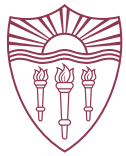


Building Spatio-Temporal Knowledge Graphs from Vectorized Topographic Historical Maps

April 5th, 2021

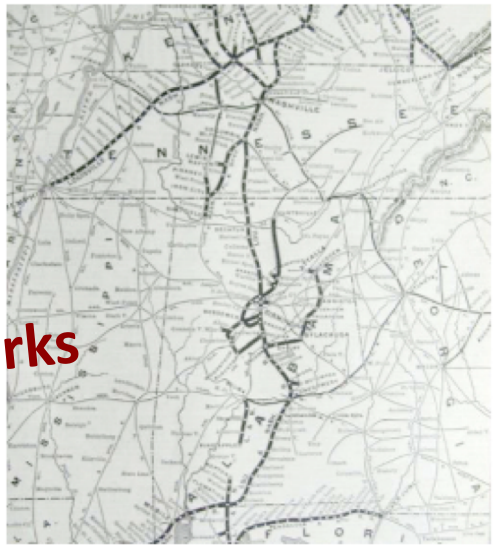
Basel Shbita[1], Craig A. Knoblock[1], Weiwei Duan[2], Yao-Yi Chiang[2], Johannes H. Uhl[3], and Stefan Leyk[3]

[1] Information Sciences Institute and Department of Computer Science, USC
[2] Spatial Sciences Institute and Department of Computer Science, USC
[3] Department of Geography, University of Colorado Boulder

Problem

Digitized Historical Maps
=
Rich sources of information

1886

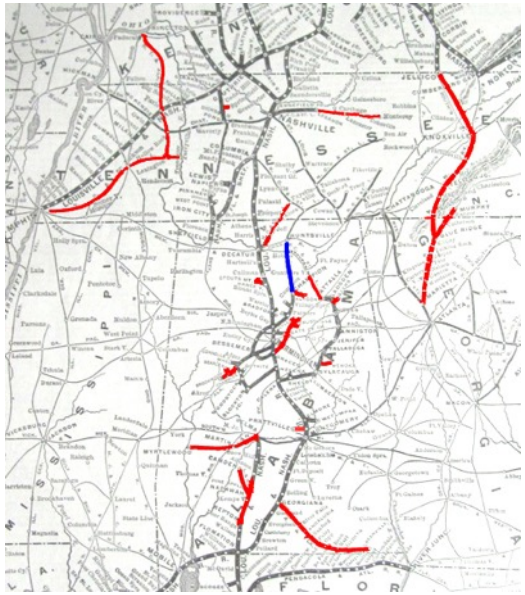


1904

Natural- & Human-made features

Wetlands

Railroad Networks



Labor-intensive to analyze across time & space

Additional discovery

Idea

Decompose

to building blocks

OpenStreetMap

then use

Linked Data & the Semantic Web

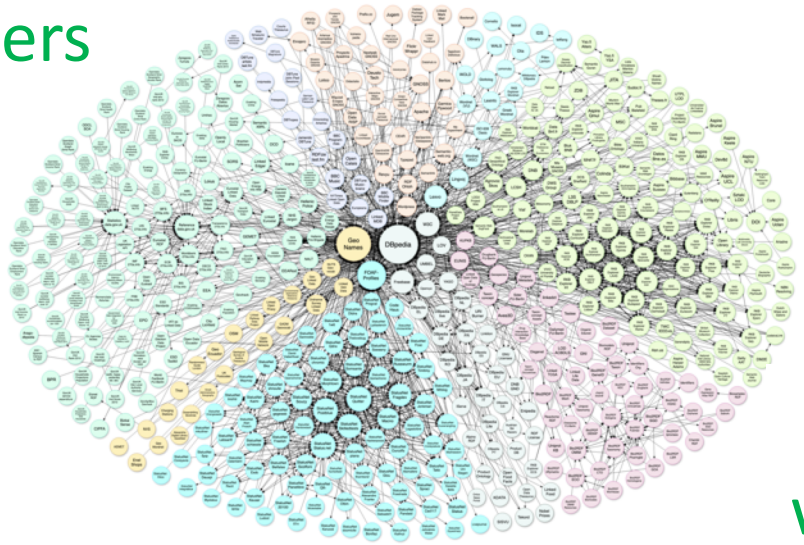
to build a

Knowledge Graph

Why Linked Data?

