Getting the Most Out of Wikidata

Markus Krötzsch
Knowledge-Based Systems, TU Dresden

Reporting on joint work with
Adrian Bielefeldt, Fredo Erxleben, Julius Gonsior,
Larry Gonzalez, Michael Günther, Stas Malyshev,
Julian Mendez, Veronica Thost, and Denny Vrandecic

and supported by the Wikimedia Foundation

Wiki Workshop 2018
“What are the world’s largest cities with a female mayor?”
“What are the world’s largest cities with a female mayor?”

<table>
<thead>
<tr>
<th>cityLabel</th>
<th>mayorLabel</th>
<th>population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokyo</td>
<td>Yuriko Koike</td>
<td>13742906</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>Carrie Lam</td>
<td>7336585</td>
</tr>
<tr>
<td>Baghdad</td>
<td>Zekra Alwach</td>
<td>6960000</td>
</tr>
<tr>
<td>Surabaya</td>
<td>Tri Rismaharini</td>
<td>4975000</td>
</tr>
<tr>
<td>Yokohama</td>
<td>Fumiko Hayashi</td>
<td>3733234</td>
</tr>
<tr>
<td>Madrid</td>
<td>Manuela Carmena</td>
<td>3182981</td>
</tr>
<tr>
<td>Rome</td>
<td>Virginia Raggi</td>
<td>2873494</td>
</tr>
<tr>
<td>Kaohsiung City</td>
<td>Chen Chu</td>
<td>2777384</td>
</tr>
<tr>
<td>Antananarivo</td>
<td>Lalao Ravalomanana</td>
<td>2610000</td>
</tr>
<tr>
<td>Paris</td>
<td>Anne Hidalgo</td>
<td>2206488</td>
</tr>
</tbody>
</table>
“Where are people born who travel to space?”
(Colour-coded by gender)
“Which days of the week do disasters occur on?”

<table>
<thead>
<tr>
<th>Date</th>
<th>Mon</th>
<th>Tue</th>
<th>Wed</th>
<th>Thu</th>
<th>Fri</th>
<th>Sat</th>
<th>Sun</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25</td>
<td>33</td>
<td>22</td>
<td>18</td>
<td>26</td>
<td>28</td>
<td>23</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>26</td>
<td>23</td>
<td>23</td>
<td>22</td>
<td>32</td>
<td>12</td>
</tr>
<tr>
<td>3</td>
<td>24</td>
<td>27</td>
<td>21</td>
<td>31</td>
<td>23</td>
<td>28</td>
<td>38</td>
</tr>
<tr>
<td>4</td>
<td>24</td>
<td>25</td>
<td>33</td>
<td>25</td>
<td>26</td>
<td>26</td>
<td>24</td>
</tr>
<tr>
<td>5</td>
<td>37</td>
<td>23</td>
<td>32</td>
<td>18</td>
<td>19</td>
<td>17</td>
<td>29</td>
</tr>
<tr>
<td>6</td>
<td>25</td>
<td>28</td>
<td>32</td>
<td>20</td>
<td>24</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>22</td>
<td>25</td>
<td>16</td>
<td>22</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>32</td>
<td>28</td>
<td>19</td>
<td>25</td>
<td>22</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>25</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>21</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>20</td>
<td>19</td>
<td>14</td>
<td>25</td>
<td>25</td>
<td>29</td>
</tr>
<tr>
<td>11</td>
<td>30</td>
<td>34</td>
<td>28</td>
<td>23</td>
<td>22</td>
<td>20</td>
<td>20</td>
</tr>
<tr>
<td>12</td>
<td>41</td>
<td>33</td>
<td>27</td>
<td>30</td>
<td>20</td>
<td>20</td>
<td>23</td>
</tr>
<tr>
<td>13</td>
<td>35</td>
<td>26</td>
<td>29</td>
<td>21</td>
<td>25</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>14</td>
<td>24</td>
<td>23</td>
<td>27</td>
<td>23</td>
<td>22</td>
<td>28</td>
<td>17</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>32</td>
<td>22</td>
<td>24</td>
<td>16</td>
<td>32</td>
<td>15</td>
</tr>
</tbody>
</table>
“Which 19th century paintings show the moon?”
“Which UK bands’ members’ average age is 70 or greater?”

