

Computing and Querying Topological Relations in Linked Geographic Data

Using Strict, Approximate, and Metrically-Refined Topology

Blake Regalia¹, Krzysztof Janowicz¹, Grant McKenzie²

2019/07/09

¹STKO Lab, University of California, Santa Barbara, USA

²Platial Analysis Lab, McGill University, Montreal, CA

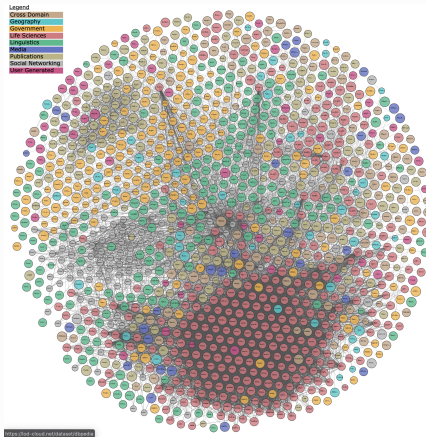


Figure 1 lod-cloud.net

The **Web of Linked Open Data** , aka the **LOD cloud**, is an open, interlinked collection of cross-domain knowledge bases from governments, industries, academic institutions and non-profit organizations.

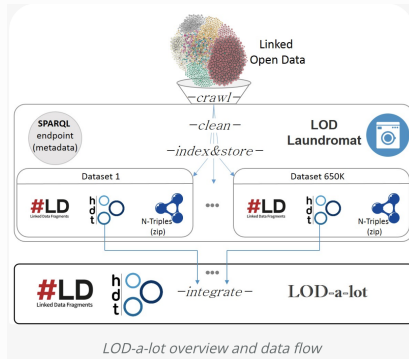


Figure 2 LOD-a-lot: lod-a-lot.lod.labs.vu.nl/

When cleaned, the cloud contains more than **28.3 billion** unique machine-readable statements derived from manual input, annotated datasets, extracted texts, sensor observations, conceptual abstractions, and so on.

Representations of **concrete objects** from the world such as persons and **places** constitute **a majority of central nodes** within the multigraph, linking statements and relations *across* datasets.

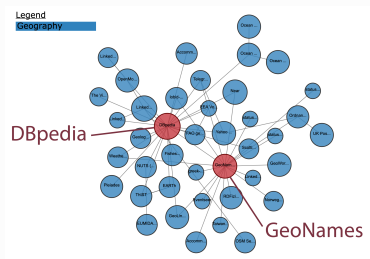


Figure 3 lod-cloud.net

DBpedia and GeoNames are the **most central dataset hubs** for **geographic data** in the cloud. DBpedia alone:

*“[...] currently describes **4.58 million things**, [...] including 1,445,000 persons, **735,000 places** (including **478,000 populated places**)”*

<https://wiki.dbpedia.org/about/facts-figures>

blake.regalia@gmail.com

phuzzy.link

SPARQL endpoint URL: <https://dbpedia.org/sparql>

Prefix JSON-LD context: <https://phuzzy.link/context/default>










Settings:
language: "en"
locale: "en-US"
limit: 128

Plugins:
phuzzy-xsd: 1 arg
phuzzy-colored-prefixes: 1 arg
phuzzy-language-filter: 1 arg
phuzzy-info: 1 arg
phuzzy-geo

Outgoing properties: 933 triples

Incoming properties: Loaded 640 triples...

dbr:San_Diego

elevation (μ)	→ "18.8976" ^{^^xsd:double}
founding_date	→  July 15, 1769 ^{^^xsd:date}  March 26, 1850 ^{^^xsd:date}
governing_body	→  San Diego City Council
government_type	→  Strong mayor
is part of	→  California  San Diego County, California
leader_name	→  Jan Goldsmith  Kevin Faulconer
leader_title	→ "City Attorney" @en "City Council" @en "Mayor" @en
maximum_elevation (μ)	→ "484.937" ^{^^xsd:double}
minimum_elevation (μ)	→ "0.0" ^{^^xsd:double}
motto	→ "Semper Vigilians (Latinfor "Ever Vigilant")"
percentage_of_area_water	→ "12.68" ^{^^xsd:float}
population_as_of	→  June 30, 2015 ^{^^xsd:date}
population_density (/sqkm)	→ "1545.4" ^{^^xsd:double}
population_metro	→ "3095313" ^{^^xsd:nonNegativeInteger}
population_total	→ "1394928" ^{^^xsd:nonNegativeInteger}