Break down data barriers

Across-domains



Structured

interpreted by **humans** & **machines**

Semantic

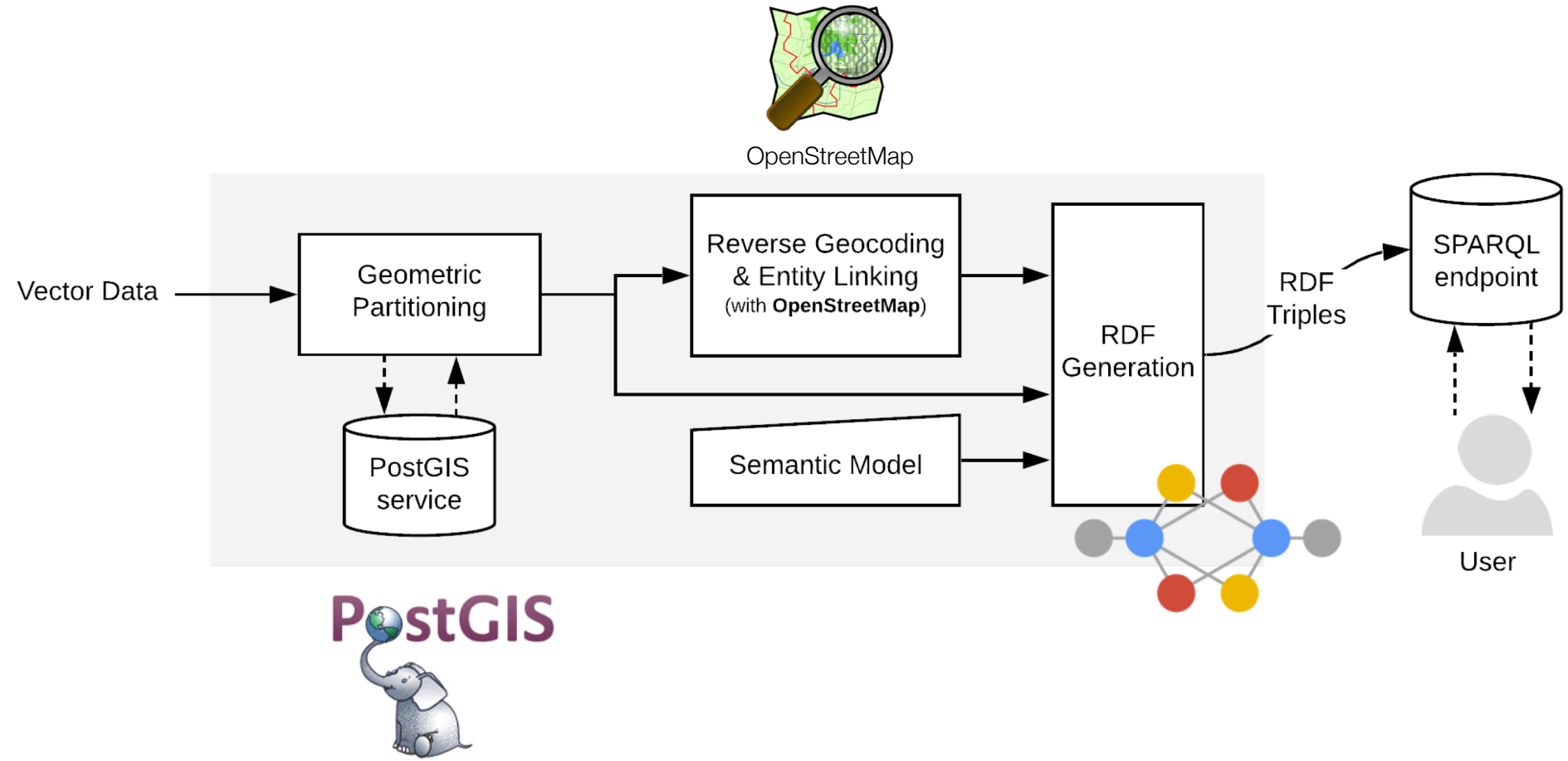
relationships, properties, metadata

Make data widely available

Fuel **Discovery**

Query & **visualize**

Our approach



Automatic feature partitioning

- Goal: create building blocks of the topographic features (geometry)
  - Entity matching/linking & entity "partitioning" task
  - "Granular Building Blocks"

- Use a spatially-enabled database service (PostGIS)
  - PostgreSQL extension
    - Manipulate & transform spatial data

- Allow incremental additions over time

Automatic feature partitioning

vector data from various map editions ("initial" building blocks)

current "building blocks"

foreach $i \in \mathcal{M}$ **do**

  foreach $k \in \mathcal{L}$ **do**

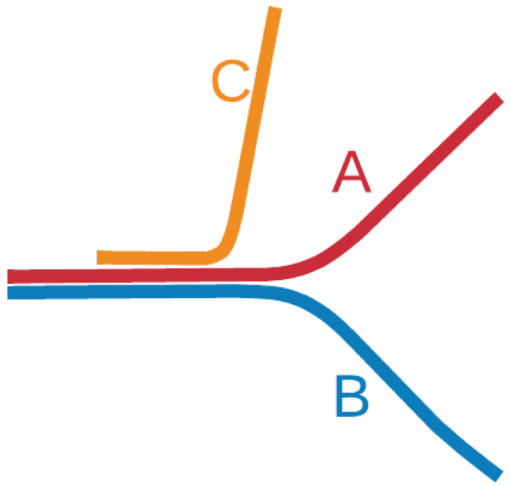
    $\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$

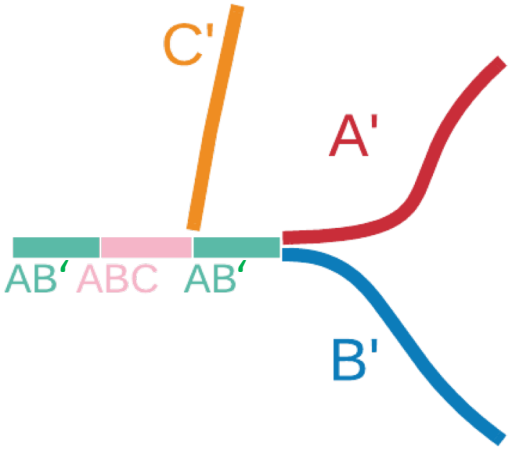
    $\mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha;$

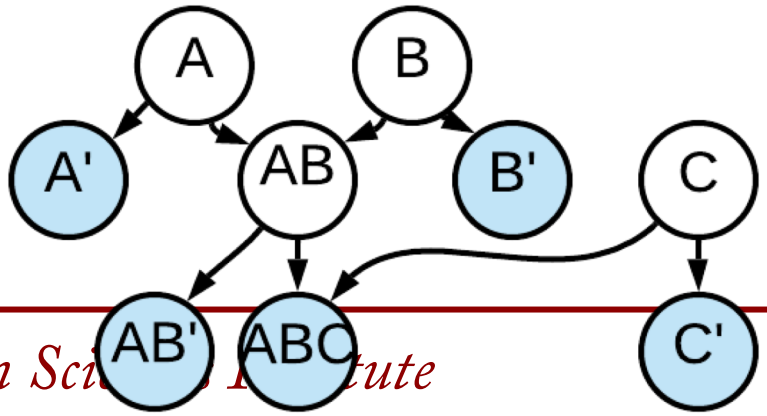
  end

  $\mathcal{F}_\delta = \mathcal{F}_i \setminus (\bigcup_{j \in \mathcal{L}} F_j);$

end







Automatic feature partitioning

vector data from various map editions ("initial" building blocks)

current "building blocks"

foreach $i \in \mathcal{M}$ **do**

  foreach $k \in \mathcal{L}$ **do**

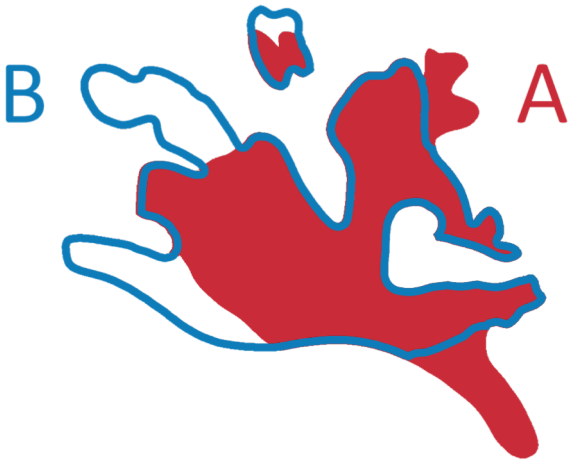
    $\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$

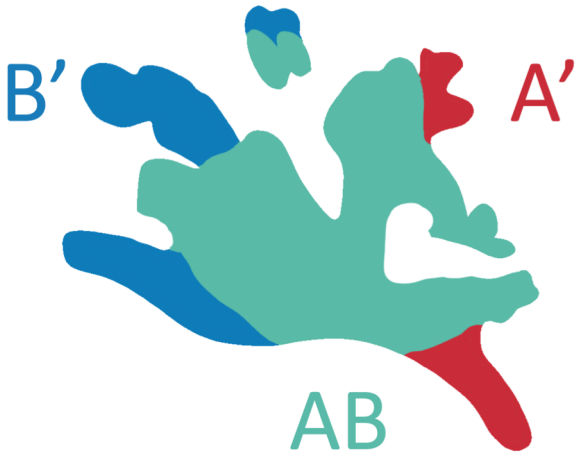
    $\mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha;$