<table>
<thead>
<tr>
<th>bandLabel</th>
<th>avgAge</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Searchers</td>
<td>80</td>
</tr>
<tr>
<td>Rory Storm</td>
<td>79</td>
</tr>
<tr>
<td>The Animals</td>
<td>77</td>
</tr>
<tr>
<td>Quintessence</td>
<td>77</td>
</tr>
<tr>
<td>The Corporation</td>
<td>76.2</td>
</tr>
<tr>
<td>The Hollies</td>
<td>76</td>
</tr>
<tr>
<td>Cream</td>
<td>75.666666666666666667</td>
</tr>
<tr>
<td>Soft Machine</td>
<td>75.5</td>
</tr>
<tr>
<td>The Moody Blues</td>
<td>75.333333333333333333</td>
</tr>
</tbody>
</table>
“The free knowledge base that anyone can edit”
Wikipedia and the Semantic Web - The Missing Links

- **Author(s):** Markus Krötzsch & Denny Vrandečić & Max Völkel
- **License:** CC-NC-SA 2.0 (for further license models, please contact the authors)
- **Slides:** collected but not uploaded yet
- **Video:** 16:44 [link] (talk given by Denny Vrandečić)
- **Note:** Presentation, paper also at [Wikipedia and the Semantic Web] (PDF, 164K)

**About the author:** The authors are research associates at the Institute of Applied Informatics and Formal Description Methods (AIFB), University of Karlsruhe, Germany, where they are members of the AIFB Research Group Knowledge Management, an interdisciplinary team of computer scientists, mathematicians, and industrial engineers that is one of the world-wide leading institutions in the Semantic Web research community. Other relevant research topics include Semantic Web, ontologies, data and text mining, logic-based knowledge representation, peer-to-peer, and Web services.

Being enthusiastic users and contributors of various Wikis, the authors also have special interest in making emerging semantic technologies available within Wikis, where computer-assisted organization and processing of knowledge plays an important role.
### Lyon

From Wikipedia, the free encyclopedia

For other uses, see Lyon (disambiguation)

**Lyon** (UK: /ˈlɔːn/,[3] US: /ˈlɔːn/; French: [liɛ̃] (listen)), locally [lijɔ̃]; Arpitan: Liyôn [lijɔ̃]), also known in British English as Lyons (/ˈlaɪənz/), is the third-largest city of France. It is located in the

### France

From Wikipedia, the free encyclopedia

"La France" redirects here. For other uses of “La France”, see France (disambiguation)

**France** (French: [frɑ̃s]), officially the **French Republic** (French: République française [ʁepəblik fʁɑ̃sɛs]), is a country whose territory consists of metropolitan France in western Europe, as well as

### Louis Néel

From Wikipedia, the free encyclopedia

**Louis Eugène Félix Néel** ForMemRS (22 November 1904 – 17 November 2000) was a French physicist born in Lyon.[2]

### Nobel Prize in Physics

From Wikipedia, the free encyclopedia

The **Nobel Prize in Physics** (Swedish: Nobelpriset i fysik) is a yearly award given by the **Royal Swedish Academy of Sciences** for those who conferred the most outstanding contributions for mankind in the field of physics. It is one of the

---


---

---

---

---
Semantic MediaWiki (2005): From Links to Graphs

Lyon → France
Louis Néel → Nobel Prize in Physics
Louis Néel → place of birth
Louis Néel → award received

country
place of birth
award received
Links are not Enough: Adding Datatypes

- Lyon
- France
- Louis Néel
- Nobel Prize in Physics

- Country
- Place of birth
- Award received
Links are not Enough: Adding Datatypes

