Berners-Lee | Timothy John Berners-Lee | TBL | Tim Berners Lee | T. Berners-Lee | T Berners-Lee | T.J. Berners-Lee

place of birth	London ▸ 1 reference	edit
		+ add value
award received	Queen Elizabeth Prize for Engineering point in time: 2013 together with: Robert Kahn, Vint Cerf, Louis Pouzin, Marc Andreessen ▸ 1 reference	edit

From Wikidata (rich graphs) to RDF (plain graphs)

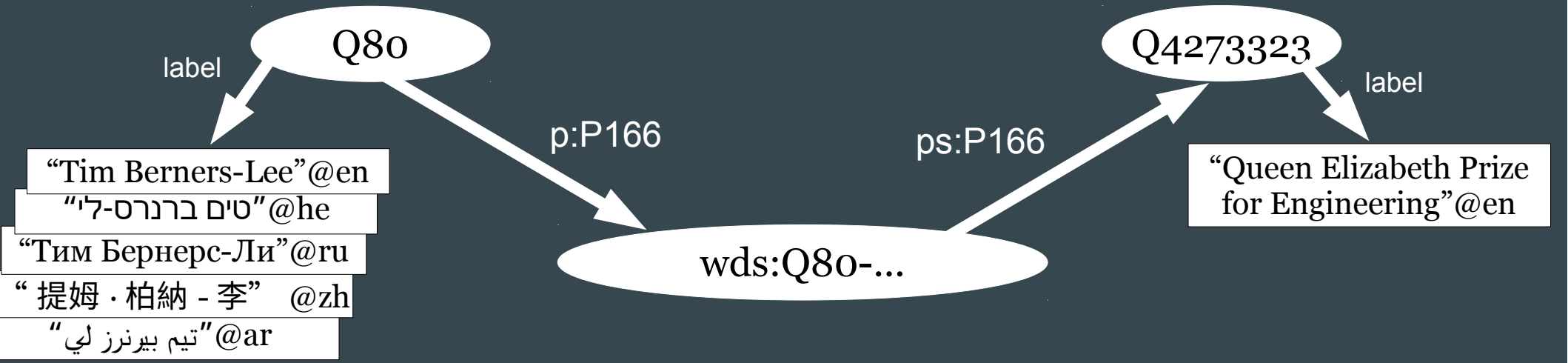
[Erxleben et al., ISWC 2014]





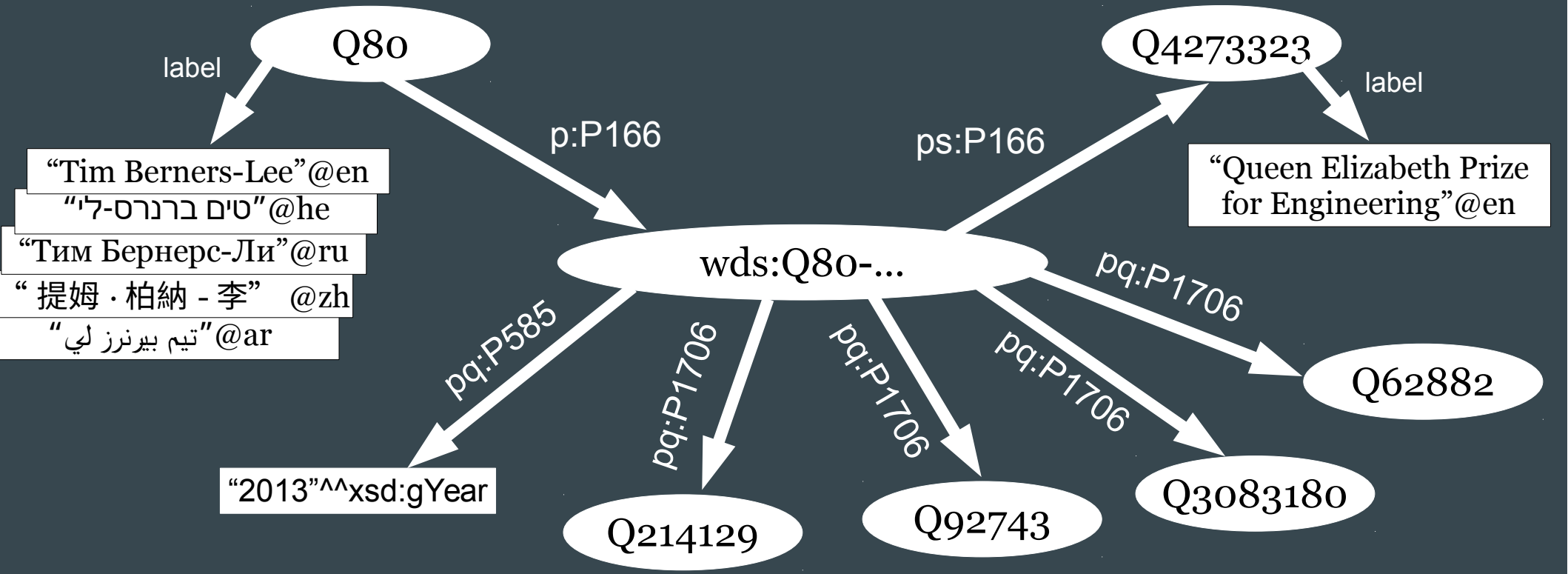
From Wikidata (rich graphs) to RDF (plain graphs)

