Creating a RDF Knowledgebase from OpenStreetMap Data

Axel Lehmann

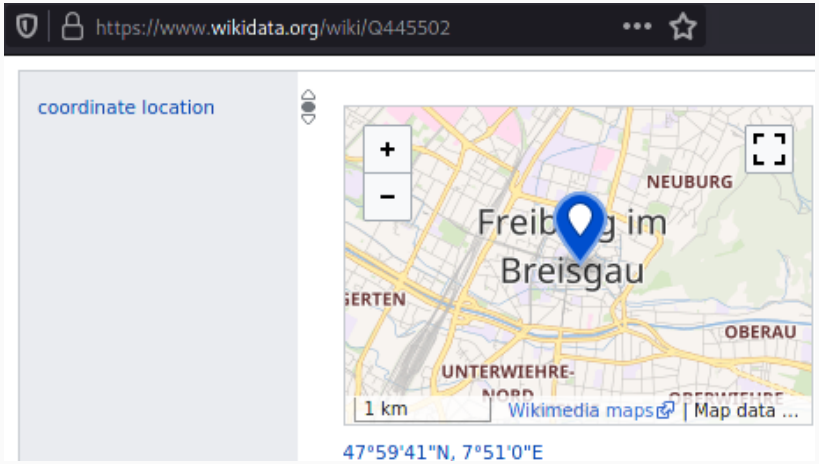
Master's Thesis

Albert Ludwig University of Freiburg
Faculty of Engineering
Department of Computer Science
Chair of Algorithms and Data Structures

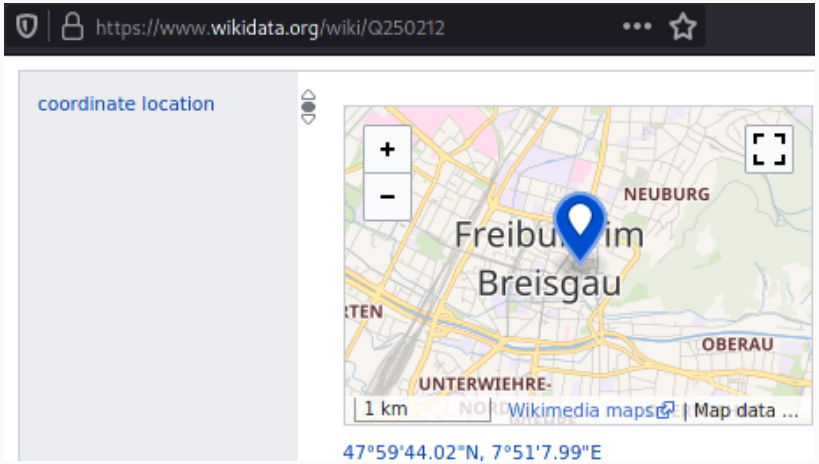
Motivation — Example

- 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*.

Motivation — Wikidata



(a) Altstadt Freiburg coordinates



(b) Freiburg Minster coordinates

Figure 1: Point coordinates of *Altstadt Freiburg* (a) and *Freiburg Minster* (b) provided by *Wikidata*.

