Practical Linked Data Access via SPARQL: The Case of Wikidata

Adrian Bielefeldt       Julius Gonsior       Markus Krötzsch

Knowledge-Based Systems
TU Dresden

Also reporting on joint work with
Stas Malyshev (Wikimedia) and Larry Gonzalez (TU Dresden)

Research supported by the Wikimedia Foundation

For the eponymous LDOW 2018 paper, see https://iccl.inf.tu-dresden.de/web/Inproceedings3196/en

Slideset published under CC-By-SA 3.0 – slides without the title slide also published as CC-By 3.0
Wikidata, the knowledge graph of Wikipedia, uses SPARQL as its main query API.

- Who is using this?
- What are those SPARQL queries like?
- What can we learn from them?
Wait! — Wikidata uses RDF?!

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
Wait! — Wikidata uses RDF?!

Louis Néel (Q155781) award received (P166) Nobel Prize in Physics (Q38104)

point in time (P585): 1970
together with (P1706): H. Alfvén (Q54945)
prize money (P2121): 200000 SEK (Q122922)

How does Wikidata’s rich graph model relate to RDF?
Wait! — Wikidata uses RDF?!

Official RDF version follows Erxleben et al. [ISWC 2014]:

Louis Néel (Q155781) – wdt:P166 – Nobel Prize in Physics (Q38104)
Wait! — Wikidata uses RDF?!

Official RDF version follows Erxleben et al. [ISWC 2014]:

- Louis Néel (Q155781) is a Nobel Prize in Physics (Q38104) in 1970.
- The official RDF version follows Erxleben et al. [ISWC 2014].
RDF for Wikidata

- Wikidata offers **all of its content** in RDF
  - Linked data live exports
    (Example: https://www.wikidata.org/wiki/Special:EntityData/Q42.nt)
  - Weekly dumps
    (See https://dumps.wikimedia.org/wikidatawiki/entities/)
- Currently **4.9B triples** (as of April 2018)
  - >415M Wikidata Statements
  - 4.5K Wikidata properties → >48K RDF properties
  - >1.5B labels/descriptions/aliases
  - >63M links to Wikipedia and friends
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 wdt: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
PREFIX wsd: <http://wikiba.se/ontology#>
PREFIX wq: <http://wikiba.se/ontology#QuantityAmount>
PREFIX wd: <http://wikiba.se/ontology#QuantityUnit>
SELECT ?district ?districtNumber ?memberOfKabupaten
WHERE {
  ?district rdfs:label "Hengersberg-Tegelberg" .
  ?memberOfKabupaten rdfs:label "Kabupaten Jelutong" .
  ?memberOfKabupaten wq:hasQuantityAmount ?amount.
  ?amount wsd:hasUnit ?unit.
  ?amount wsd:hasNumber ?number.
  FILTER (?number = 100) .
}
```

A not-so-simple SPARQL query