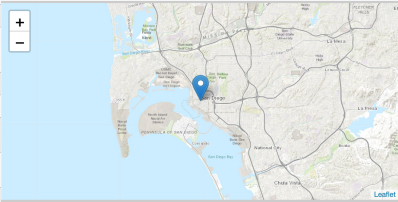


Figure 4 DBpedia excerpt from phuzzy.link

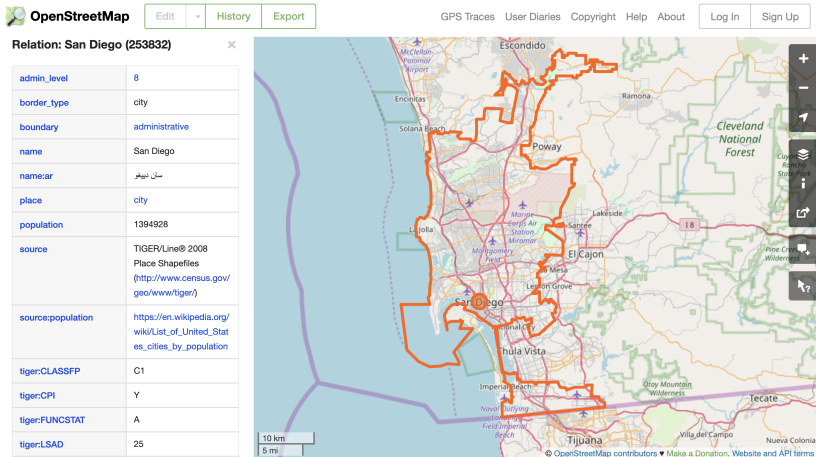


Figure 5 San Diego on OpenStreetMap

Motivation

The majority of geographic identifiers on LOD cloud are represented geometrically as **point coordinates** , severely limiting their potential for **spatial analysis**.

However, simply integrating more **complex geometries** into the LOD cloud **will be of limited use** to the Geospatial Linked Data community because:

- **Graph queries** involving spatial operations on high-resolution geometries **do not scale** well over large datasets.
- Geometries are merely a **means to an end** for spatial analysis, the proper geometric representation of real-world entities varies by place type, scale, and task.
- Spatial extensions to RDF triplestores, such as GeoSPARQL, compute topology *on-demand* and **without context** , meaning that topology is computed from the geometries alone; **place types are ignored**.
- Computing topology **requires pre-processing** steps to **clean geometries** for errors such as sliver polygons, something not currently supported by Linked Data based frameworks.

Topology Matters, Metric Refines:

"In geographic space, topology is considered to be first-class information, whereas metric properties, such as distances and shapes, are used as refinements that are frequently less exactly captured."

Egenhofer and Mark 1995b

Problem Statement:

Following the slogan that “Topology Matters, Metric Refines”, knowledge graphs will benefit from **explicit topological relations** in addition to **complex geometries** and other place-specific properties.

Computing topology **based on geometry alone** is not sufficient in the context of Linked Data for two reasons we focus on here:

1. Heterogeneous data sources and crowd-sourced datasets propagate digitization errors along complex geometries, meaning that open data clouds **will always require pre-processing** steps in order to compute topology.
2. Topology also depends on domain knowledge, **vagueness** and uncertainty principles (Bennet, 2001).

Research Objective:

Enrich the geographic LOD cloud with **topological relations** rather than purely complex geometries.

Contribution 1:

Align OpenStreetMap features with **DBpedia** places to combine **complex geometries** (polylines and polygons) with rich **place type information** .