Motivation — OpenStreetMap

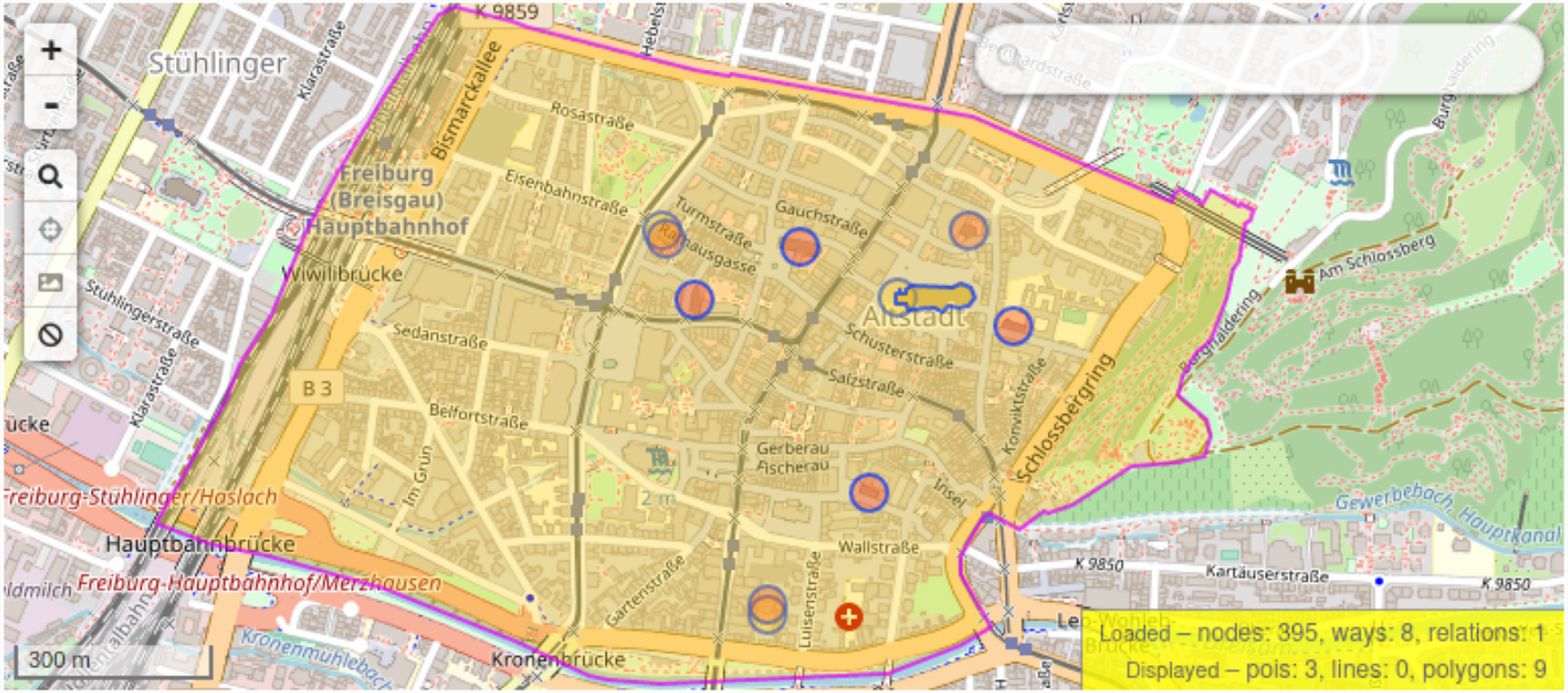


Figure 2: *Places of worship* inside *Altstadt Freiburg* provided by *Overpass*. © Overpass API; Base map and data from OpenStreetMap and OpenStreetMap Foundation

Motivation — Expected Result

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

$\Rightarrow$ Convert OpenStreetMap data into RDF Triples with explicit spatial relations

Problem — OpenStreetMap

- Volunteers mapping the whole world.
- Generic information as key-value string pairs
- Spatial Entities
  - $\approx$ 6.5 B Nodes (Longitude, Latitude)
  - $\approx$ 0.7 B Ways (List of Nodes)
  - $\approx$ 8.3 M Relations (Lost of Nodes and ways with type information)
- Virtual Entities (libosmium)
  - $\approx$ 0.5 B Areas (Regions which can contain other elements, not explicitly stored, derived from ways and relations)
- Spatial relations (contains, intersects, etc.) not explicitly stored

Problem — Resource Description Framework

- 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?