[Erxleben et al., ISWC 2014]



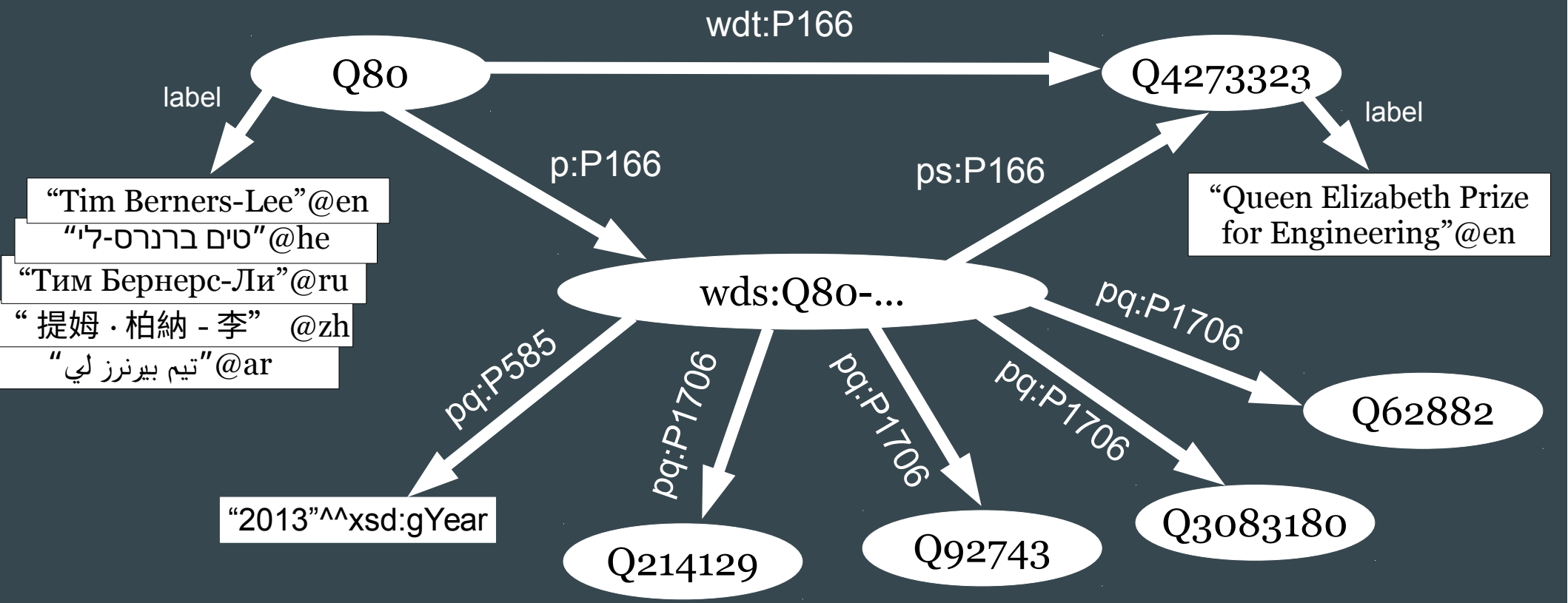
From Wikidata (rich graphs) to RDF (plain graphs)

[Erxleben et al., ISWC 2014]



From Wikidata (rich graphs) to RDF (plain graphs)

[Erxleben et al., ISWC 2014]