- Lyon: country
- France: population (66628000)
- Louis Néel: date of birth (1904-11-22), place of birth
- Nobel Prize in Physics: award received
Louis Néel was born on November 22, 1904. He was awarded the Nobel Prize in Physics in 1970.
Wikidata: One Graph for Many Languages

Q456 (country) — Q142 (award received) — Q155781 (date of birth) — Q38104 (population)

1904-11-22

"Louis Néel" @en
"Louis Néel" @fr
"Нэль, Луи Эжен Феликс" @ru
"ليوس نيل" @ur

"Nobel Prize in Physics" @en
"פרס נובל לפיזיקס" @he
"prix Nobel de physique" @fr
"诺贝尔物理学奖" @zh
"جائزة نوبل في الفيزياء" @ar
Louis Néel (Q155781)

French physicist
Louis Neel | Louis Eugène Felix Néel

award received

Nobel Prize in Physics

point in time 1970
together with Hannes Alfvén
prize money 200,000 Swedish krona

2 references

reference URL http://www.nobelprize.org/nobel_prizes/physics/laurcates
Wikidata in April 2018

- >400M statements on >45M entities
- >60M links to Wikipedia articles
- >200M labels and aliases
- >1,200M disambiguating descriptions
- >200K registered contributors
“How to query Wikidata?”
The Wikidata Query Service

www.wikidata.org

Relational Database (MySQL)

Wiki Website
The Wikidata Query Service

Wiki Website

Relational Database (MySQL)

Linked data export

Change monitoring

Graph Database (BlazeGraph)

Load balancing/caching

Query Service

Wiki Website

Linked data export

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

**Q155781**

P166 (award received)

**Q38104**

P585 (point in time): 1970
P1706 (together with): Q54945 (H. Alfvén)
P2121 (prize money): 200000 Q122922 (SEK)
From Wikidata (rich graphs) to RDF (plain graphs)

Q155781 wdt:P166 Q38104
From Wikidata (rich graphs) to RDF (plain graphs)

Q155781

wds:Q155781...

Q38104

wdt:P166

"1970"^^xsd:gYear

Q54945

pq:P1706

"200000"^^xsd:decimal

pq:P585

pq:P2121

p:P166

p:Q155781

ps:P166
From Wikidata (rich graphs) to RDF (plain graphs)

- Statements get own objects in graph
- Some simple statements also stored directly
- Each Wikidata property becomes many RDF properties
- Complex values get own objects too (not shown)
Wikidata RDF Exports