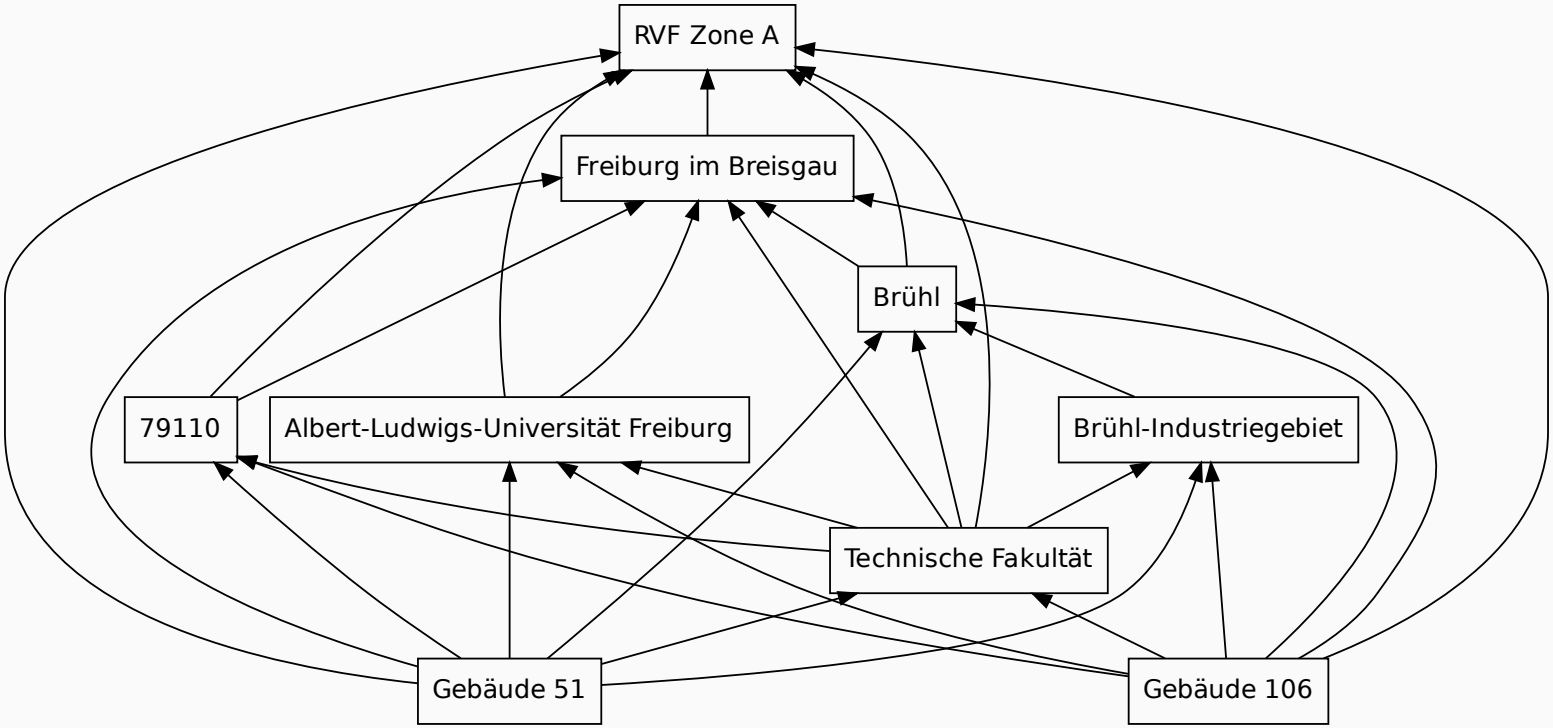
Approach

Approach — Converting into RDF

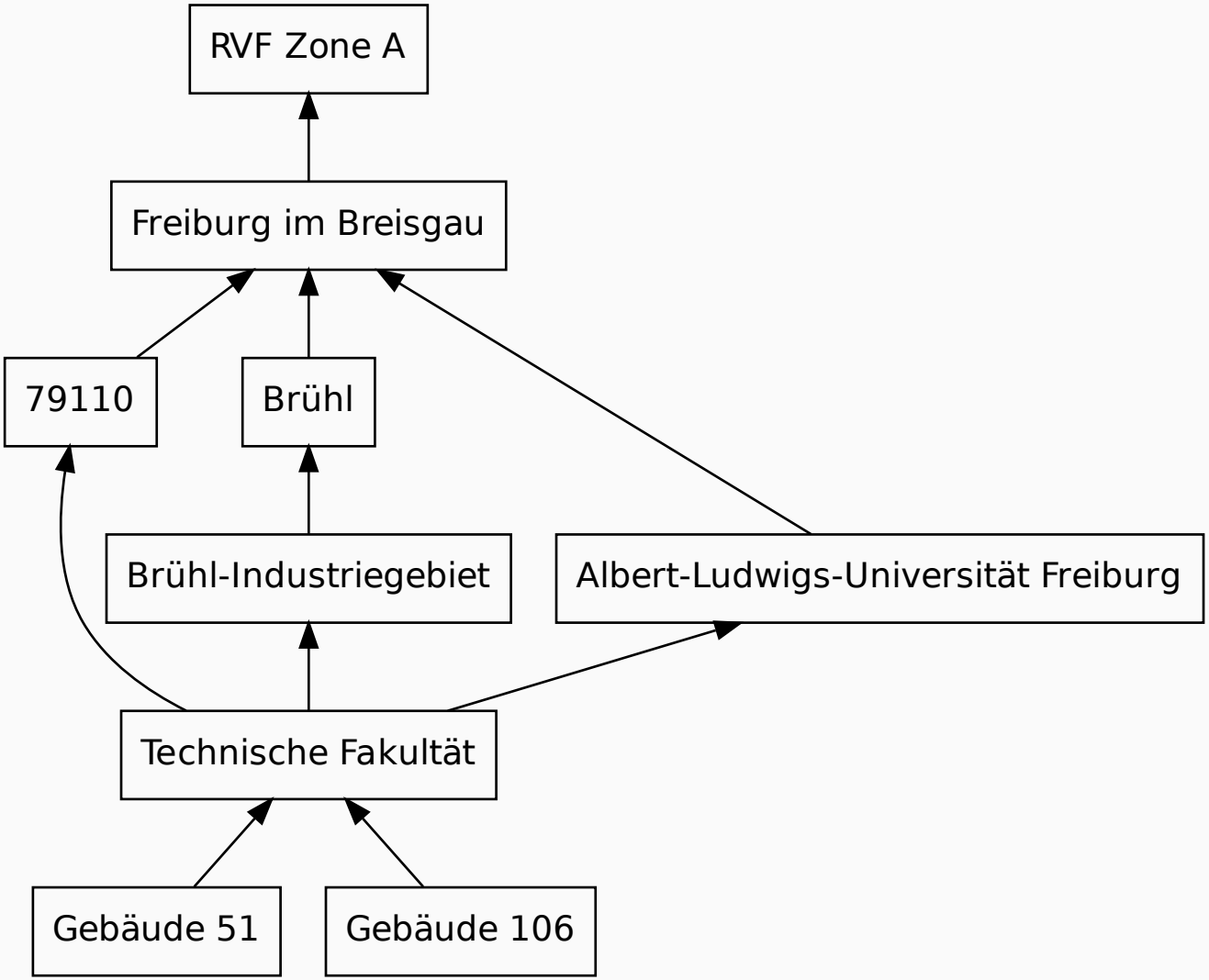
1. Convert key-value data into *RDF Triples*
   - Free form data → no structural guarantees
   - Multilingual (e.g. japanese) → UTF-8 support required
2. Conversion of spatial features into *well-known text*
   - Well-known text imposes stricter order requirements for elements
3. Calculating spatial relations and storing them as Triples
   - contains, intersects
   - Only tagged entities