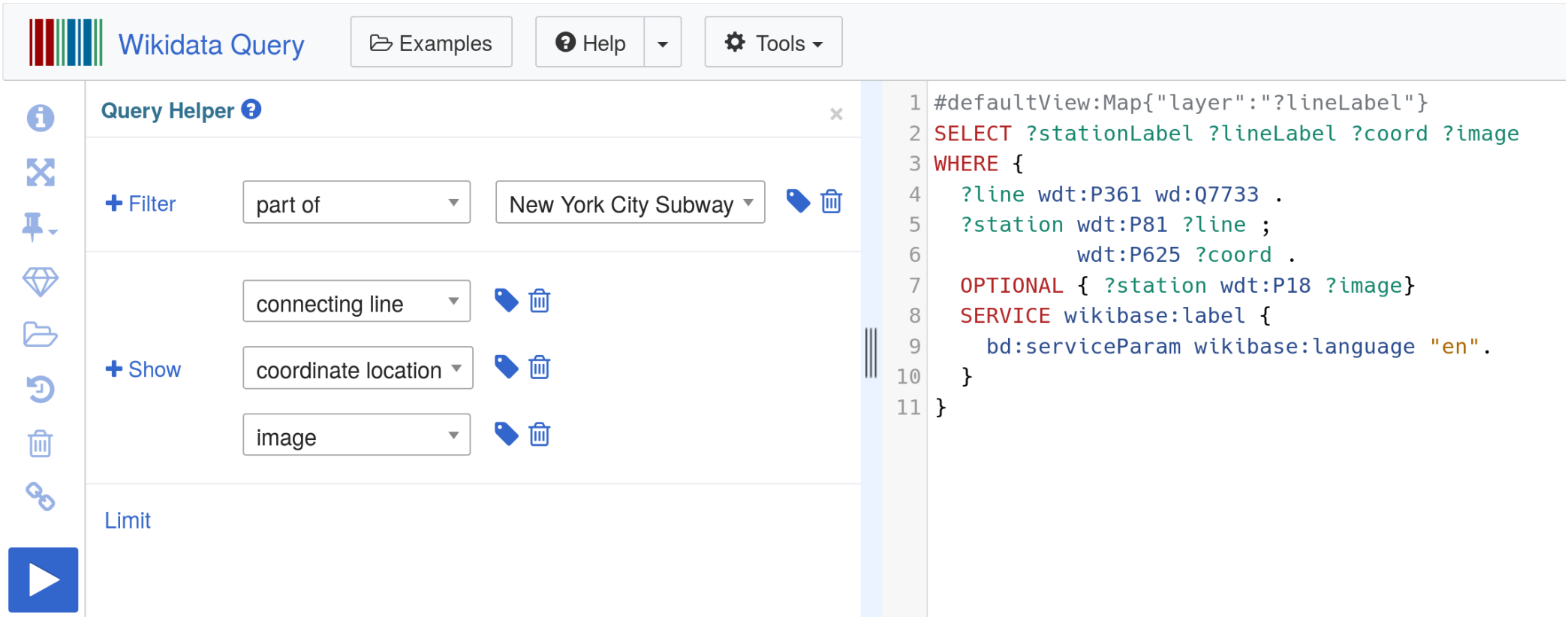
Wikidata RDF Exports

- Weekly full dumps
  - Currently 6.2 billion triples (42 GB Turtle gzip compressed)
  - At https://dumps.wikimedia.org/wikidatawiki/entities/
- Linked Data Exports
  - Live data in many formats
  - E.g., http://www.wikidata.org/wiki/Special:EntityData/Q42.nt