- **Weekly full dumps**
  - Currently 4.9 billion triples (32 GBit Turtle 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 (4.9B triples), live (latency<60s)
- No limits (well, almost):
  - 60sec timeout
  - No limit on result size (!)
  - No limit on query numbers per IP
  - Clients might be paused after too many parallel requests
A simple SPARQL query

```sparql
#defaultView:Map{"layer":)?lineLabel"}

SELECT ?stationLabel ?lineLabel ?coord ?image
WHERE {
  ?line wdt:P361 wd:Q1552 .
  ?station wdt:P81 ?line;
      wdt:P625 ?coord .
  OPTIONAL {?station wdt:P18 ?image}
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en"
  }
}
```
A simple SPARQL query

```sparql
#defaultView:Map{"layer":?lineLabel}
SELECT ?stationLabel ?lineLabel ?coord ?image
WHERE {
  ?line wdt:P361 wd:Q1552 .
  ?station wdt:P81 ?line;
  wdt:P625 ?coord .
  OPTIONAL {?station wdt:P18 ?image}
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en"
  }
}
```
A not-so-simple SPARQL query

```sparql
# Constituencies for the election to the German Bundestag 2017, with winning candidate and party

SELECT DISTINCT ?district ?districtLabel WHERE {
  SERVICE wikibase:label {
    bd:serviceConfig wikibase:language "en".
    SERVICE wikibase:label { bd:serviceConfig wikibase:language "en".
      FILTER(lang(?partyLabel) = "en")
    }.
    ?district wdt:P532 ?districtNumber.
    FILTER(lang(?districtLabel) = "en")
    FILTER(lang(?partyLabel) = "en")
  }
  FILTER(lang(?districtLabel) = "en")
  FILTER(lang(?partyLabel) = "en")
}.

# Find districts with shape


#_successful candidate for 19th German Bundestag with party and % votes


# District number in 2017 Bundestag constituencies

p:P528 ?districtNumberString.

# Turn string district number into integer

BIND(?districtNumberString AS ?districtNumber).

# Sanity check

OPTIONAL {
  p:P400 ?party.
}. 

BIND(true AS ?sanityCheckMB).

# Find original color of party

party wdt:P4627/SPARQL P485 trgbOriginal.

# Shade color depending on % votes, knowing that the original colors are only composed of FF, BB, GG: shift BB to AB or CB, and BB to AB or BB

BIND(?tvsowsPercentage == (100/2).

trgbOriginal,
   IF(?tvsowsPercentage >= (100/3),
   REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLACE(REPLAC

ORDER BY ?districtNumber
```
A not-so-simple SPARQL query
An advanced SPARQL query

```
# films starring more than one future head of government
WHERE {
  ?film wdt:P31 wdt:P279+ wd:Q11424;
  ?film wdt:P577 ?publicationDate;
  FILTER(?startTime > ?publicationDate) # *future* head of government
  FILTER NOT EXISTS {
    FILTER(?otherStartTime < ?publicationDate) # not already a head of government
  }.
  SERVICE wikibase:label {
    bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en".
  }
  schema:description ?filmDescription.
  BIND(If(BOUND(?characterLabel), CONCAT(?characterLabel, " / ") + @en, ?positionLabel)
  GROUP BY ?film ?filmLabel ?filmDescription
  HAVING(?count > 1)
}
```

http://www.bigdata.com/queryHints#Prior
http://www.bigdata.com/queryHints#runLast
"false"^^http://www.w3.org/2001/XMLSchema#boolean
film publication date publicationDate
film cast member headOfGovernmentStatement
You expect normal people to use SPARQL?! 

- If they want ... it's really not that difficult
  - Extensive online documentation
  - Over 300 example queries
  - Tutorials and workshops at community events

- But SPARQL is often hidden from users
  - Embedded results on Web pages (incl. Wikipedia)
    Mobile apps and online apps
  - Crowdsourcing platforms
Wikidata:Request a query

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?
Current Usage

- SPARQL is widely used
  - >100M requests per month (3.8M per day) in 2018
- 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
- All software/customisations free & open source
  - See https://github.com/wikimedia/wikidata-query-rdf
“What can we learn from all these SPARQL queries?”
SPARQL Queries Are Interesting

- Which data is actually asked for?
- Which SPARQL features are most important?
- Who is using SPARQL through which tools?

We have analysed complete Wikidata SPARQL query logs (Wikimedia Research Collaboration)

[Bielefeldt et al., “Linked Data on The Web” @ WWW 2018]
Analysing SPARQL logs: The Bot Problem
Analysing SPARQL logs: The Bot Problem

- Query traffic is **ruled** by a few bots

Fig.: Wikidata SPARQL traffic Jun-Sep 2017
Analysing SPARQL logs: The Bot Problem

- Query traffic is **ruled** by a few bots

Fig.: Wikidata SPARQL traffic Jun-Sep 2017

- 41% of all Wikidata query traffic from June – September 2017 caused by one super-power user (Magnus Manske)
Analysing SPARQL logs: The Bot Problem

- Query traffic is **ruled** by a few bots

Fig.: Wikidata SPARQL traffic Jun-Sep 2017

- 41% of all Wikidata query traffic from June – September 2017 caused by one super-power user (Magnus Manske)

- The effect does **not** average out, and it affects other sites too

Fig.: Usage of DISTINCT on DBpedia [Bonifati et al. 2017]
Analysing SPARQL logs: The Bot Problem

- Query traffic is **ruled** by a few bots

Fig.: Wikidata SPARQL traffic Jun-Sep 2017

- 41% of all Wikidata query traffic from June – September 2017 caused by one super-power user (Magnus Manske)

Fig.: Usage of DISTINCT on DBpedia [Bonifati et al. 2017]

---

No trends! No predictability! No insights!
Are SPARQL queries interesting after all?