Approach — Spatial area relations example

Spatial location of *Building 51* and *Building 106*



Approach — Directed Acyclic Graph — I



Approach — Directed Acyclic Graph — II

Benefits:

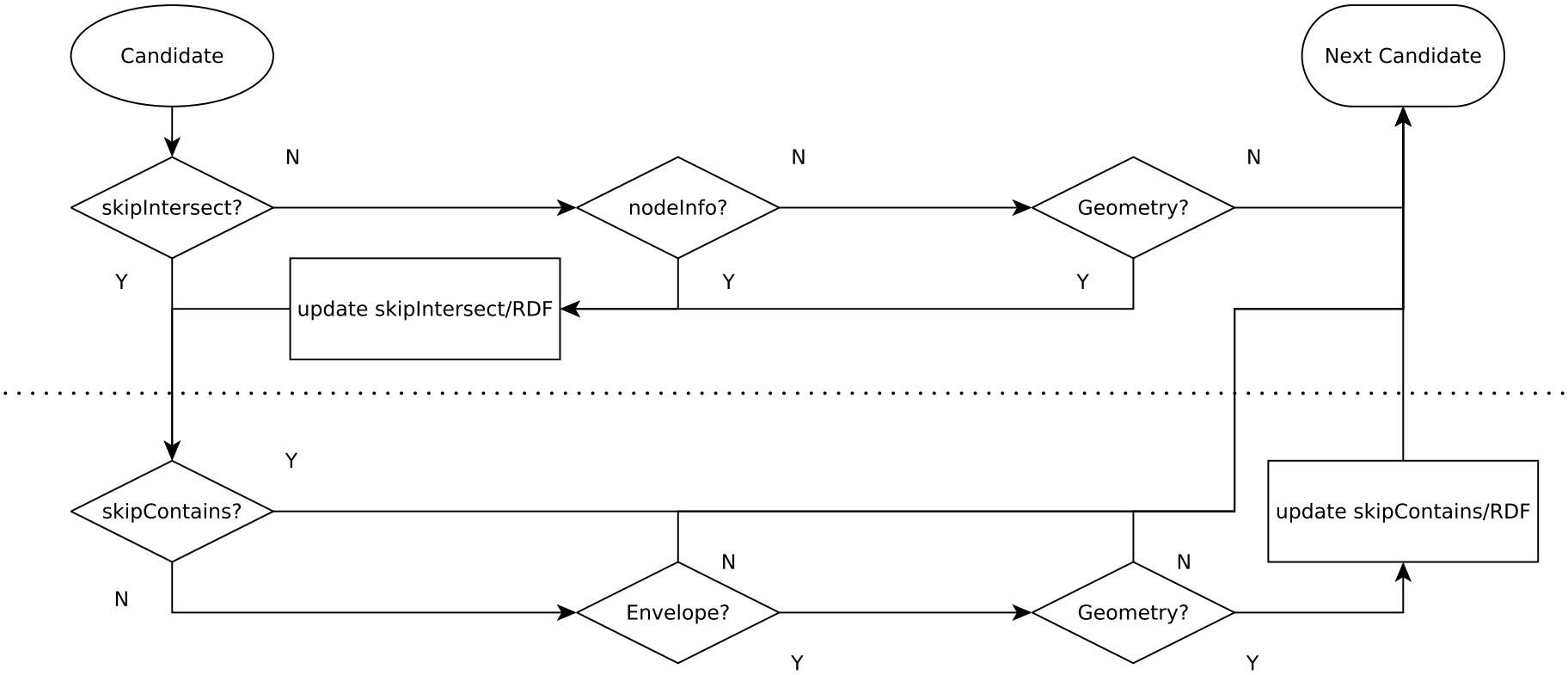
- Reduction of spatial comparisons
- Reduction of spatial triples (transitivity)