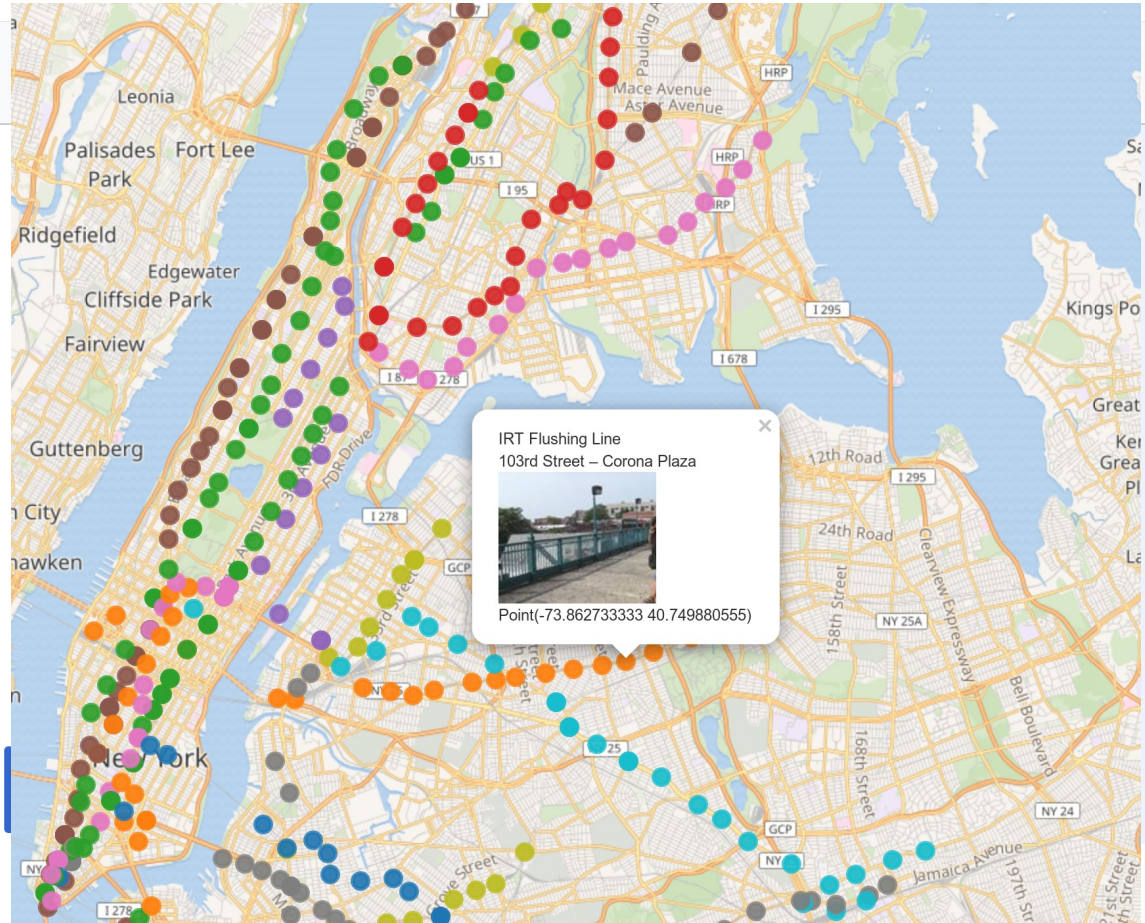
Wikidata SPARQL Query Service

- Official query service since mid 2015
  - User interface at https://query.wikidata.org/
- All the data (6.2B triples), live (latency<60s)
- No limits (well, almost):
  - 60sec timeout
  - No limit on result size (!)
  - No limit on parallel queries, but CPU-time budget per client
- Extra SERVICEs in SPARQL (geo, Wikipedia API, labels, ...)

A simple SPARQL query

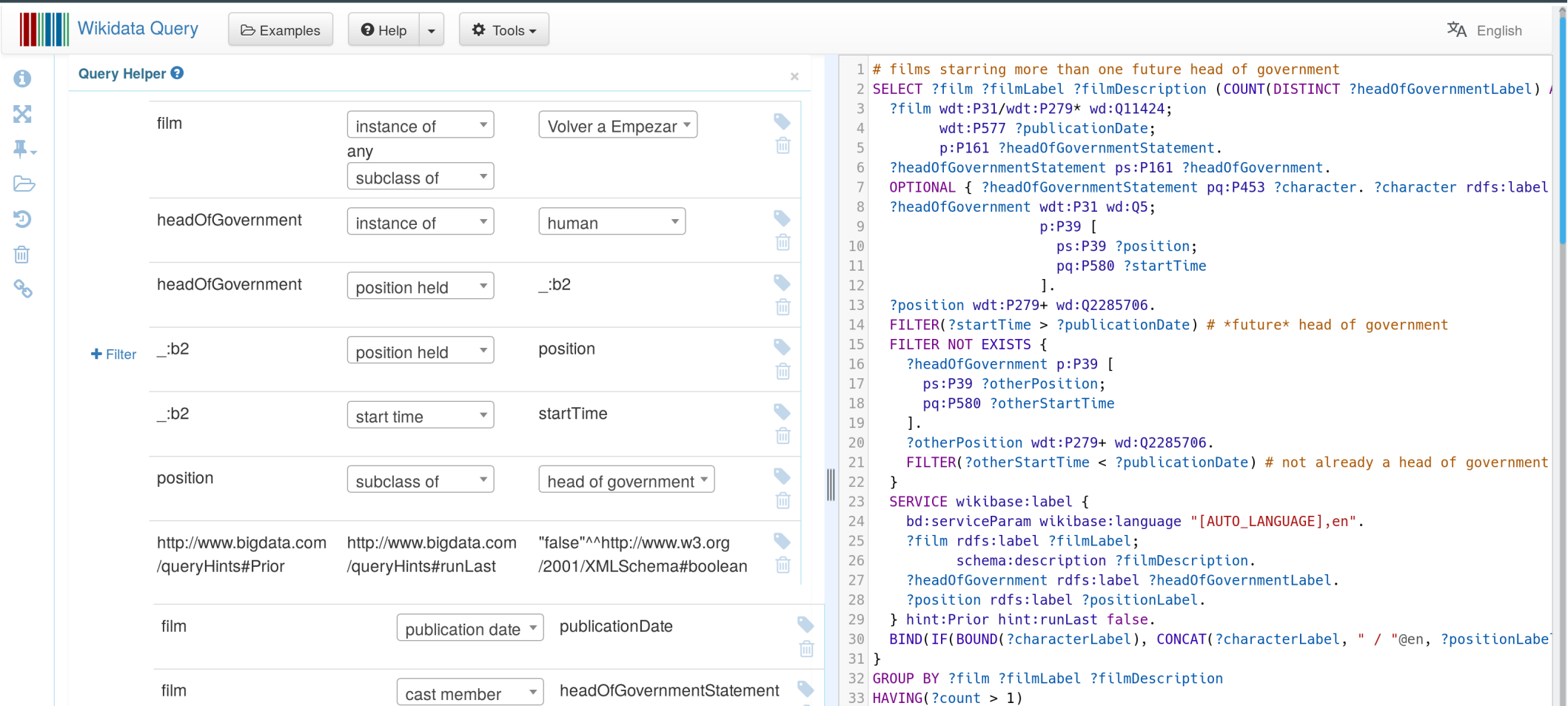


A simple SPARQL query




```
#defaultView:Map{"layer":"?lineLabel"}
SELECT ?stationLabel ?lineLabel ?coord ?image
WHERE {
  ?line wdt:P361 wd:Q7733 .
  ?station wdt:P81 ?line ;
           wdt:P625 ?coord .
  OPTIONAL { ?station wdt:P18 ?image}
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en".
  }
}
```