```sparql
# Constituencies for the election to the German Bundestag 2017, with winning candidate and party
    # Find districts with shape
    ?district wdt:P3898 ?shape;
    # Successful candidate for 19th German Bundestag with party and % votes
    # District number in 2017 Bundestag constituencies
    # Turn string district number into integer
    BIND(?wdt:Q329) AS ?districtNumber;
    # Sanity check
    BIND(true AS ?sanityCheckMbo)
    # Find original color of party
    ?party wd:P4645 "P465" "rgbOriginal",
    # Fade color depending on % votes, knowing that the original colors are only composed of FF, BB, 00; shift BB to AB or C6, and 00 to 80 or 88; using separate calls to replace R, G, and B components so that the replacements are aligned to them
} SERVICE wikibase:label {
    bd:serviceParam wikibase:language "en",
} ORDER BY ?districtNumber
```
Some metrics

- Running on BlazeGraph database engine
  - 3 servers (+3 as backup) Intel Xeon E5-2620 8 core/128G mem/800G SSD
  - Standard caching (Varnish) and load balancing (LVS)
  - Some custom tools, extension and tunings

All available online: https://github.com/wikimedia/wikidata-query-rdf
Some metrics

- **Running on BlazeGraph database engine**
  - 3 servers (+3 as backup) Intel Xeon E5-2620 8 core/128G mem/800G SSD
  - Standard caching (Varnish) and load balancing (LVS)
  - Some custom tools, extension and tunings
    All available online: https://github.com/wikimedia/wikidata-query-rdf

- **Serving >100M requests/month (3.8M/day)**
  - 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
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)

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

  Fig.: Usage of DISTINCT on DBpedia

  [Bonifati et al. 2017]
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 triples
  organic 17%: 1, 97%: ≤11 vs. robotic 57%: 1, 96%: ≤7
- More varied (vocabulary, SPARQL features)

Temporal distribution of organic queries (12 weeks / time of day)
Insights on SPARQL Usage

- General: more features than reported elsewhere
- Typically organic: LIMIT, DISTINCT, OPTIONAL, ORDER BY, subqueries, aggregates, services
- Typically robotic: BIND, UNION, VALUES
- Conjunctive regular path queries with converse (C2RPQs)
  - Main query fragment for robotic queries (75% when allowing VALUES)
- OPTIONAL:
  - Important mostly for organic queries
  - Recent data (2018) also shows shift to C2RPQ+OPTIONAL (up to 82%)
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**
Conclusion and Outlook

Wikidata relies on RDF and SPARQL for some of its core features – a fascinating use case!

- **Conclusions**
  - SPARQL log analysis is **methodologically difficult**
  - **Organic traffic** can be extracted based on User Agent and timestamps
  - SPARQL queries are **more varied and more complex** than reported elsewhere
  - After Joins, **path queries** are the second most important feature

- **Outlook**
  - Publishing anonymised datasets: under review; stay tuned
  - Documenting Wikidata’s SPARQL deployment insights
  -Wikidata will expand further … (Dictionary content! Media meta-data!)
## SPARQL Feature Distribution (2017/2018)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Limit</td>
<td>31.08%</td>
<td>39.55%</td>
<td>46.56%</td>
<td>52.31%</td>
<td>51.23%</td>
<td>36.87%</td>
<td></td>
<td>21.12%</td>
<td>16.86%</td>
<td>17.42%</td>
<td>20.38%</td>
<td>11.47%</td>
<td>15.17%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distinct</td>
<td>26.50%</td>
<td>31.40%</td>
<td>19.05%</td>
<td>59.30%</td>
<td>60.42%</td>
<td>63.78%</td>
<td></td>
<td>15.84%</td>
<td>5.48%</td>
<td>4.27%</td>
<td>4.32%</td>
<td>7.54%</td>
<td>12.25%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order By</td>
<td>17.29%</td>
<td>14.75%</td>
<td>13.22%</td>
<td>46.89%</td>
<td>46.99%</td>
<td>34.53%</td>
<td></td>
<td>12.97%</td>
<td>8.01%</td>
<td>6.78%</td>
<td>8.76%</td>
<td>7.68%</td>
<td>17.46%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offset</td>
<td>0.40%</td>
<td>2.92%</td>
<td>0.37%</td>
<td>0.09%</td>
<td>0.08%</td>
<td>0.06%</td>
<td></td>
<td>7.73%</td>
<td>6.07%</td>
<td>6.29%</td>
<td>0.10%</td>
<td>0.07%</td>
<td>0.10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Join</td>
<td>67.99%</td>
<td>87.82%</td>
<td>89.76%</td>
<td>82.50%</td>
<td>91.70%</td>
<td>87.02%</td>
<td></td>
<td>88.46%</td>
<td>78.53%</td>
<td>67.41%</td>
<td>73.26%</td>
<td>61.39%</td>
<td>70.19%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Optional</td>
<td>42.36%</td>
<td>46.24%</td>
<td>55.92%</td>
<td>50.90%</td>
<td>41.30%</td>
<td>41.15%</td>
<td></td>
<td>25.08%</td>
<td>11.63%</td>
<td>11.45%</td>
<td>12.73%</td>
<td>15.41%</td>
<td>30.71%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Filter</td>