Problem:

- Where to look in the DAG?
  R-tree as index structure $\rightarrow$ candidate ids based on envelope

Approach — Spatial relation calculation — Ways

Candidates from R-tree → check intersects and contains (multihreaded using OpenMP)



Questions?

Experiments

Experiments — Overview

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

*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

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

*Preliminary code used.

Experiments — Spatial relation calculations

		freiburg	bawue	germany	europe
DAG	vertices	32.79 K	146.46 K	991.31 K	5.54 M
	edges	40.19 K	181.20 K	1.23 M	7.69 M
unnamed area	∩ skipped (DAG)	60.94 K	272.69 K	1.61 M	14.95 M
	∩	179.71 K	678.05 K	4.44 M	32.66 M
	∩ yes	56.37 K	213.19 K	1.48 M	7.61 M
	⊆ skipped (DAG)	46.81 K	219.58 K	1.24 M	12.85 M
	⊆ envelope	70.51 K	266.31 K	1.85 M	9.72 M
	⊆	19.84 K	59.89 K	398.03 K	4.76 M
	⊆ yes	13.86 K	37.92 K	253.33 K	3.63 M
node	⊆ skipped (DAG)	2.67 M	15.34 M	96.58 M	658.15 M
	⊆	3.54 M	17.68 M	118.04 M	903.54 M
	⊆ yes	740.68 K	2.53 M	17.62 M	134.71 M

Experiments — Spatial relation calculations

		freiburg	bawue	germany	europe
DAG	vertices	32.79 K	146.46 K	991.31 K	5.54 M
	edges	40.19 K	181.20 K	1.23 M	7.69 M
way	in DAG	31.12 K	140.18 K	935.95 K	4.93 M
	∩ skipped (member)	109.72 K	535.81 K	4.32 M	23.28 M
	∩ skipped (DAG)	7.37 M	55.66 M	376.04 M	2.21 B
	∩	10.37 M	66.95 M	453.93 M	2.98 B
	∩ yes	2.06 M	8.72 M	62.31 M	404.87 M
	⊆ skipped (DAG)	7.19 M	54.80 M	370.28 M	2.18 B
	⊆ envelope	2.35 M	10.11 M	72.40 M	461.92 M
	⊆	2.27 M	9.77 M	69.83 M	446.00 M
	⊆ yes	2.11 M	9.04 M	64.95 M	418.67 M

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 ...