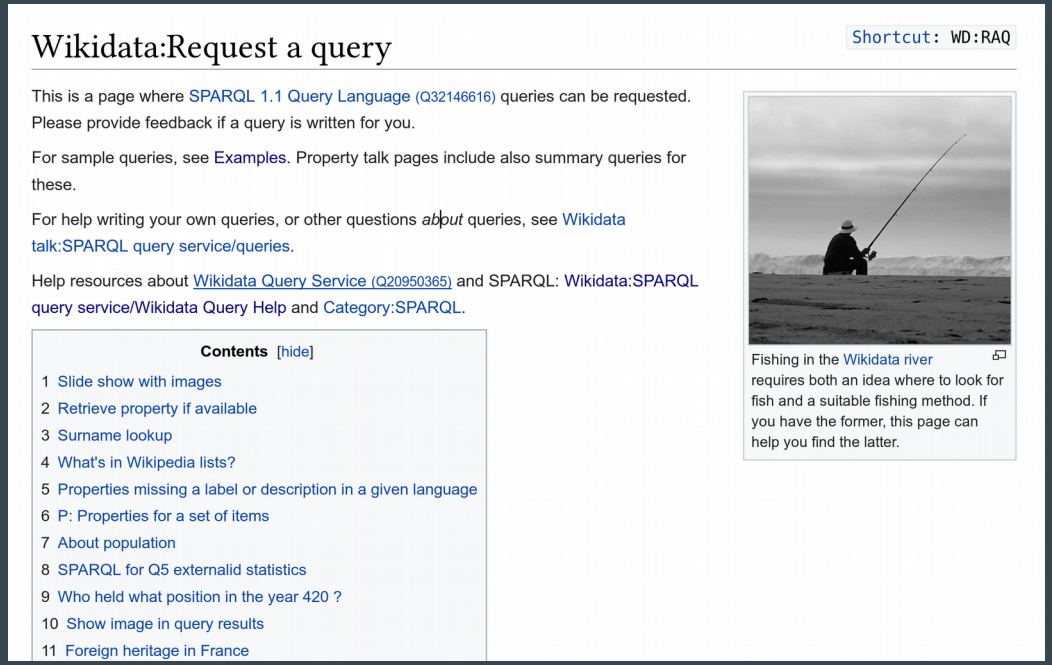
An advanced SPARQL query



“It’s too complicated!”

“It’s not too complicated!”

- SPARQL is widely used
  - >100M requests per month (3.8M per day) in 2018
- It’s an API – most users are not in direct contact
- The community offers tutorials, workshops and support services



Wikidata:Request a query

Shortcut: WD:RAQ

This is a page where SPARQL 1.1 Query Language (Q32146616) queries can be requested. Please provide feedback if a query is written for you.

For sample queries, see Examples. Property talk pages include also summary queries for these.

For help writing your own queries, or other questions *about* queries, see Wikidata talk:SPARQL query service/queries.

Help resources about Wikidata Query Service (Q20950365) and SPARQL: Wikidata:SPARQL query service/Wikidata Query Help and Category:SPARQL.

Contents [hide]

1 Slide show with images
2 Retrieve property if available
3 Surname lookup
4 What's in Wikipedia lists?
5 Properties missing a label or description in a given language
6 P: Properties for a set of items
7 About population
8 SPARQL for Q5 externalid statistics
9 Who held what position in the year 420 ?
10 Show image in query results
11 Foreign heritage in France

Fishing in the Wikidata river requires both an idea where to look for fish and a suitable fishing method. If you have the former, this page can help you find the latter.

“It does not scale!”

“It does ~~not~~ scale!”

- Excellent availability and performance
  - 50% of queries answered in <40ms (95% in <440ms; 99% in <40s)
  - Less than 0.05% of queries time out
  - Service has never been down so far
- Affordable system setup:
  - Three commodity servers (+three for geo-redundancy)
  - Standard Linux load balancing + standard HTTP cache
- All software/customisations free & open source
  - See https://github.com/wikimedia/wikidata-query-rdf

So what are those 100Ms of queries?

- We looked at 481,716,280 queries logged during 24 weeks
- Analysing SPARQL query activity is hard
  - Extreme influence of scripts and bots
  - Does not average out over time, each month looks rather different!
  - → Classify major sources (bots) and isolate “organic” part of the traffic

Robotic and organic traffic

Robotic traffic

- Dominates (60% of queries by top-3 bots)
- Mostly data integration and data download
- More uniform, shorter

Organic traffic

- Much smaller volume (0.6% of all queries)
- Browsers, mobile apps, miscellaneous
- More diverse, longer

Path queries are very important

- Reified statements in 4%–10% of queries

See for yourself!

- We have released complete, timestamped query logs
  - Anonymised to avoid user identification
  - With limited user agent information
  - Full dataset, no sample!
- Currently 12 weeks in 2017 – more to come soon

https://kbs.inf.tu-dresden.de/WikidataSPARQL

Conclusions

- Semantic web technology – it works!
  - Interactive analytics and query is affordable for dynamic knowledge graphs with >10^9 edges
  - Usable for large, open communities without prior RDF/SPARQL experience
  - We want more applications & more research!