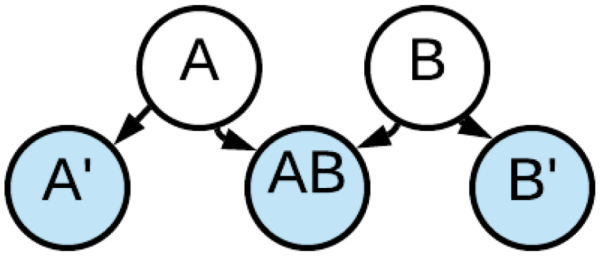
  end

  $\mathcal{F}_\delta = \mathcal{F}_i \setminus (\bigcup_{j \in \mathcal{L}} F_j);$

end







Geo entity-linking

- Goal: map building blocks to Open KBs
  - Entity matching
  - Enrich data to fuel discovery

- Use a reverse geocoding service (OpenStreetMap) for initial filtering

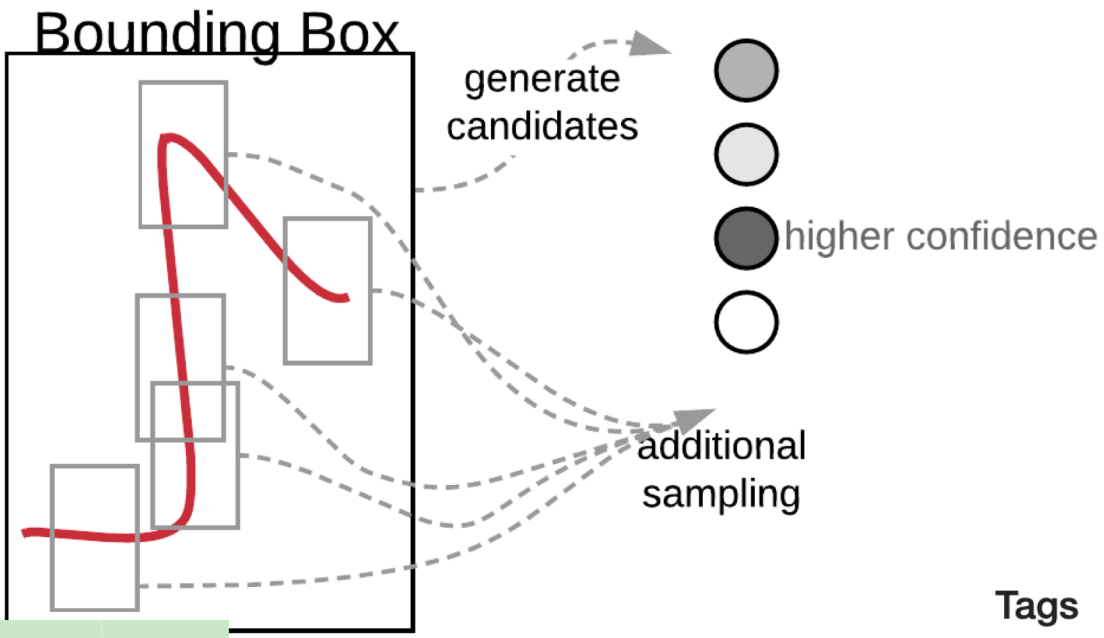
- Determine confidence by (random) sample match frequency

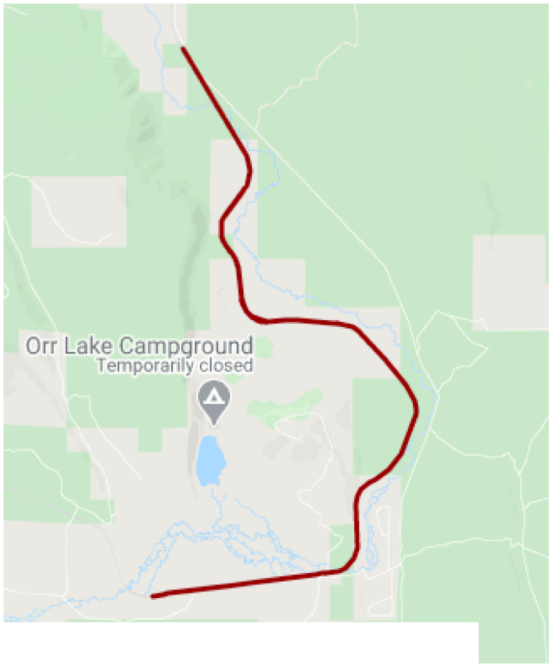
OpenStreetMap

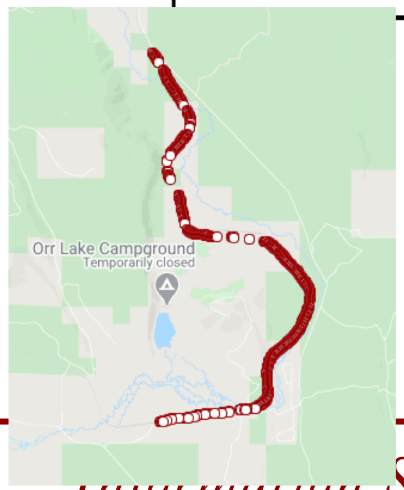
Geo entity-linking

B_s = bounding box wrapping s;
$\mathcal{L}$ = reverse-geocoding(B_s, T);
for 1...N **do**
    e = randomly sample a `Point` in segment s;
    E = reverse-geocoding(e, T);
    $\mathcal{L}$.add(E);
end
filter out instances with a single appearance in $\mathcal{L}$;
return $\mathcal{L}$;