Conclusion — Result

- 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`

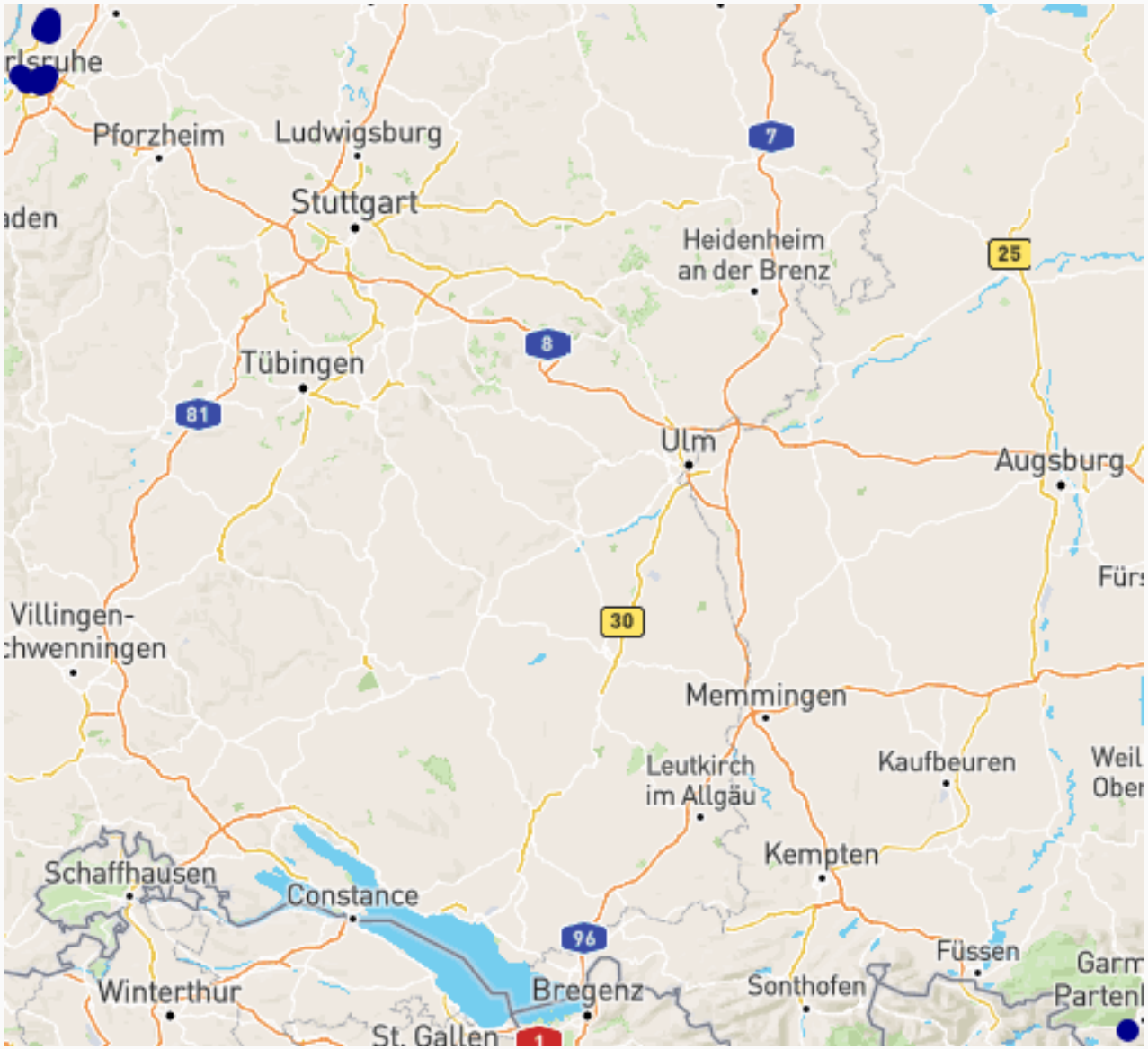
Questions?

References

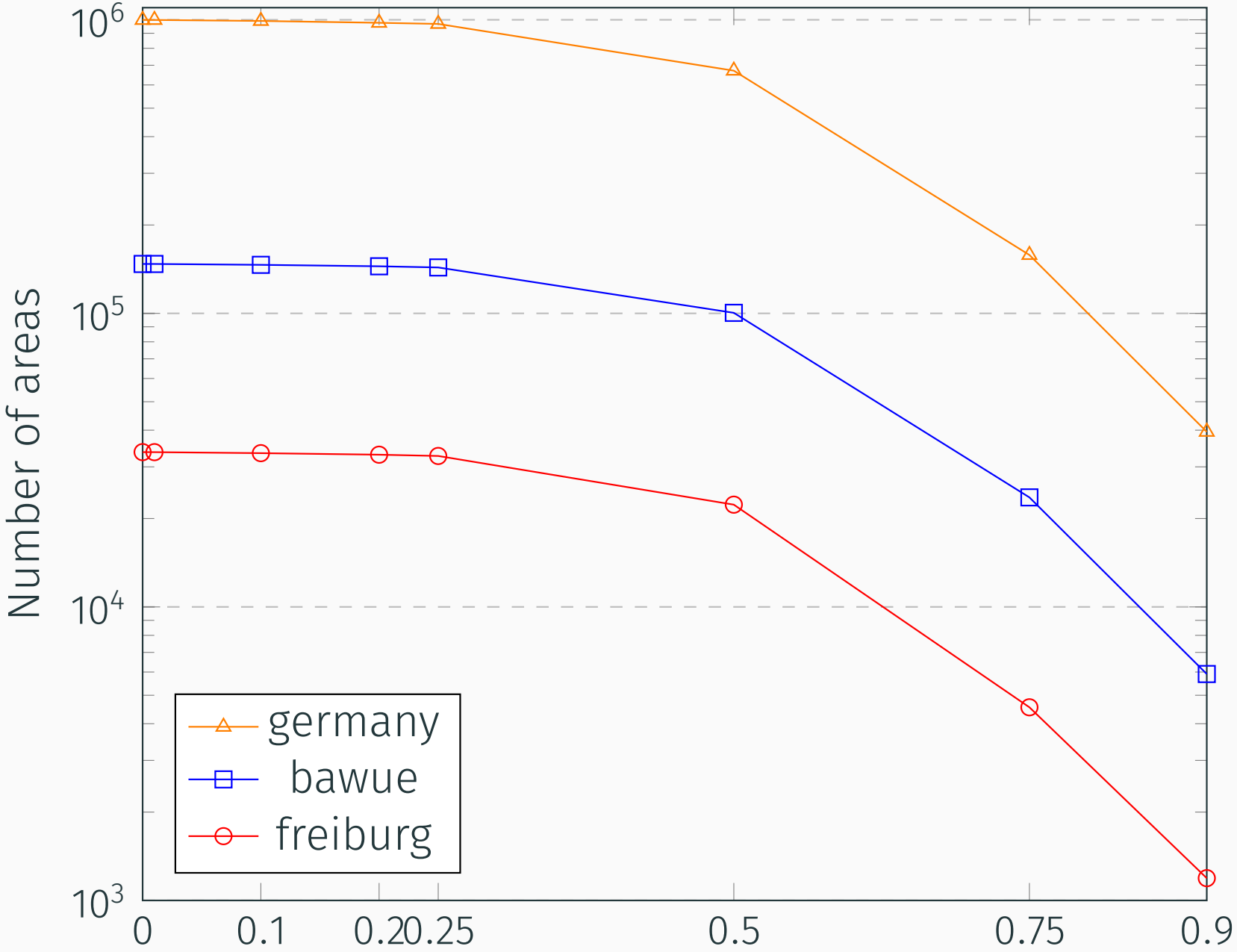
Milos Jovanovik, Timo Homburg, and Mirko Spasic. “A GeoSPARQL Compliance Benchmark”. In: *CoRR* abs/2102.06139 (2021). arXiv: 2102.06139. URL: https://arxiv.org/abs/2102.06139.