Knowledge Base Alignment:

Starting with all **DBpedia places** in the contiguous United States; select those with the **following place types** and align them with their matching feature in **OpenStreetMap**:

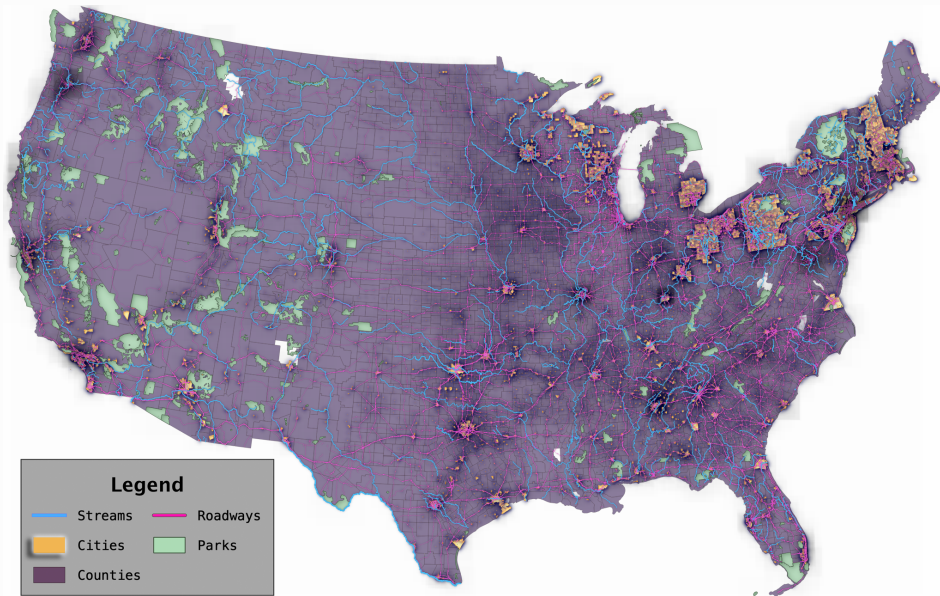
Polylines:

- Roadways
- Streams

Polygons:

- Cities
- Counties
- Parks

36,520 features in total .



Contribution 2:

Precompute and materialize **strict, approximate, and metrically refined** topological relations between cleaned geometries using place type knowledge.

Strict Topology:

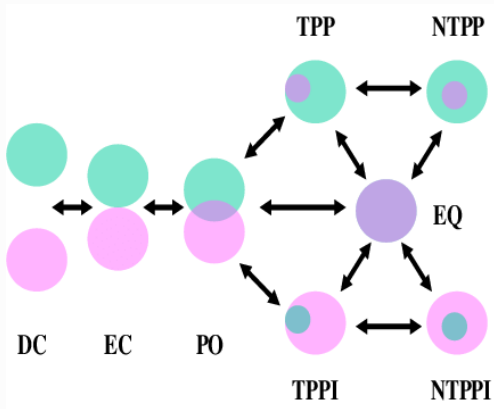


Figure 6 Region Connection Calculus (RCC8)

Approximate Topology:

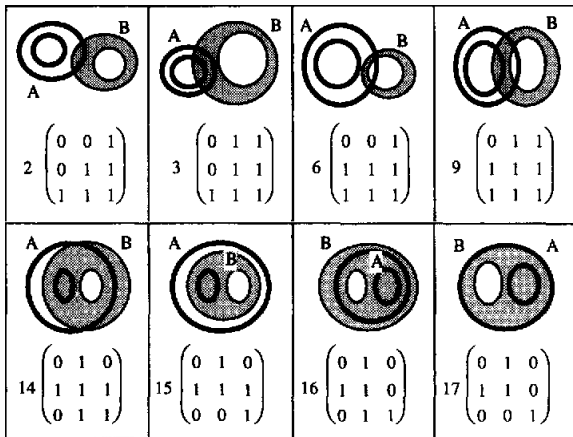


Figure 7 Broad boundaries (Clementini et al.)