Geo entity-linking







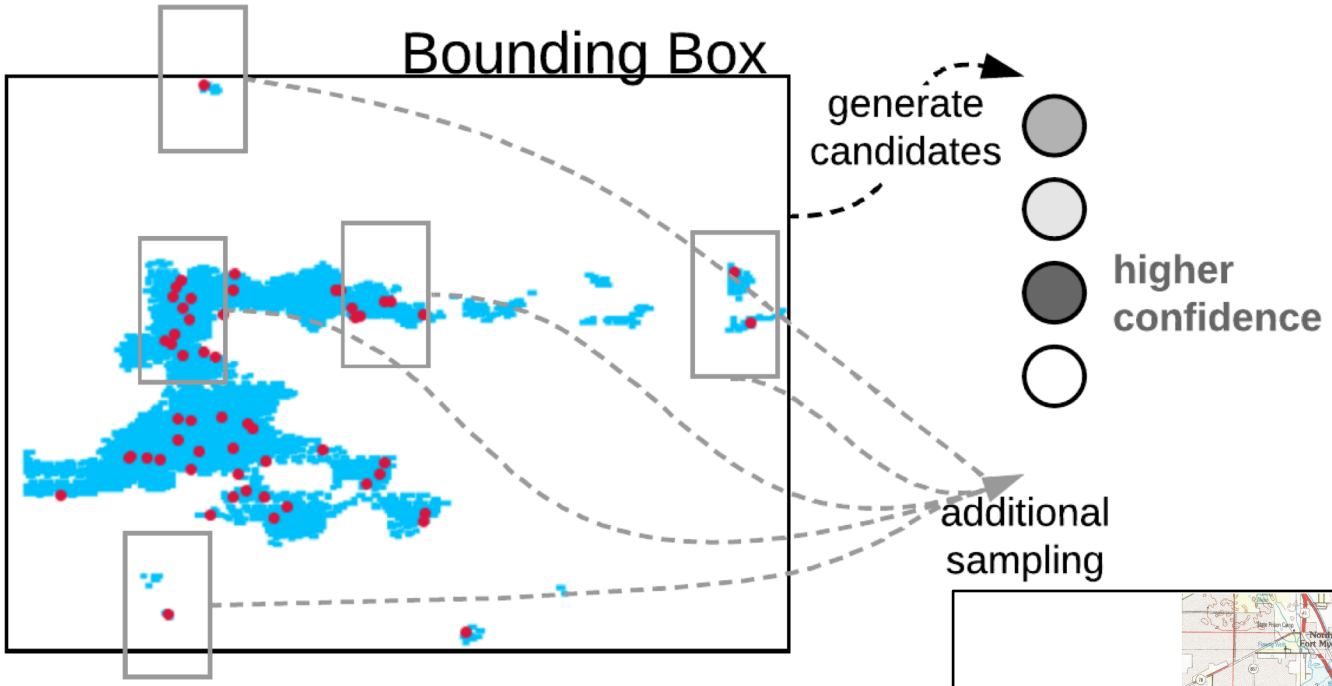
Way: Black Butte Subdivision (322131253)

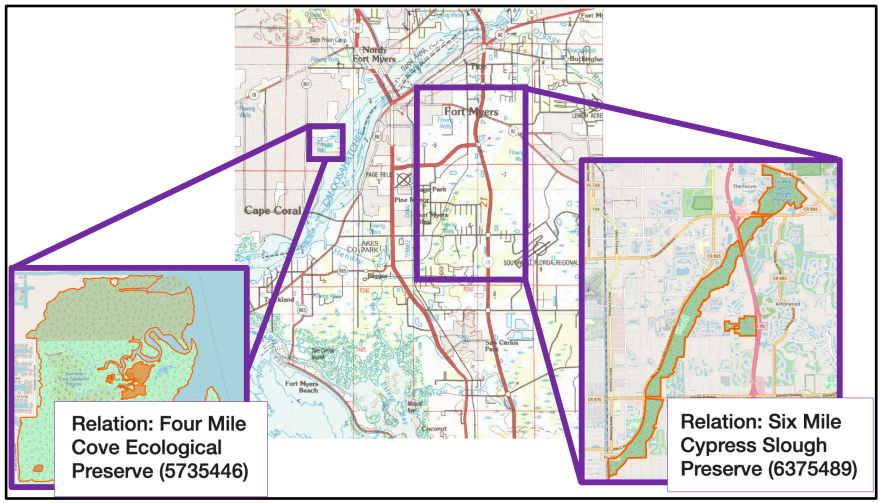
Version #6

Tags

name	Black Butte Subdivision
operator	Union Pacific Railroad;Amtrak
owner	Union Pacific Railroad
railway	rail