OpenStreetMap Wiki. *Sophox — OpenStreetMap Wiki*, [Online; accessed 15-March-2021]. 2021. URL: https://wiki.openstreetmap.org/w/index.php?title=Sophox&oldid=2105584.

Area/Envelope Ratio Example



Minimal Area/Envelope Ratio



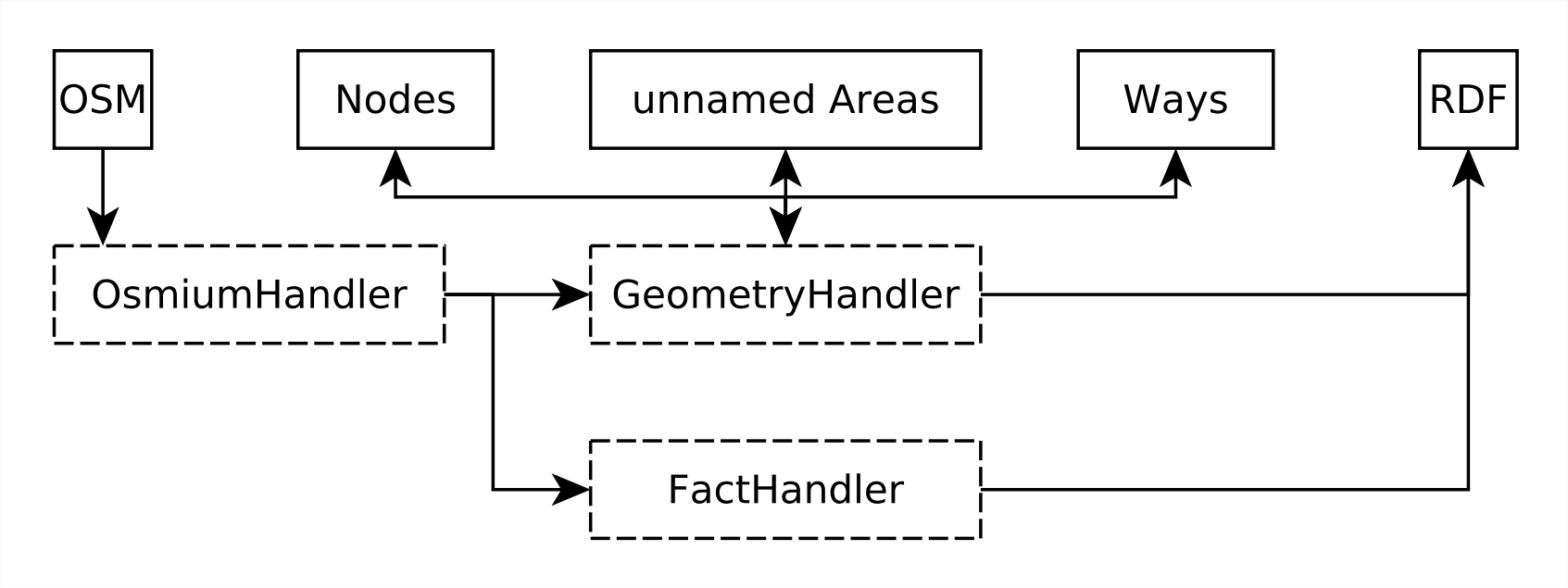
WKT — Syntax

- Node (Singe 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, …), (…)), (…))