Metrically Refined Topology:

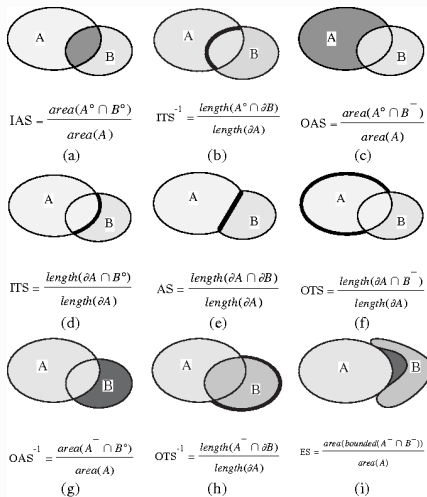


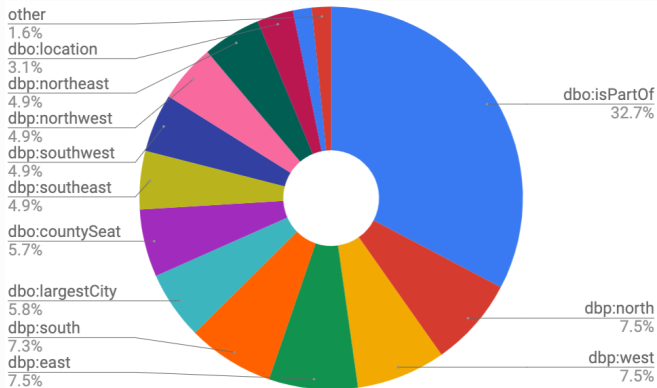
Figure 8 Nine Splitting Measures (Egenhofer et al.)

Code	Types	Description
	L	Refers to (Multi)Polyline geometry types.
	G	Refers to (Multi)Polygon geometry types.
	E	Refers to <i>either</i> of the two aforementioned geometry types.
		Crisp Boundary Relations for G/G pairs – RCC8 (Cohn et al., 1997)
DC	G/G	Disconnected
EC	G/G	Externally Connected
PO	G/G	Partially Overlaps
EQ	G/G	Equals
TPP/i	G/G	Tangential Proper Part \cup Tangential Proper Part Inverse
NTPP/i	G/G	Non-Tangential Proper Part \cup Non-Tangential Proper Part Inverse
		Crisp Boundary Relations for L/G pairs – as used by Formica et al. (2012).
TCH	L/E	Touches
PTH	L/G	Passes Through
INC	E/L	Inclusion
		Crisp Boundary Relations for L/L pairs Formica et al. (2018).
CRS	L/L	Crosses
TCS	L/L	Touch Crosses $\subset TCH$
		Broad Boundary Relations – as defined by Clementini and Di Felice (2001).
nM	E/G	Nearly Meets $\subset DC$
nCt	G/G	Nearly Contains $\subset PO$
nE	G/G	Nearly Equals $\subset (PO \cup TPP/i \cup NTPP/i)$
		Metrically-Refined Topological Relations
mW	G/G	Mostly Within $\subset PO$: the area of intersection is greater than or equal to 80% of P_1 's area.
bT	G/G	Barely Touches $\subset EC$: the spheroidal length of the intersecting boundary is less than 10m.
RAL RAS	L/E	Runs Along (L/L), Runs Alongside (L/G): the area of intersection between the features' broad boundary buffers is greater than some threshold value as described in Section 3.2.
CON	L/L	Connects $\subset TCH$: at least one of the points where the polylines intersect is collocated with one of the points that either polyline starts or ends.

Table 1 Topological operator codes as defined by related works as well as our custom *metrically-refined* operator codes. P_1 refers to the polygon with lesser area and P_2 the polygon with greater area.