Geo entity-linking



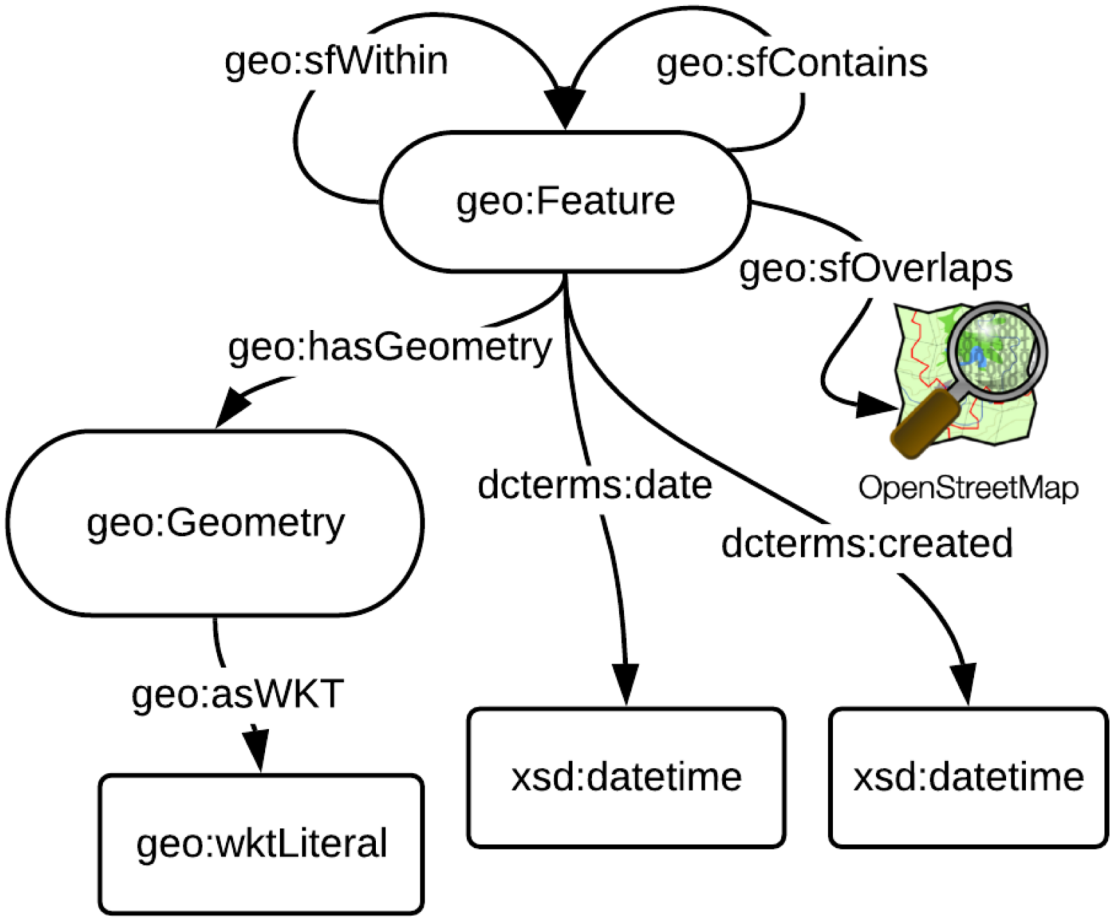


Generate RDF

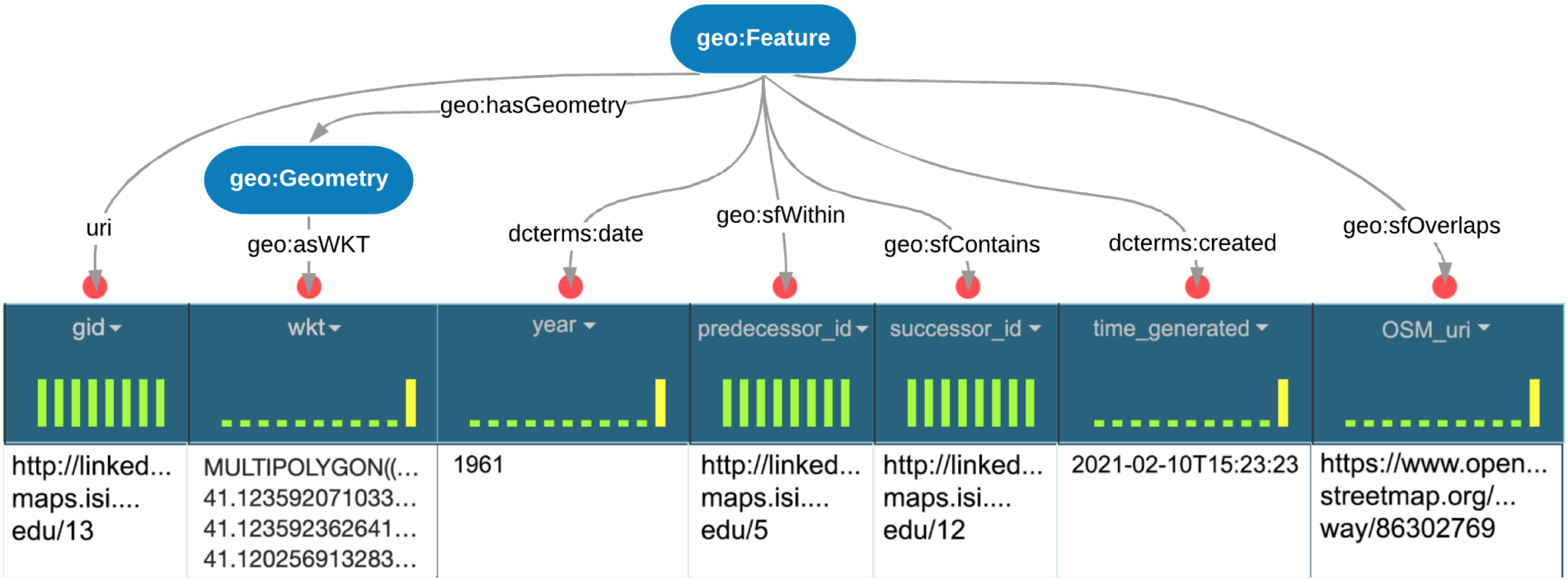
- Goal: construct a KG from the data we generated & collected
  - Useful semantic representation
  - Support downstream **spatial reasoners**

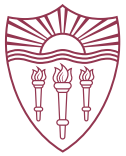
- Construct a meaningful semantic model
  - OGC GeoSPARQL standard
  - Universal conventions
  - Easily queried

