Creating a RDF Knowledgebase from OpenStreetMap Data

Axel Lehmann
Master’s Thesis

Albert Ludwig University of Freiburg
Faculty of Engineering
Department of Computer Science
Chair of Algorithms and Data Structures
• Query: All places of worship with gothic architecture in the Altstadt of Freiburg.

• Combine factual data with spatial constraints.
  • factual data provided by Wikidata.
  • spatial constraints provided by OpenStreetMap.
Figure 1: Point coordinates of Altstadt Freiburg (a) and Freiburg Minster (b) provided by Wikidata. © Wikidata: All structured data from the main, Property, Lexeme, and EntitySchema namespaces is available under the Creative Commons CC0 License; text in the other namespaces is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply.
Figure 2: Places of worship inside Altstadt Freiburg provided by Overpass. © Overpass API; Base map and data from OpenStreetMap and OpenStreetMap Foundation
Figure 3: Expected result for: All places of worship with gothic architecture in the Altstadt of Freiburg — The Freiburg Minster.
Problem
Problem — Situation

- OpenStreetMap data not available as RDF
  - Sophox [Wik21] currently defunct
- Jovanovik, Homburg, and Spasic [JHS21] found GeoSPARQL support in SPARQL engines to be lacking

⇒ Convert OpenStreetMap data into RDF Triples with explicit spatial relations
• Volunteers mapping the whole world.
• Generic information as key-value string pairs
• Spatial Entities
  \( \approx 6.5 \text{ B Nodes (Longitude, Latitude)} \)
  \( \approx 0.7 \text{ B Ways (List of Nodes)} \)
  \( \approx 8.3 \text{ M Relations (Lost of Nodes and ways with type information)} \)
• Virtual Entities (libosmium)
  \( \approx 0.5 \text{ B Areas (Regions which can contain other elements, not explicitly stored, derived from ways and relations)} \)
• Spatial relations (contains, intersects, etc.) not explicitly stored
• Describes knowledge as a graph
• Subject-Predicate-Object Triples
  • Predicates are edges in the knowledge graph
• N-Triples and Turtle dialects store all data as annotated strings
Questions?
APPROACH
1. Convert key-value data into *RDF Triples*
   - Free form data → no structural guarantees
   - Multilingual (e.g. japanese) → UTF-8 support required

2. Conversion of spatial features into *well-known text*
   - Well-known text imposes stricter order requirements for elements

3. Calculating spatial relations and storing them as Triples
   - contains, intersects
   - Only tagged entities
Spatial location of Building 51 and Building 106
Approach — Directed Acyclic Graph — I

- RVF Zone A
- Freiburg im Breisgau
- 79110
- Brühl
- Brühl-Industriegebiet
- Albert-Ludwigs-Universität Freiburg
- Technische Fakultät
- Gebäude 51
- Gebäude 106
Benefits:

- Reduction of spatial comparisons
- Reduction of spatial triples (transitivity)

Problem:

- Where to look in the DAG?
  R-tree as index structure → candidate ids based on envelope
Candidates from R-tree → check intersects and contains (multithreaded using OpenMP)
Questions?
Experiments
# Experiments — Overview

<table>
<thead>
<tr>
<th></th>
<th>freiburg</th>
<th>bawue</th>
<th>germany</th>
<th>europe</th>
<th>planet*</th>
</tr>
</thead>
<tbody>
<tr>
<td>input (.pbf)</td>
<td>123 M</td>
<td>482 M</td>
<td>3.4 G</td>
<td>23 G</td>
<td>55 G°</td>
</tr>
<tr>
<td>output (.bz2)</td>
<td>257 M</td>
<td>1.1 G</td>
<td>7.5 G</td>
<td>48 G</td>
<td>100 G*</td>
</tr>
<tr>
<td>ram usage</td>
<td>2.47 G</td>
<td>2.82 G</td>
<td>13.51 G</td>
<td>90.11 G</td>
<td>108 G°</td>
</tr>
<tr>
<td>runtime facts</td>
<td>2.50 m</td>
<td>10.07 m</td>
<td>71.51 m</td>
<td>7.56 h</td>
<td>16.68 h</td>
</tr>
<tr>
<td>runtime geometry</td>
<td>50.13 s</td>
<td>44.96 m</td>
<td>15.30 h</td>
<td>12.28 d</td>
<td>47.02 d</td>
</tr>
<tr>
<td>runtime</td>
<td>3.34 m</td>
<td>55.02 m</td>
<td>16.49 h</td>
<td>12.59 d</td>
<td>47.72 d</td>
</tr>
<tr>
<td>fact triples</td>
<td>22.05 M</td>
<td>89.51 M</td>
<td>598.73 M</td>
<td>3.38 B</td>
<td>4.67 B</td>
</tr>
<tr>
<td>geometry triples</td>
<td>11.83 M</td>
<td>47.99 M</td>
<td>342.09 M</td>
<td>2.28 B</td>
<td>4.88 B</td>
</tr>
</tbody>
</table>

*Preliminary code used.