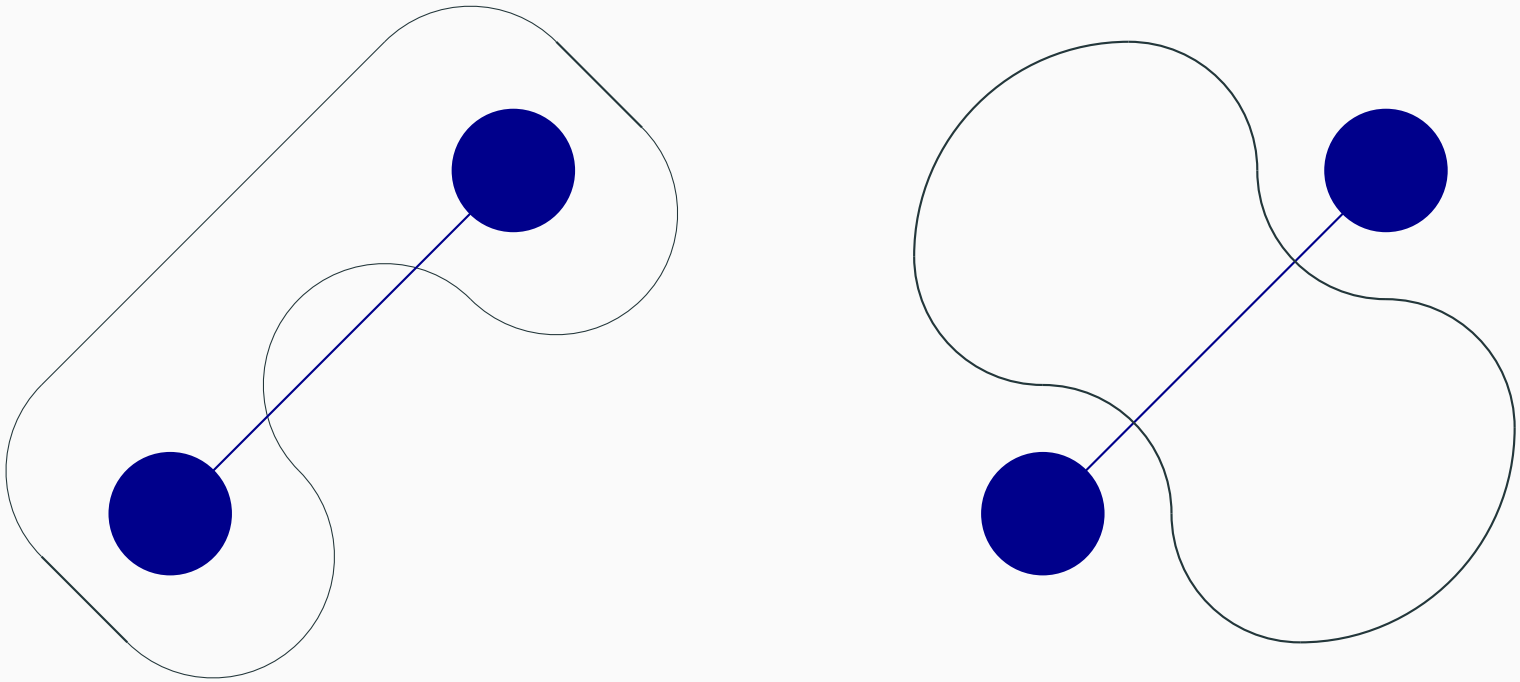
WKT — OpenStreetMap

- OpenStreetMap represents geometries explizit
  - Closed ways $(n_0, n_1, \ldots n_k) \mid n_0 = n_k$
  - 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

Data flow


OSM
Nodes
unnamed Areas
Ways
RDF
OsmiumHandler
GeometryHandler
FactHandler

Why explicit spatial calculations



Sub optimal intersection

Provided by: OpenStreetMap | Map data © OpenStreetMap contributors

The highlighted way crosses multiple administrative levels and intersects *Baden-Württemberg* and *Hesse*. It is contained in neither of them but in *Germany*.

DAG — Links — I

- 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`

DAG — Links — II

- 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`