Semantic Model



Data Modeling

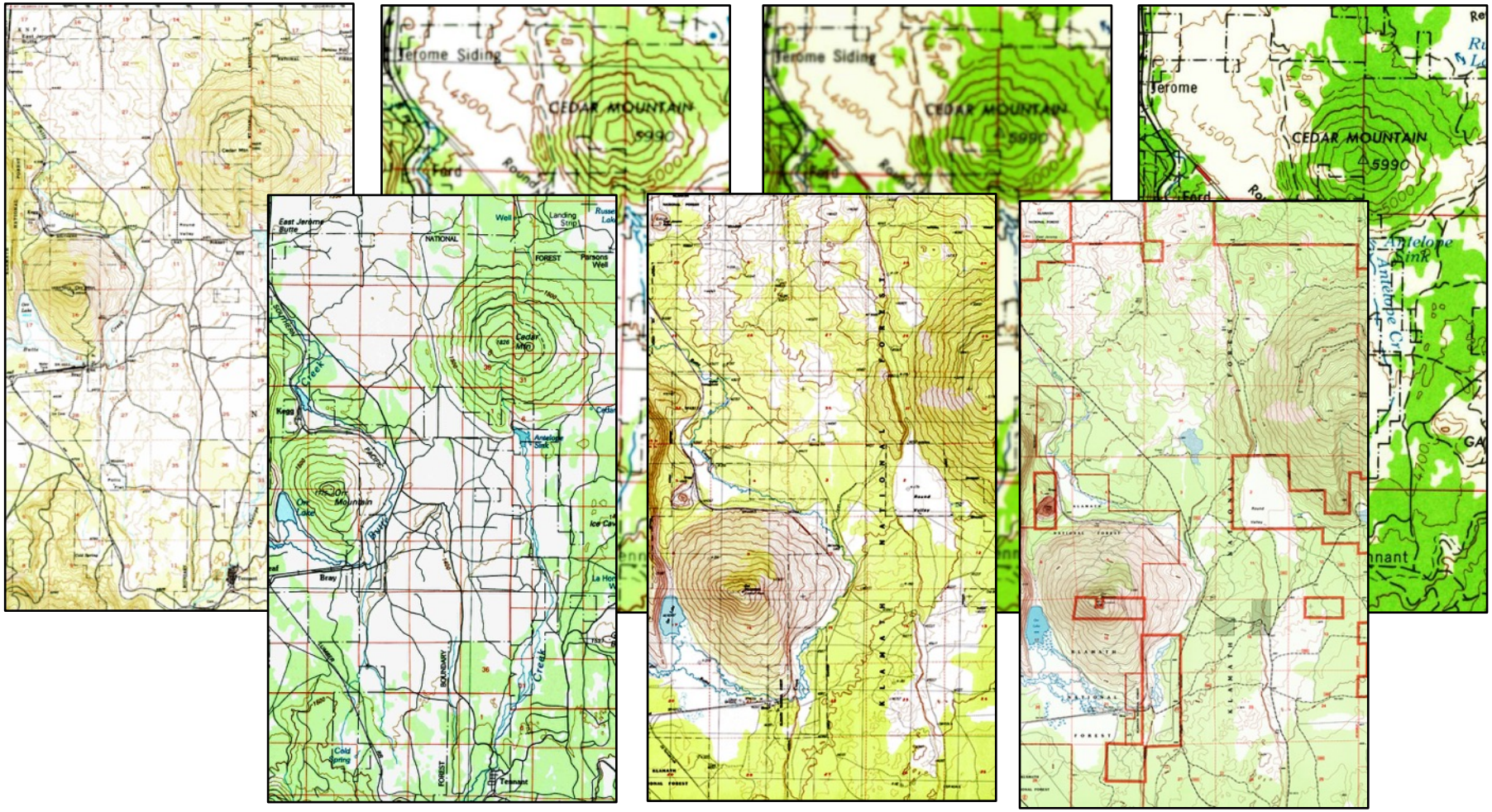




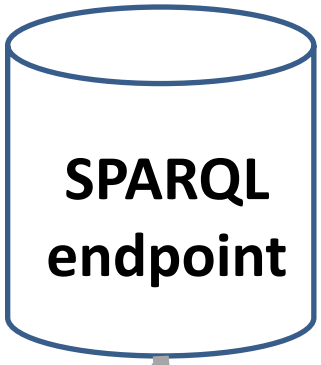
We constructed a KG

Now what?

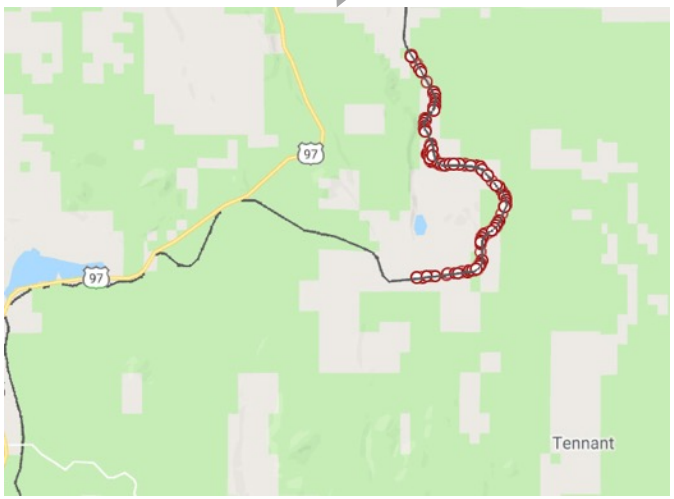
Example 1: railroads (CA)



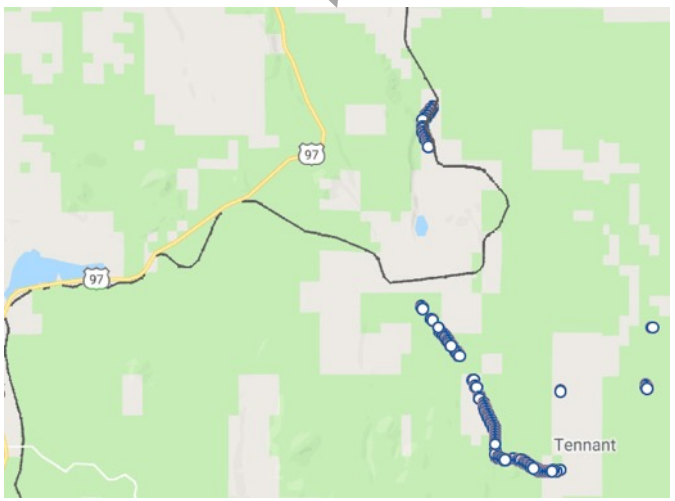
Example 1



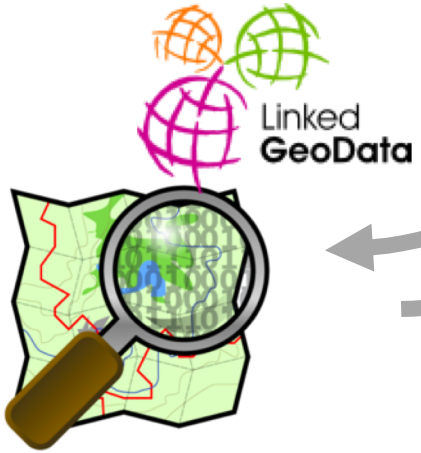
Railroad segments that are similar in 1962 and 2001



Railroad segments that are present in 1962 but are not present in 2001



Example 1



OpenStreetMap

SPARQL endpoint

Can you show me a subset of what's abandoned?

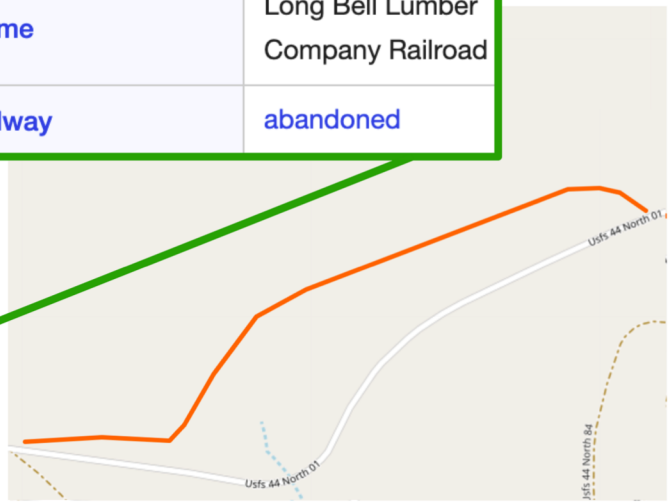
Way: Long Bell Lumber Company Railroad (177559134)