- Observation: Robotic traffic dominates
  - May not represent any real interest
  - Governed by very few sources
  - Random changes – not uniform on any observed scale
Are SPARQL queries interesting after all?

- **Observation: Robotic traffic dominates**
  - May not represent any real interest
  - Governed by very few sources
  - Random changes – not uniform on any observed scale

- **Hypothesis: Organic traffic also exists**
  - Representing human information need during some interaction
  - Composed of many diverse sources
  - Continuous change over months

Note: “Organic” ≠ “hand-written SPARQL” (user apps might use SPARQL to get user-requested data without users actually writing queries)
Extracting organic traffic

- Main signal: User Agents
  - Assumption: organic traffic generally from browser-like agents
Extracting organic traffic

- **Main signal: User Agents**
  - Assumption: organic traffic generally from browser-like agents
- **2nd signal: query comments**
  - Some browser-based tools mark queries using comments
- **3rd signal: activity spikes**
  - Group queries by query pattern (following [Raghuveer, USEWOD’12])
  - Find agent-pattern pairs that spike (>2K requests/month)
  - Manually inspect these queries to decide if organic or robotic
  - About 300 further browser-based sources classified “robotic”
Results: Organic component

- Jun–Sep 2017: 658,890 queries (<0.5%)

More varied (vocabulary, SPARQL features)

More triples
organic 17%: 1, 97%: ≤11 vs. robotic 57%: 1, 96%: ≤7
Insights on Wikidata Usage

- Robotic traffic:
  - Mainly information integration bots (comparing database contents)
  - Potentially also selective data download (spider-like)
  - Most queries from a few dominant bots (>60% from top-three bots)

- Organic traffic:
  - Data browsers (often general-purpose)
  - Mobile apps (often topical)
  - Most queries from unidentified “small” sources

- Reified statements in 4%–10% of queries
What’s Next?
More data

- Wikidata: >45M items with >400M statements
More data

- Wikidata: >45M items with >400M statements
  - OSM: >4B nodes, >230M buildings, >10M trees
  - WDC: >9.5B entities, >38B RDF triples
- Why don’t we just import everything?!
More data

- Wikidata: >45M items with >400M statements
  - OSM: >4B nodes, >230M buildings, >10M trees
  - WDC: >9.5B entities, >38B RDF triples
- Why don’t we just import everything?!
  - Notability? Well, sometimes …
  - Community support! Who will maintain this?
More data: current efforts

- Data donation guidelines
- Wikidata aligns with >2500 databases and catalogues
- Supervised data alignment with crowdsourcing (Mix’n’Match)

<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>AcademiaNet</td>
<td>Database for excellent female scientists</td>
<td>99%</td>
</tr>
<tr>
<td>Austrian Parliament ID</td>
<td>Austrian Parliament’s “Who’s Who” database</td>
<td>99%</td>
</tr>
<tr>
<td>International World Games Ass</td>
<td>Sportspeople</td>
<td>24% 59%</td>
</tr>
<tr>
<td>botanist author abbreviation</td>
<td>standard form (official abbreviation) of a personal name for use in article</td>
<td>99%</td>
</tr>
<tr>
<td>Mactutor</td>
<td>identifier of the person’s biography in the MacTutor History of Mathematics</td>
<td>98%</td>
</tr>
<tr>
<td>South Australian Football Hall</td>
<td>Australian rules football players</td>
<td>72%</td>
</tr>
<tr>
<td>AIBL members</td>
<td>Membres de l’Académie des Inscriptions et Belles Lettres (AIBL)</td>
<td>70%</td>
</tr>
<tr>
<td>Lotsawa House Tibetan author</td>
<td>Tibetan authors in the Lotsawa House library</td>
<td>40%</td>
</tr>
<tr>
<td>parliament.uk</td>
<td>UK MP or Peer's biography</td>
<td>95%</td>
</tr>
<tr>
<td>EPHE</td>
<td>identifier of a researcher on the online prosopographical dictionary of</td>
<td>74% 16%</td>
</tr>
<tr>
<td>Sport Australia Hall of Fame</td>
<td>Sportspeople linked to Australia</td>
<td>73% 10%</td>
</tr>
<tr>
<td>North Carolina Sports Hall of F.</td>
<td>Sportspeople linked to North Carolina</td>
<td>54% 16%</td>
</tr>
</tbody>
</table>
Soccer donna