○ Input contains metadata.

* 12 decimal places, no additional metadata.

○ Node locations (during dump) stored on disk and not in RAM.

○ Added relations in both directions.
## Experiments — Overview

<table>
<thead>
<tr>
<th></th>
<th>freiburg</th>
<th>bawue</th>
<th>germany</th>
<th>europe</th>
<th>planet*</th>
</tr>
</thead>
<tbody>
<tr>
<td>nodes (src)</td>
<td>12.42 M</td>
<td>46.71 M</td>
<td>337.22 M</td>
<td>2.72 B</td>
<td>6.50 B</td>
</tr>
<tr>
<td>nodes (fact)</td>
<td>605.65 K</td>
<td>2.03 M</td>
<td>14.05 M</td>
<td>96.76 M</td>
<td>164.42 M</td>
</tr>
<tr>
<td>nodes (geom)</td>
<td>605.65 K</td>
<td>2.03 M</td>
<td>14.05 M</td>
<td>96.76 M</td>
<td>164.42 M</td>
</tr>
<tr>
<td>ways (s)</td>
<td>1.82 M</td>
<td>7.72 M</td>
<td>55.03 M</td>
<td>326.43 M</td>
<td>718.51 M</td>
</tr>
<tr>
<td>ways (f)</td>
<td>1.80 M</td>
<td>7.65 M</td>
<td>54.61 M</td>
<td>319.97 M</td>
<td>706.17 M</td>
</tr>
<tr>
<td>ways (g)</td>
<td>1.80 M</td>
<td>7.65 M</td>
<td>54.61 M</td>
<td>319.97 M</td>
<td>706.17 M</td>
</tr>
<tr>
<td>relations (s)</td>
<td>31.60 K</td>
<td>101.67 K</td>
<td>677.84 K</td>
<td>5.51 M</td>
<td>8.38 M</td>
</tr>
<tr>
<td>relations (f)</td>
<td>31.58 K</td>
<td>101.63 K</td>
<td>677.56 K</td>
<td>5.51 M</td>
<td>8.37 M</td>
</tr>
<tr>
<td>areas (s)</td>
<td>1.24 M</td>
<td>5.43 M</td>
<td>39.62 M</td>
<td>234.80 M</td>
<td>499.37 M</td>
</tr>
<tr>
<td>areas (f)</td>
<td>1.24 M</td>
<td>5.43 M</td>
<td>39.62 M</td>
<td>234.80 M</td>
<td>499.37 M</td>
</tr>
<tr>
<td>areas (g)</td>
<td>1.24 M</td>
<td>5.43 M</td>
<td>39.62 M</td>
<td>234.80 M</td>
<td>499.37 M</td>
</tr>
</tbody>
</table>

*Preliminary code used.
<table>
<thead>
<tr>
<th></th>
<th>freiburg</th>
<th>bawue</th>
<th>germany</th>
<th>europe</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vertices</td>
<td>32.79 K</td>
<td>146.46 K</td>
<td>991.31 K</td>
<td>5.54 M</td>
</tr>
<tr>
<td>edges</td>
<td>40.19 K</td>
<td>181.20 K</td>
<td>1.23 M</td>
<td>7.69 M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>∩ skipped (DAG)</td>
<td>60.94 K</td>
<td>272.69 K</td>
<td>1.61 M</td>
<td>14.95 M</td>
</tr>
<tr>
<td>∩</td>
<td>179.71 K</td>
<td>678.05 K</td>
<td>4.44 M</td>
<td>32.66 M</td>
</tr>
<tr>
<td>∩ yes</td>
<td>56.37 K</td>
<td>213.19 K</td>
<td>1.48 M</td>
<td>7.61 M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⊆ skipped (DAG)</td>
<td>46.81 K</td>
<td>219.58 K</td>
<td>1.24 M</td>
<td>12.85 M</td>
</tr>
<tr>
<td>⊆ envelope</td>
<td>70.51 K</td>
<td>266.31 K</td>
<td>1.85 M</td>
<td>9.72 M</td>
</tr>
<tr>
<td>⊆</td>
<td>19.84 K</td>
<td>59.89 K</td>
<td>398.03 K</td>
<td>4.76 M</td>
</tr>
<tr>
<td>⊆ yes</td>
<td>13.86 K</td>
<td>37.92 K</td>
<td>253.33 K</td>
<td>3.63 M</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>⊆ skipped (DAG)</td>
<td>2.67 M</td>
<td>15.34 M</td>
<td>96.58 M</td>
<td>658.15 M</td>
</tr>
<tr>
<td>⊆</td>
<td>3.54 M</td>
<td>17.68 M</td>
<td>118.04 M</td>
<td>903.54 M</td>
</tr>
<tr>
<td>⊆ yes</td>
<td>740.68 K</td>
<td>2.53 M</td>
<td>17.62 M</td>
<td>134.71 M</td>
</tr>
</tbody>
</table>
### Experiments — Spatial relation calculations