Version #3

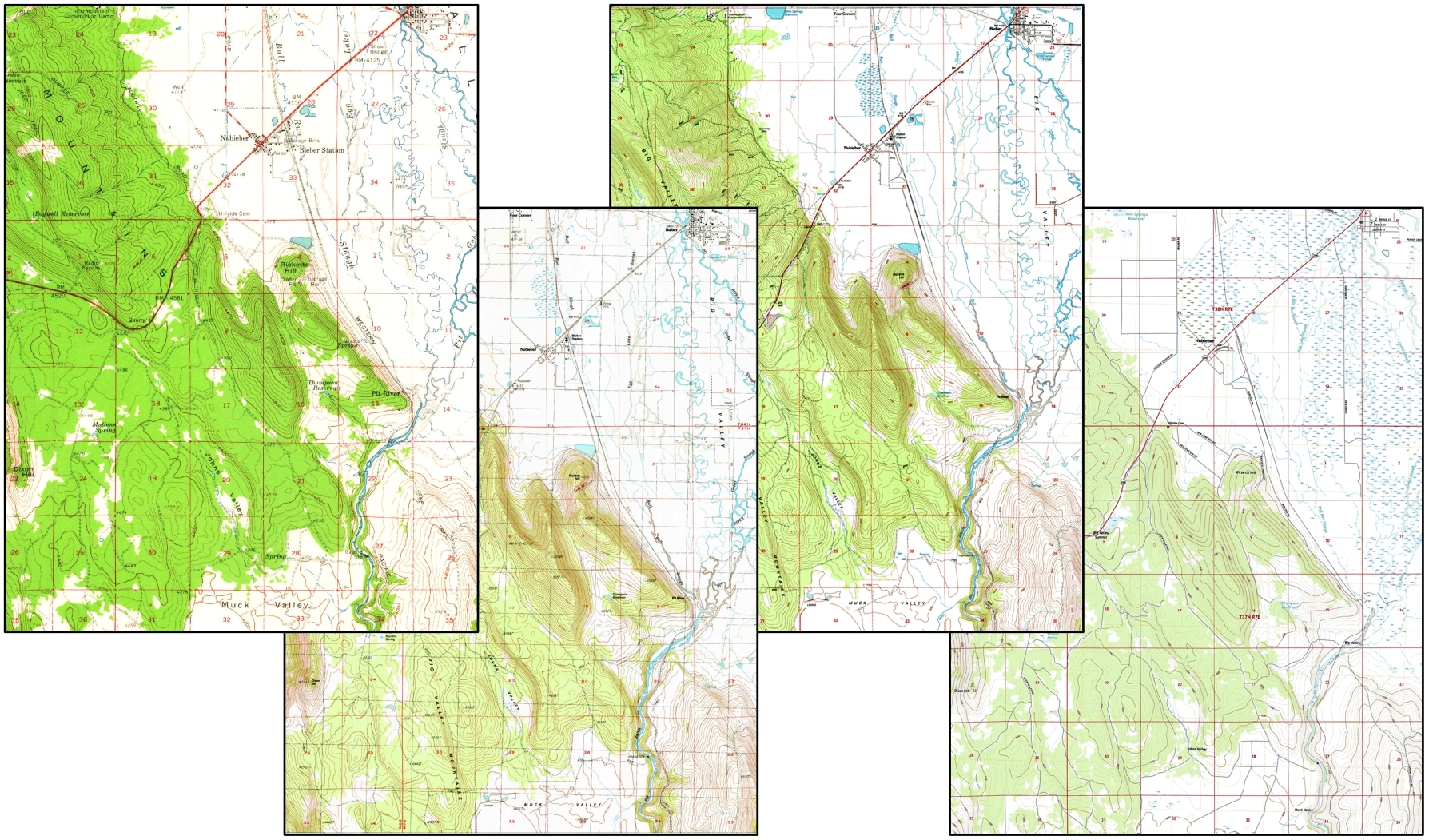
Tags

name	Long Bell Lumber Company Railroad
railway	abandoned
tiger:cfcc	B11
tiger:county	Siskiyou, CA
tiger:name_base	Long Bell Lumber Company RR

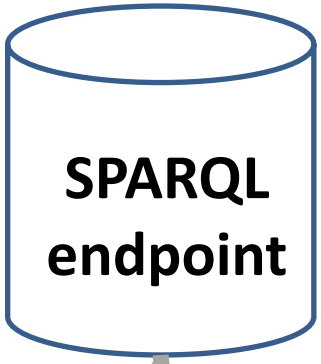
name	Long Bell Lumber Company Railroad
railway	abandoned



Example 2: wetlands (CA)