Soccer donna website female association football player db

Casey Short

' player born 23.08.1990 at Naperville, Illinois plays'

Markus Krötzsch

Load next entry on empty search results

Find

Casey Short [Q16766251]
US-American association football player [1990]?

Alaska – Die rau e Eiswei [Q925785]
1997 short film from United States of America
Q17525877 [Q17525877]
1912 silent short film from United States of America
Q17520601 [Q17520601]
2014 silent short film by James Young from United States of America
Q24258544 [Q24258544]
1913 short film by Edward Dillon from United States of America

wikidata

Casey Short (Q16766251)

American association football player

Statements

instance of

human

0 references
New kinds of data

- Coming soon: **lexical data** (dictionary/thesaurus)
  - Exciting & dangerous
- Planned: **media (meta-)data** (Wikimedia Commons)
- Factual knowledge that is not in catalogues?
- Common sense?

In many cases: technical changes/extensions needed
Quality!

- Errors, spam, vandalism
- Global coherency of modelling
- Sources & alignments
- Incompleteness
- Change & data rot
Germany

Federal parliamentary republic in central-western Europe

Potential issues

conflicts-with constraint
An entity should not have a statement for `basic form of government` if it also has a statement for `instance of` with value `republic`.

one-of constraint
The value for `basic form of government` should be one of the following:

- republic
- constitutional monarchy
- federal republic
- representative democracy
- parliamentary system
- soviet republic
A rule of inference:

(?headOfState.position held ^{P39} = ?headOffice)@X,
(?country.office held by head of state ^{P1906} = ?headOffice)@Y
→ (?country.head of state ^{P35} = ?headOfState)@{start time ^{P580} = ?X.start time ^{P580},
end time ^{P582} = ?X.end time ^{P582}}
Conclusion and Outlook

- Wikidata is a fascinating, fast-moving project
  - Large amounts of quality data & much more to come
  - Data export and analysis services for all needs
  - Innovation-friendly community
- Many unsolved questions for research
  - Quality, provenance, social aspects, performance challenges, data integration, internationalisation, ...
Literature