<table>
<thead>
<tr>
<th></th>
<th>freiburg</th>
<th>bawue</th>
<th>germany</th>
<th>europe</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>DAG</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>vertices</td>
<td>32.79 K</td>
<td>146.46 K</td>
<td>991.31 K</td>
<td>5.54 M</td>
</tr>
<tr>
<td>edges</td>
<td>40.19 K</td>
<td>181.20 K</td>
<td>1.23 M</td>
<td>7.69 M</td>
</tr>
<tr>
<td><strong>way</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>in DAG</td>
<td>31.12 K</td>
<td>140.18 K</td>
<td>935.95 K</td>
<td>4.93 M</td>
</tr>
<tr>
<td>(\cap) skipped (member)</td>
<td>109.72 K</td>
<td>535.81 K</td>
<td>4.32 M</td>
<td>23.28 M</td>
</tr>
<tr>
<td>(\cap) skipped (DAG)</td>
<td>7.37 M</td>
<td>55.66 M</td>
<td>376.04 M</td>
<td>2.21 B</td>
</tr>
<tr>
<td>(\subseteq) yes</td>
<td>2.06 M</td>
<td>8.72 M</td>
<td>62.31 M</td>
<td>404.87 M</td>
</tr>
<tr>
<td>(\subseteq) skipped (DAG)</td>
<td>7.19 M</td>
<td>54.80 M</td>
<td>370.28 M</td>
<td>2.18 B</td>
</tr>
<tr>
<td>(\subseteq) envelope</td>
<td>2.35 M</td>
<td>10.11 M</td>
<td>72.40 M</td>
<td>461.92 M</td>
</tr>
<tr>
<td>(\subseteq)</td>
<td>2.27 M</td>
<td>9.77 M</td>
<td>69.83 M</td>
<td>446.00 M</td>
</tr>
<tr>
<td>(\subseteq) yes</td>
<td>2.11 M</td>
<td>9.04 M</td>
<td>64.95 M</td>
<td>418.67 M</td>
</tr>
</tbody>
</table>
Conclusion
Conclusion — Future work

- Runtime
  - Spatial comparisons
    - better approximations (inner + outer shapes)
    - better index structure (R-tree axis aligned)
    - split areas with low area/envelope ratio
  - RDF / WKT

- OpenStreetMap relations
  - e.g. public transportation (Bus/Rail/Ferry)

- Spatial relations
  - touches, disjoint, ...
  - cardinal direction, distance ...
• We provide OpenStreetMap data as RDF
• We enable non GeoSPARQL capable SPARQL engines to answer some spatial queries
  • intersects
  • contains
• DAG reduces number of explicit spatial relation triples
• Source code (GPLv3+): https://github.com/ad-freiburg/osm2ttl
• Temporary demo: https://qlever.cs.uni-freiburg.de/OSM_Test
Questions?

WKT — Syntax

• Node (Single Point/Location)
  POINT(x, y)

• Way
  LINESTRING(x_0 y_0, x_1 y_1, ...)

• Area (solid, single part)
  POLYGON((x_0 y_0, x_1 y_1, ...))

• Area (single part with holes)
  POLYGON((x_0 y_0, x_1 y_1, ...), (...))

• Area (multiple parts with optional holes)
  MULTIPOLYGON(((x_0 y_0, x_1 y_1, ...), (...)), (...))
• OpenStreetMap represents geometries explizit
  • Closed ways \((n_0, n_1, \ldots n_k) | n_0 = n_k\)
  • Order and number of inner and outer not relevant

• WKT uses different representations
  • Closed ways as POLYGON (no repeat of first node)
  • Order of inner and outer is relevant
  • MULTIPOLYGON required for multiple outer parts
WHY EXPLICIT SPATIAL CALCULATIONS
The highlighted way crosses multiple administrative levels and intersects Baden-Württemberg and Hesse. It is contained in neither of them but in Germany.
• Gebäude 51
  https://www.openstreetmap.org/way/98284318
• Gebäude 106
  https://www.openstreetmap.org/way/33903567
• Technische Fakultät
  https://www.openstreetmap.org/way/4498466
• Albert-Ludwigs-Universität Freiburg
  https://www.openstreetmap.org/relation/1590189
• Brühl-Industriegebiet
  https://www.openstreetmap.org/relation/294855
• Brühl
  https://www.openstreetmap.org/relation/1956119
• 79110
  https://www.openstreetmap.org/relation/1112757

• Freiburg im Breisgau
  https://www.openstreetmap.org/relation/62768

• RVF Zone A
  https://www.openstreetmap.org/relation/4221993