<td>25.89%</td>
<td>29.12%</td>
<td>22.24%</td>
<td>12.56%</td>
<td>11.76%</td>
<td>11.76%</td>
<td></td>
<td>21.64%</td>
<td>17.92%</td>
<td>13.79%</td>
<td>14.70%</td>
<td>16.83%</td>
<td>29.02%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Path with *</td>
<td>15.02%</td>
<td>15.59%</td>
<td>12.88%</td>
<td>40.92%</td>
<td>32.43%</td>
<td>30.34%</td>
<td></td>
<td>16.43%</td>
<td>19.19%</td>
<td>14.80%</td>
<td>20.56%</td>
<td>17.26%</td>
<td>24.81%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subquery *</td>
<td>13.09%</td>
<td>15.30%</td>
<td>12.79%</td>
<td>6.45%</td>
<td>5.07%</td>
<td>5.39%</td>
<td></td>
<td>0.34%</td>
<td>0.28%</td>
<td>0.33%</td>
<td>0.09%</td>
<td>0.13%</td>
<td>0.11%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bind</td>
<td>9.85%</td>
<td>9.23%</td>
<td>8.68%</td>
<td>4.72%</td>
<td>3.99%</td>
<td>4.15%</td>
<td></td>
<td>16.29%</td>
<td>12.07%</td>
<td>9.60%</td>
<td>11.94%</td>
<td>13.79%</td>
<td>24.03%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Union</td>
<td>5.10%</td>
<td>5.76%</td>
<td>12.62%</td>
<td>2.56%</td>
<td>2.07%</td>
<td>3.39%</td>
<td></td>
<td>11.26%</td>
<td>8.63%</td>
<td>7.61%</td>
<td>13.96%</td>
<td>13.05%</td>
<td>18.57%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Values</td>
<td>4.44%</td>
<td>3.07%</td>
<td>10.88%</td>
<td>3.29%</td>
<td>3.23%</td>
<td>3.20%</td>
<td></td>
<td>35.72%</td>
<td>30.74%</td>
<td>28.92%</td>
<td>29.82%</td>
<td>23.80%</td>
<td>11.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Exists</td>
<td>3.31%</td>
<td>3.37%</td>
<td>2.46%</td>
<td>1.24%</td>
<td>0.94%</td>
<td>0.69%</td>
<td></td>
<td>0.19%</td>
<td>0.21%</td>
<td>0.19%</td>
<td>0.27%</td>
<td>0.29%</td>
<td>0.35%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Minus</td>
<td>2.04%</td>
<td>2.91%</td>
<td>1.60%</td>
<td>0.82%</td>
<td>0.57%</td>
<td>0.71%</td>
<td></td>
<td>0.53%</td>
<td>0.92%</td>
<td>1.07%</td>
<td>1.46%</td>
<td>1.26%</td>
<td>1.78%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service (lang)</td>
<td>44.63%</td>
<td>42.09%</td>
<td>54.78%</td>
<td>50.88%</td>
<td>41.71%</td>
<td>42.95%</td>
<td></td>
<td>10.40%</td>
<td>6.15%</td>
<td>4.27%</td>
<td>7.15%</td>
<td>7.91%</td>
<td>8.90%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Service (other)</td>
<td>11.49%</td>
<td>10.53%</td>
<td>10.32%</td>
<td>7.30%</td>
<td>13.14%</td>
<td>2.31%</td>
<td></td>
<td>4.51%</td>
<td>0.19%</td>
<td>0.16%</td>
<td>0.17%</td>
<td>0.18%</td>
<td>0.51%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group By</td>
<td>17.12%</td>
<td>19.93%</td>
<td>13.04%</td>
<td>7.00%</td>
<td>5.40%</td>
<td>5.07%</td>
<td></td>
<td>0.41%</td>
<td>0.37%</td>
<td>0.48%</td>
<td>0.22%</td>
<td>0.23%</td>
<td>0.39%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sample</td>
<td>8.85%</td>
<td>10.93%</td>
<td>4.60%</td>
<td>1.61%</td>
<td>1.63%</td>
<td>1.69%</td>
<td></td>
<td>0.04%</td>
<td>0.04%</td>
<td>0.06%</td>
<td>0.05%</td>
<td>0.04%</td>
<td>0.10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Count</td>
<td>7.55%</td>
<td>7.60%</td>
<td>8.15%</td>
<td>5.22%</td>
<td>3.88%</td>
<td>3.73%</td>
<td></td>
<td>1.15%</td>
<td>4.30%</td>
<td>3.00%</td>
<td>1.52%</td>
<td>0.65%</td>
<td>0.89%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GroupConcat</td>
<td>1.80%</td>
<td>2.79%</td>
<td>1.17%</td>
<td>0.86%</td>
<td>0.86%</td>
<td>0.74%</td>
<td></td>
<td>0.06%</td>
<td>0.09%</td>
<td>0.02%</td>
<td>0.03%</td>
<td>0.02%</td>
<td>0.28%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Having</td>
<td>1.17%</td>
<td>1.14%</td>
<td>0.72%</td>
<td>0.65%</td>
<td>0.26%</td>
<td>0.33%</td>
<td></td>
<td>0.01%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.04%</td>
<td></td>
<td></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></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(none)</td>
<td>8.04</td>
<td>9.22</td>
<td>19.67</td>
<td>27.67</td>
</tr>
<tr>
<td>J</td>
<td>13.29</td>
<td>31.35</td>
<td>11.26</td>
<td>10.09</td>
</tr>
<tr>
<td>F</td>
<td>1.10</td>
<td>0.98</td>
<td>1.92</td>
<td>1.31</td>
</tr>
<tr>
<td>J F</td>
<td>6.68</td>
<td>2.39</td>
<td>2.61</td>
<td>1.68</td>
</tr>
<tr>
<td>J P</td>
<td>2.98</td>
<td>1.62</td>
<td>13.50</td>
<td>13.94</td>
</tr>
<tr>
<td>J F P</td>
<td>2.48</td>
<td>0.58</td>
<td>0.39</td>
<td>0.07</td>
</tr>
<tr>
<td>J V</td>
<td>0.39</td>
<td>2.01</td>
<td>30.42</td>
<td>17.47</td>
</tr>
<tr>
<td>O</td>
<td>1.26</td>
<td>1.64</td>
<td>0.11</td>
<td>0.63</td>
</tr>
<tr>
<td>J O</td>
<td>22.32</td>
<td>7.04</td>
<td>1.86</td>
<td>1.95</td>
</tr>
<tr>
<td>J O P</td>
<td>2.07</td>
<td>29.10</td>
<td>0.35</td>
<td>0.05</td>
</tr>
</tbody>
</table>

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