- Maximilian Marx, Markus Krötzsch, Veronika Thost: “Logic on MARS: Ontologies for generalised property graphs” Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI'17), 1188-1194, 2017
<table>
<thead>
<tr>
<th>Film Title</th>
<th>Year</th>
<th>Director</th>
<th>Year in Office</th>
<th>Future Heads of Government</th>
</tr>
</thead>
<tbody>
<tr>
<td>Star in the Dust</td>
<td>1956</td>
<td>Charles F. Haas</td>
<td>2</td>
<td>Clint Eastwood, mayor; George Wallace, Governor of Alabama</td>
</tr>
<tr>
<td>The Two Who Stole the Moon</td>
<td>1962</td>
<td>Jan Batory</td>
<td>2</td>
<td>Jarosław Kaczyński, Prime Minister of Poland; Lech Kaczyński, Mayor of Warsaw</td>
</tr>
<tr>
<td>Ragasiya Police 115</td>
<td>1968</td>
<td>B. R. Panthulu</td>
<td>2</td>
<td>M. G. Ramachandran, Chief Minister of Tamil Nadu; Jayalalithaa, Chief Minister of Tamil Nadu</td>
</tr>
<tr>
<td>Québec : Duplessis et après...</td>
<td>documentary</td>
<td></td>
<td>2</td>
<td>Bernard Landry, Premier of Quebec; René Lévesque, Premier of Quebec</td>
</tr>
<tr>
<td>Q3541438</td>
<td>1994</td>
<td>Claude Lanzmann</td>
<td>2</td>
<td>Ariel Sharon, Prime Minister of Israel; Ehud Barak, Prime Minister of Israel</td>
</tr>
<tr>
<td>Batman &amp; Robin</td>
<td>1997</td>
<td></td>
<td>2</td>
<td>Arnold Schwarzenegger, Mr. Freeze / Governor of California; Jesse Ventura, Governor of Minnesota</td>
</tr>
<tr>
<td>------------------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td><strong>Limit</strong></td>
<td>31.08%</td>
<td>39.55%</td>
<td>46.56%</td>
<td>52.31%</td>
</tr>
<tr>
<td><strong>Distinct</strong></td>
<td>26.50%</td>
<td>31.40%</td>
<td>19.05%</td>
<td>59.30%</td>
</tr>
<tr>
<td><strong>Order By</strong></td>
<td>17.29%</td>
<td>14.75%</td>
<td>13.22%</td>
<td>46.89%</td>
</tr>
<tr>
<td><strong>Offset</strong></td>
<td>0.40%</td>
<td>2.92%</td>
<td>0.37%</td>
<td>0.09%</td>
</tr>
<tr>
<td><strong>Join</strong></td>
<td>67.99%</td>
<td>87.82%</td>
<td>89.76%</td>
<td>82.50%</td>
</tr>
<tr>
<td><strong>Optional</strong></td>
<td>42.36%</td>
<td>46.24%</td>
<td>55.92%</td>
<td>50.90%</td>
</tr>
<tr>
<td><strong>Filter</strong></td>
<td>25.89%</td>
<td>29.12%</td>
<td>22.24%</td>
<td>12.59%</td>
</tr>
<tr>
<td><strong>Path with</strong></td>
<td>15.02%</td>
<td>15.59%</td>
<td>12.88%</td>
<td>40.92%</td>
</tr>
<tr>
<td><strong>Subquery</strong></td>
<td>13.09%</td>
<td>15.30%</td>
<td>12.79%</td>
<td>6.45%</td>
</tr>
<tr>
<td><strong>Bind</strong></td>
<td>9.85%</td>
<td>9.23%</td>
<td>8.68%</td>
<td>4.72%</td>
</tr>
<tr>
<td><strong>Union</strong></td>
<td>5.10%</td>
<td>5.76%</td>
<td>12.62%</td>
<td>2.56%</td>
</tr>
<tr>
<td><strong>Values</strong></td>
<td>4.44%</td>
<td>3.07%</td>
<td>10.88%</td>
<td>3.29%</td>
</tr>
<tr>
<td><strong>Not Exists</strong></td>
<td>3.31%</td>
<td>3.37%</td>
<td>2.46%</td>
<td>1.24%</td>
</tr>
<tr>
<td><strong>Minus</strong></td>
<td>2.04%</td>
<td>2.91%</td>
<td>1.60%</td>
<td>0.82%</td>
</tr>
<tr>
<td><strong>Service (lang)</strong></td>
<td>44.63%</td>
<td>42.09%</td>
<td>54.78%</td>
<td>40.88%</td>
</tr>
<tr>
<td><strong>Service (other)</strong></td>
<td>11.49%</td>
<td>10.53%</td>
<td>10.32%</td>
<td>7.30%</td>
</tr>
<tr>
<td><strong>Group By</strong></td>
<td>17.12%</td>
<td>19.93%</td>
<td>13.04%</td>
<td>7.00%</td>
</tr>
<tr>
<td><strong>Sample</strong></td>
<td>8.85%</td>
<td>10.93%</td>
<td>4.60%</td>
<td>1.61%</td>
</tr>
<tr>
<td><strong>Count</strong></td>
<td>7.55%</td>
<td>7.60%</td>
<td>8.15%</td>
<td>5.22%</td>
</tr>
<tr>
<td><strong>GroupConcat</strong></td>
<td>1.80%</td>
<td>2.79%</td>
<td>1.17%</td>
<td>0.86%</td>
</tr>
<tr>
<td><strong>Having</strong></td>
<td>1.17%</td>
<td>1.14%</td>
<td>0.72%</td>
<td>0.65%</td>
</tr>
</tbody>
</table>
Triples per query: organic (blue) / robotic (yellow)
Languages of labels in organic queries
### SPARQL feature co-occurrence