Thanks

- Denny Vrandecic, Lydia Pintscher, and the whole Wikimedia Deutschland e.V. team in Berlin who made Wikidata possible
- Brad Bebee, Bryan Thompson, and all of the BlazeGraph team
- Anyone contributing to RDF and SPARQL libraries
- All who contributed to W3C standards used here, esp. SPARQL
- The Wikidata community
- TimBL ;-)

Films with future heads of government

Star in the Dust	1956 film by Charles F. Haas	2	Clint Eastwood, mayor; George Wallace, Governor of Alabama
The Two Who Stole the Moon	1962 Polish film by Jan Batory	2	Jarosław Kaczyński, Prime Minister of Poland; Lech Kaczyński, Mayor of Warsaw
Ragasiya Police 115	1968 film by B. R. Panthulu	2	M. G. Ramachandran, Chief Minister of Tamil Nadu; Jayalalithaa, Chief Minister of Tamil Nadu
Québec : Duplessis et après…	documentary	2	Bernard Landry, Premier of Quebec; René Lévesque, Premier of Quebec
Q3541438	1994 film by Claude Lanzmann	2	Ariel Sharon, Prime Minister of Israel; Ehud Barak, Prime Minister of Israel
Batman & Robin	1997 American superhero film based on the DC Comics character Batman	2	Arnold Schwarzenegger, Mr. Freeze / Governor of California; Jesse Ventura, Governor of Minnesota