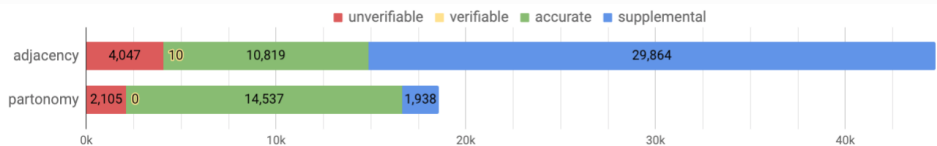
The **resulting topological relations dataset** produces **120,681 distinct RDF statements** covering various topological relations between features of the selected place types within the contiguous United States.

Putting these topological relations to use, we attempt to **validate** any existing relations between places on DBpedia.



Distribution of all place-place relation predicates:

1. **50% Adjacency**: primary and inter-primary cardinal direction relations, as well as “adjacent communities” linksets.
2. **48% Partonomy**: `dbo:isPartOf`, `dbo:largestCity`, `dbo:countySeat`, `dbo:location` predicates.



1. **Unverifiable:** Relation is absent from our dataset
2. **Verifiable:** Relation exists in both datasets, but the topological relation we observe does not align with the expected predicate.
3. **Accurate:** Relation exists in both datasets and the topological relation aligns with the expected predicate.
4. **Supplemental:** Relation is absent from DBpedia; demonstrates volume of our contribution to enriching the knowledge base.

How does this compare to GeoSPARQL?

```

1  # Using GeoSPARQL
2  select ?countyA ?borderingCounties {
3      select ?countyA (count(?countyB) as ?borderingCounties) {
4          ?countyA a experiment:County ; geosparql:hasGeometry ?geomA .
5          ?countyB a experiment:County ; geosparql:hasGeometry ?geomB .
6          filter(geof:sfTouches(?geomA, ?geomB))
7      } group by ?countyA
8  } order by desc(?borderingCounties)

```

Listing 1.4 Query for all bordering counties using GeoSPARQL's *extensible value testing* function `geof:sfTouches`, which computes the **EC** topological relation on-the-fly.

```

1  # Using our precomputed topological dataset
2  select ?countyA ?borderingCounties where {
3      select ?countyA (count(?countyB) as ?borderingCounties) {
4          ?countyA a experiment:County . ?countyB a experiment:County .
5          { ?countyA agt:touches ?countyB }
6          union { ?countyB agt:touches ?countyA }
7      } group by ?countyA
8  } order by desc(?borderingCounties)

```

Listing 1.5 Query for all bordering counties using our `agt:touches` predicate, which represents the **EC** topological relation that was materialized by precomputing topology for all features.

Figure 9 Selecting bordering counties when topology must be **computed on-demand** creates a combinatorial explosion. **Our precomputed approach** is optimized for graph queries since the topological relations are already materialized in the triplestore.



Figure 10 **GeoSPARQL fails** to capture any relations due to tiny sliver polygons. **Our approach** **captures all 8** *externally connected* relations and qualifies one as *barely touches*.

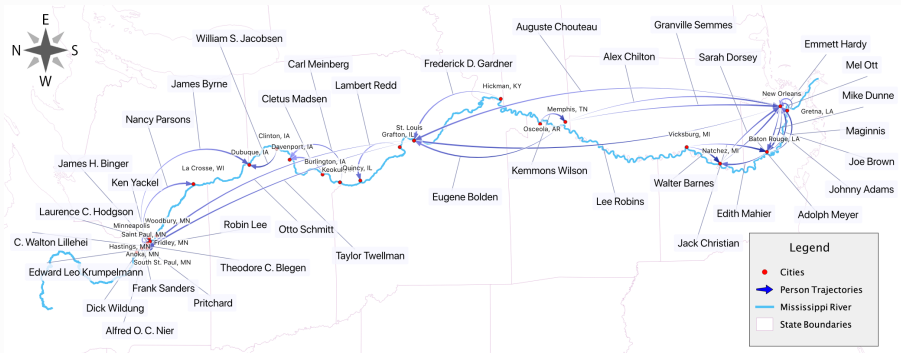
How can topology be used on geospatial linked data in practice?

```

1  select ?person ?placeBorn ?placeDied where {
2      ?placeBorn a :City .  ?placeDied a :City .
3
4      dbr:Mississippi_River ?interactsA ?placeBorn .
5      values ?interactsA { agt:touches agt:crosses agt:nearlyMeets }
6
7      dbr:Mississippi_River ?interactsB ?placeDied .
8      values ?interactsB { agt:touches agt:crosses agt:nearlyMeets }
9
10     filter(?placeBorn != ?placeDied)
11
12     service <http://dbpedia.org/sparql/> {
13         ?person a dbo:Person ;
14         dbo:birthPlace ?placeBorn ;
15         dbo:deathPlace ?placeDied .
16     }
17 }