<table>
<thead>
<tr>
<th>J F O U P V S</th>
<th>organic</th>
<th>robotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>(none)</td>
<td>I1–I3</td>
<td>I4–I6</td>
</tr>
<tr>
<td></td>
<td>8.04</td>
<td>9.22</td>
</tr>
<tr>
<td></td>
<td>19.67</td>
<td>27.67</td>
</tr>
<tr>
<td>J</td>
<td>13.29</td>
<td>31.35</td>
</tr>
<tr>
<td></td>
<td>11.26</td>
<td>10.09</td>
</tr>
<tr>
<td>F</td>
<td>1.10</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>1.92</td>
<td>1.31</td>
</tr>
<tr>
<td>J F</td>
<td>6.68</td>
<td>2.39</td>
</tr>
<tr>
<td></td>
<td>2.61</td>
<td>1.68</td>
</tr>
<tr>
<td>J P</td>
<td>2.98</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td>13.50</td>
<td>13.94</td>
</tr>
<tr>
<td>J F P</td>
<td>2.48</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.39</td>
<td>0.07</td>
</tr>
<tr>
<td>J V</td>
<td>0.39</td>
<td>2.01</td>
</tr>
<tr>
<td></td>
<td>30.42</td>
<td>17.47</td>
</tr>
<tr>
<td>J O</td>
<td>1.26</td>
<td>1.64</td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>0.63</td>
</tr>
<tr>
<td>J O P</td>
<td>22.32</td>
<td>7.04</td>
</tr>
<tr>
<td></td>
<td>1.86</td>
<td>1.95</td>
</tr>
<tr>
<td>J O P</td>
<td>2.07</td>
<td>29.10</td>
</tr>
<tr>
<td></td>
<td>0.35</td>
<td>0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>J F O U P V S</th>
<th>organic</th>
<th>robotic</th>
</tr>
</thead>
<tbody>
<tr>
<td>J</td>
<td>2.66</td>
<td>1.32</td>
</tr>
<tr>
<td></td>
<td>2.13</td>
<td>1.18</td>
</tr>
<tr>
<td>J O U</td>
<td>3.49</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>J O V</td>
<td>3.38</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>0.11</td>
<td>0.43</td>
</tr>
<tr>
<td>J O P V</td>
<td>1.01</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>0.16</td>
<td>0.07</td>
</tr>
<tr>
<td>J S</td>
<td>2.76</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>0.06</td>
<td>0.01</td>
</tr>
<tr>
<td>J O S</td>
<td>4.78</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>J F S</td>
<td>3.19</td>
<td>2.28</td>
</tr>
<tr>
<td></td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>J F O S</td>
<td>1.02</td>
<td>0.13</td>
</tr>
<tr>
<td></td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J F O P</td>
<td>0.79</td>
<td>0.31</td>
</tr>
<tr>
<td></td>
<td>0.64</td>
<td>1.58</td>
</tr>
<tr>
<td>J U P V</td>
<td>0.01</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>0.05</td>
<td>1.92</td>
</tr>
</tbody>
</table>