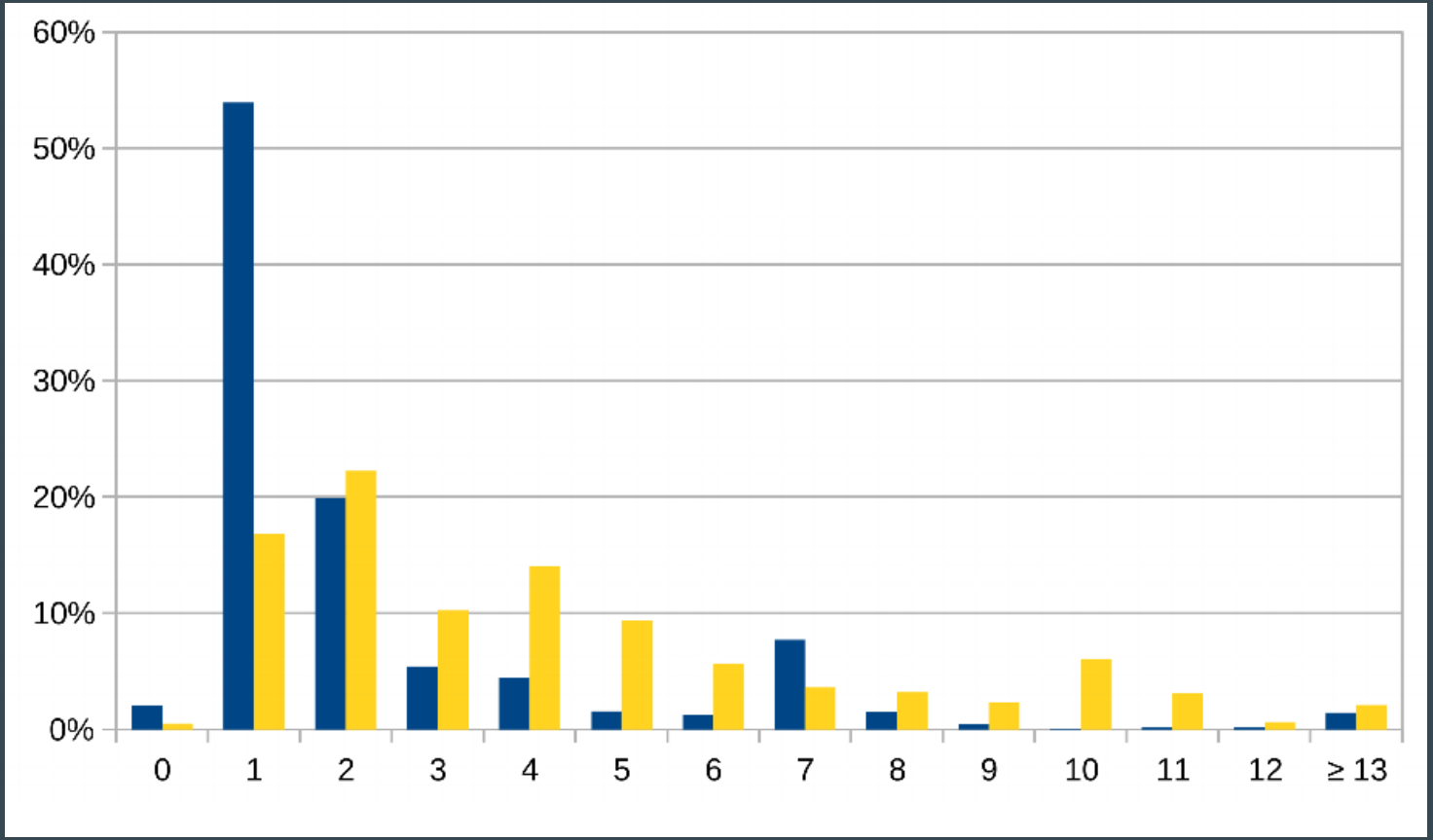
Literature

- Stanislav Malyshev, Markus Krötzsch, Larry González, Julius Gonsior, Adrian Bielefeldt: “Getting the Most out of Wikidata: Semantic Technology Usage in Wikipedia’s Knowledge Graph” In Denny Vrandečić, et al., eds., Proceedings of the 17th International Semantic Web Conference (ISWC'18)
- Adrian Bielefeldt, Julius Gonsior, Markus Krötzsch: “Practical Linked Data Access via SPARQL: The Case of Wikidata” Proceedings of the WWW2018 Workshop on Linked Data on the Web (LDOW-18), CEUR Workshop
- Fredo Erxleben, Michael Günther, Markus Krötzsch, Julian Mendez, Denny Vrandečić: “Introducing Wikidata to the Linked Data Web” In Proceedings of the 13th International Semantic Web Conference (ISWC 2014)

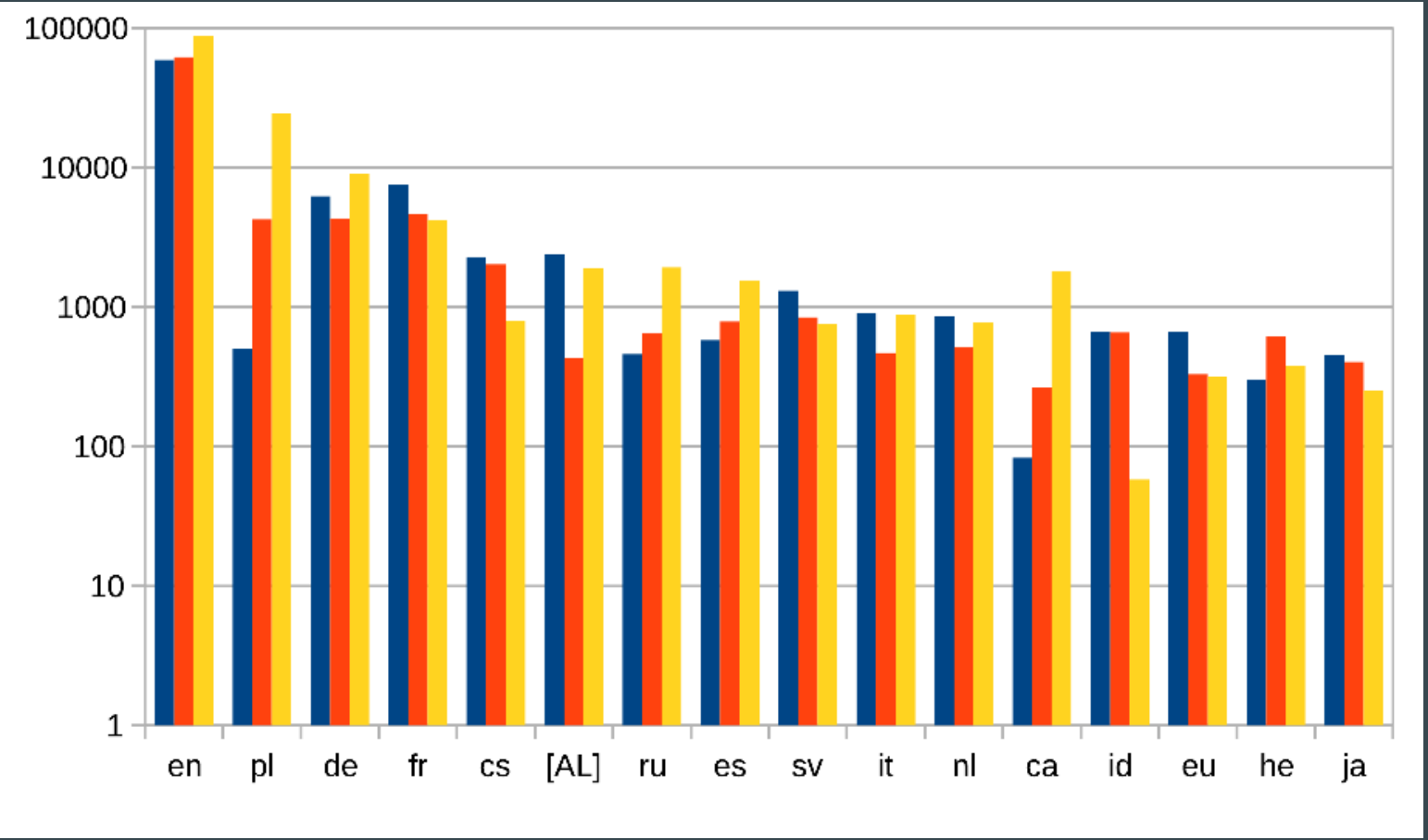
SPARQL Feature Distribution (2017/2018)