```

Figure 11 SPARQL query asking: What persons of historical significance were born in **a city along the Mississippi River** and then died in another city also along the river?



Topological Reasoning:

“Coastal cities” are a class of cities that have an adjacency relation to the ocean.

“Landlocked regions” are administrative regions that lack direct access to some resource due to being entirely surrounded by other administrative regions.

“Accessibility” reasoning:

$$\begin{aligned}\text{County} &\sqcap (\exists \text{PO.County} \sqcup \exists \text{NTPP.County} \sqcup \exists \text{TPP.County}) \sqsubseteq \perp \\ \text{State} &\sqcap (\exists \text{PO.State} \sqcup \exists \text{NTPP.State} \sqcup \exists \text{TPP.State}) \sqsubseteq \perp \\ \text{NTC} &\equiv \text{County} \sqcap \exists \text{NTPP.State} \\ \text{ParksInNTC} &\equiv \text{Park} \sqcap (\forall \text{PO.NTC} \sqcup \exists \text{NTPP.NTC})\end{aligned}$$

Parks that can be accessed via counties that are non-tangential proper parts to their state boundary.

Thank you

												avg. area of...	
region-region	EQ	EC	PO	TPP/i	NTPP/i	nE	nM	nCt	mW	bT		smaller polygon	larger polygon
park-park	1	220	9	10	49	0	84	0	3	4		477km ²	3,952km ²
park-city	0	283	160	79	740	0	120	14	47	291		22km ²	617km ²
park-county	0	516	512	135	1,645	0	15	1	74	439		411km ²	4,971km ²
city-city	0	11,827	48	58	189	0	386	0	20	27		65km ²	170km ²
city-county	1	6,768	1,046	3,397	12,496	0	84	5	280	880		40km ²	2,694km ²
county-county	0	9,117	0	1	9	0	25	0	0	0		2,048km ²	3,302km ²

Table 2 Number of region-to-region relations materialized for each place type combination by row, and each topological relation by column using codes defined in Table 1.

							<i>avg. length/area of...</i>	
polyline-region	PTH	TCH	INC	nM	bT	RAS	<i>polyline</i>	<i>polygon</i>
road-park	137	984	155	13,928	10	11	316km	1,169km ²
road-city	3,072	17,302	3,303	19,528	106	100	425km	137km ²
road-county	7,041	5,597	3,751	4,579	213	9	383km	2,739km ²
stream-park	156	220	123	1,303	3	5	293km	4,180km ²
stream-city	708	2,516	285	2,973	106	382	408km	258km ²
stream-county	1,502	1,491	1,718	828	118	241	418km	4,221km ²

Table 3 Number of polyline-to-region relations materialized for each place type combination by row, and each topological relation by column using codes defined in Table 1.

						<i>avg. length of...</i>	
polyline-polyline	CRS	TCS	CON	nM	RAL	<i>shorter polyline</i>	<i>longer polyline</i>
road-road	9,861	2	658	100,790	65	20km	127km
road-stream	4,109	0	84	7,922	94	79km	556km
stream-stream	12	0	237	4,573	2	5km	24km

Table 4 Number of polyline-to-polyline relations materialized for each place type combination by row, and each topological relation by column using codes defined in Table 1.