	organic						robotic					
	Jun 2017	Jul 2017	Aug 2017	Jan 2018	Feb 2018	Mar 2018	Jun 2017	Jul 2017	Aug 2017	Jan 2018	Feb 2018	Mar 2018
Limit	31.08%	39.55%	46.56%	52.31%	51.23%	36.87%	21.12%	16.86%	17.42%	20.38%	11.47%	15.17%
Distinct	26.50%	31.40%	19.05%	59.30%	60.42%	63.78%	15.84%	5.48%	4.27%	4.32%	7.54%	12.25%
Order By	17.29%	14.75%	13.22%	46.89%	46.99%	34.53%	12.97%	8.01%	6.78%	8.76%	7.68%	17.46%
Offset	0.40%	2.92%	0.37%	0.09%	0.08%	0.06%	7.73%	6.07%	6.29%	0.10%	0.07%	0.10%
Join	87.59%	87.82%	89.76%	82.50%	91.70%	87.02%	88.48%	78.53%	67.41%	73.26%	61.39%	70.19%
Optional	42.36%	46.24%	55.92%	50.90%	41.30%	41.15%	25.08%	11.63%	11.45%	12.73%	15.41%	30.71%
Filter	25.89%	29.12%	22.24%	12.59%	11.76%	11.76%	21.64%	17.92%	13.79%	14.70%	16.83%	29.02%
Path with *	15.02%	15.59%	12.88%	40.92%	32.43%	30.34%	16.43%	19.19%	14.80%	20.56%	17.26%	24.81%
Subquery	13.09%	15.30%	12.79%	6.45%	5.07%	5.39%	0.34%	0.28%	0.33%	0.09%	0.13%	0.11%
Bind	9.85%	9.23%	8.68%	4.72%	3.99%	4.15%	16.29%	12.07%	9.60%	11.94%	13.79%	24.03%
Union	5.10%	5.76%	12.62%	2.56%	2.07%	3.39%	11.26%	8.63%	7.61%	13.96%	13.05%	18.57%
Values	4.44%	3.07%	10.88%	3.29%	3.23%	3.20%	35.72%	30.74%	28.92%	29.82%	23.80%	11.90%
Not Exists	3.31%	3.37%	2.46%	1.24%	0.94%	0.69%	0.19%	0.21%	0.19%	0.27%	0.29%	0.35%
Minus	2.04%	2.91%	1.60%	0.82%	0.57%	0.71%	0.53%	0.92%	1.07%	1.46%	1.26%	1.78%
Service (lang)	44.63%	42.09%	54.78%	50.88%	41.71%	42.95%	10.40%	6.15%	4.27%	7.15%	7.91%	8.90%
Service (other)	11.49%	10.53%	10.32%	7.30%	13.14%	2.31%	4.51%	0.19%	1.16%	0.17%	0.18%	0.51%
Group By	17.12%	19.93%	13.04%	7.00%	5.40%	5.07%	0.41%	0.37%	0.48%	0.22%	0.23%	0.39%
Sample	8.85%	10.93%	4.60%	1.61%	1.63%	1.69%	0.04%	0.04%	0.06%	0.05%	0.04%	0.10%
Count	7.55%	7.60%	8.15%	5.22%	3.88%	3.73%	1.15%	4.30%	0.30%	1.52%	0.65%	0.89%
GroupConcat	1.80%	2.79%	1.17%	0.86%	0.86%	0.74%	0.06%	0.09%	0.02%	0.03%	0.02%	0.28%
Having	1.17%	1.14%	0.72%	0.65%	0.26%	0.33%	0.01%	0.01%	0.00%	0.00%	0.00%	0.04%

Triples per query: organic (blue)/robotic (yellow)



Languages of labels in organic queries



SPARQL feature co-occurrence

Table 3. Co-occurrence of SPARQL features in percent of total queries per dataset (Join, Filter, Optional, Union, Path, Values, Subquery)

							organic		robotic	
J	**F**	**O**	**U**	**P**	**V**	**S**	I1–I3	I4–I6	I1–I3	I4–I6
		(none)					8.18	9.35	19.67	27.96
J							15.23	32.31	10.81	10.10
	F						1.09	0.94	1.95	1.27
J	F						8.87	2.37	2.61	1.50
J				P			2.93	1.63	13.72	14.09
J	F			P			2.49	0.58	0.39	0.06
J					V		0.41	2.04	30.91	17.65
		O					1.28	1.61	0.12	0.64
J		O					25.97	7.16	1.88	1.95
J		O		P			2.10	28.41	0.36	0.05
J	F	O					3.30	1.30	1.78	1.19
J		O	U				3.57	0.25	0.02	0.00
J		O			V		3.45	0.40	0.11	0.43
J		O		P	V		1.01	0.06	0.16	0.04
J						S	0.86	1.43	0.06	0.01
J		O				S	1.64	0.63	0.00	0.01
J	F					S	0.64	2.17	0.02	0.01
J	F	O		P			0.87	0.31	0.65	1.60
J			U	P	V		0.01	0.01	0.05	1.94
						All cases shown	**83.90**	**92.96**	**85.27**	**80.50**