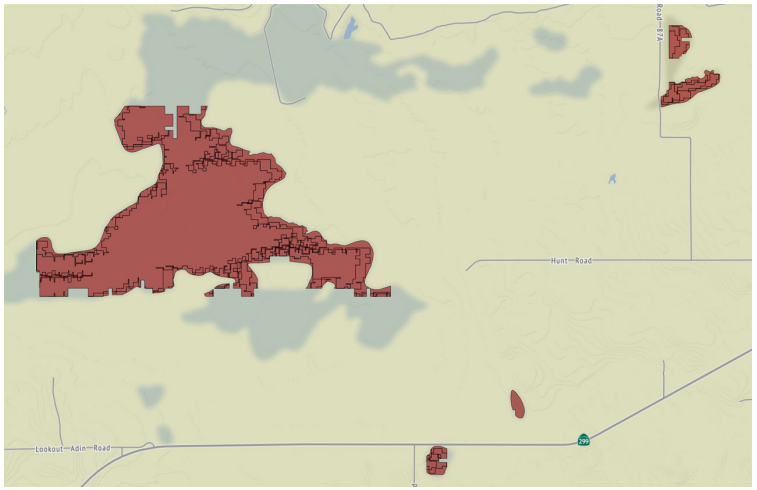


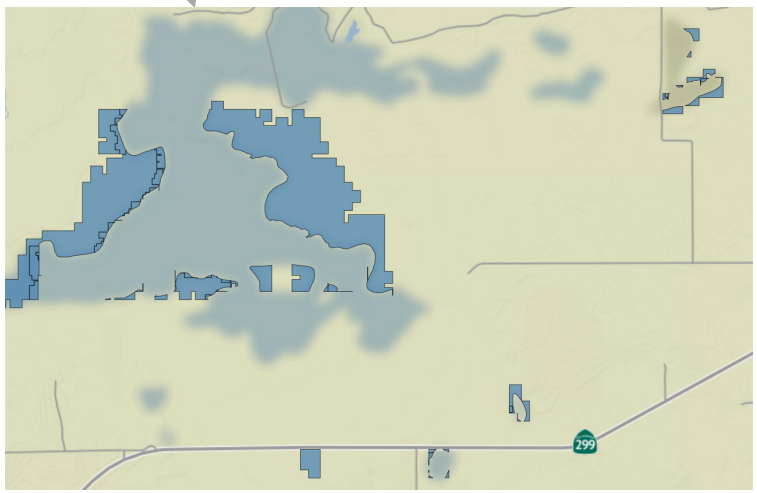
Example 2



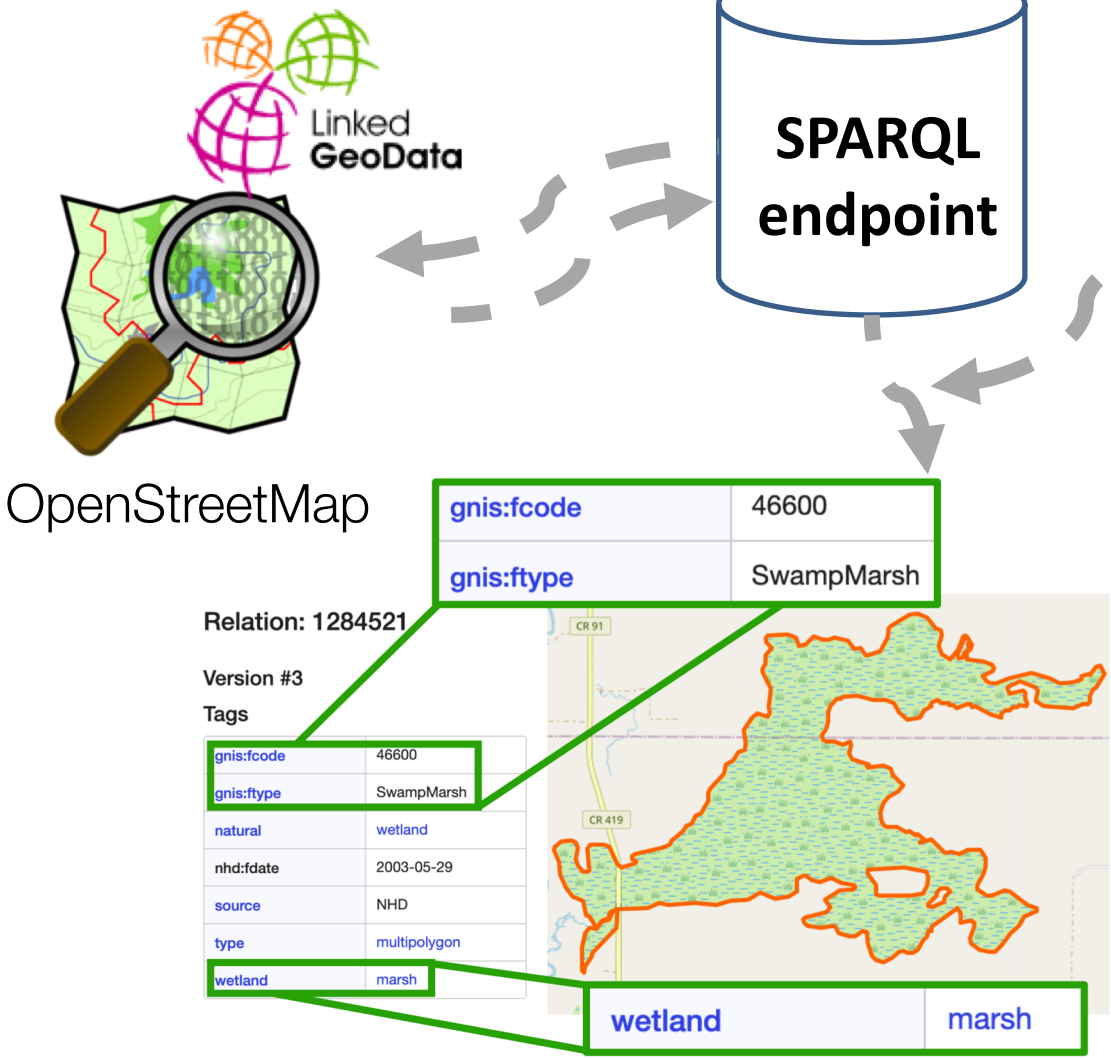
wetland blocks that are similar in 1961 and 2018

wetland blocks that are present in 1961 but are not present in 2018





Example 2


Linked GeoData
SPARQL endpoint
Can you provide additional attributes from a data subset?
OpenStreetMap
gnis:fcode 46600
gnis:ftype SwampMarsh
Relation: 1284521
Version #3
Tags
gnis:fcode 46600
gnis:ftype SwampMarsh
natural wetland
nhd:fdate 2003-05-29
source NHD
type multipolygon
wetland marsh
CR 91
CR 419
wetland marsh

Evaluation

- **Data** (vectorized historical maps)
  - Railroads:
    - Bray, CA (7)
    - Louisville, CO (4)
  - Wetlands:
    - Bieber, CA (4)
    - Palm Beach, FL (3)
    - Duncanville, TX (3)

- **Evaluation**:
  - Partitioning
    - Runtime
    - Number of nodes
  - geo EL
    - Runtime
    - Correctness
      - Precision, Recall & F1
  - RDF
    - Query time
    - Query complexity
    - Query robustness

Results: Partitioning

Partitioning statistics for **CA railroads**

Year	# vecs	Runtime (s)	# nodes
1954	2382	<1	1
1962	2322	36	5
1988	11134	1047	11
1984	11868	581	24
1950	11076	1332	43
2001	497	145	57
1958	1860	222	85

Partitioning statistics for **CA wetlands**

Year	# vecs	Runtime (s)	# nodes
1961	12	<1	1
1993	17	<1	5
1990	27	6	11
2018	9	6	24

Partitioning statistics for **FL wetlands**

Year	# vecs	Runtime (s)	# nodes
1987	184	<1	1
1956	531	180	5
2020	5322	1139	13

Partitioning statistics for **CO railroads**

Year	# vecs	Runtime (s)	# nodes
1965	838	<1	1
1950	418	8	5
1942	513	5	8
1957	353	4	10

Partitioning statistics for **TX wetlands**

Year	# vecs	Runtime (s)	# nodes
1959	8	<1	1
1995	6	<1	5
2020	1	1	10

Results: geo EL

Geo-entity linking results; Area is in square kilometers

		Area	Precision	Recall	F_1
Railroads	CA-baseline	420.39	0.193	1.000	0.323
	CA		0.800	0.750	**0.774**
	CO-baseline	132.01	0.455	1.000	0.625
	CO		0.833	1.000	**0.909**
Wetlands	CA-baseline	224.05	0.556	1.000	0.714
	CA		1.000	1.000	**1.000**
	FL-baseline	27493.98	0.263	1.000	**0.417**
	FL		0.758	0.272	0.400
	TX*	16.62	-	-	-

Results: RDF

```
SELECT ?f ?wkt WHERE {
  ?f a geo:Feature ;
    geo:hasGeometry [ geo:asWKT ?wkt ] ;
    dcterms:date 1962^^xsd:gYear ;
    dcterms:date 2001^^xsd:gYear .
  FILTER NOT EXISTS { ?f geo:sfContains _:_ } }
```

Fig. 15. Query similar feature geometries in both 1962 and 2001

```
SELECT ?f ?wkt WHERE {
  ?f a geo:Feature ;
    geo:hasGeometry [ geo:asWKT ?wkt ] ;
    dcterms:date 1962^^xsd:gYear .
  FILTER NOT EXISTS { ?f geo:sfContains _:_ }
  MINUS { ?f dcterms:date 2001^^xsd:gYear . } }
```

Fig. 16. Query feature geometries present in 1962 but not in 2001

```
SELECT ?f ?wkt WHERE {
  ?f a geo:Feature ;
    geo:hasGeometry [ geo:asWKT ?wkt ] ;
    dcterms:date 1958^^xsd:gYear .
  FILTER NOT EXISTS { ?f geo:sfContains _:_ }
  ?f dcterms:date ?date . }
GROUP BY ?f ?wkt
HAVING (COUNT(DISTINCT ?date) = 1)
```

Fig. 17. Query unique feature geometries from 1958

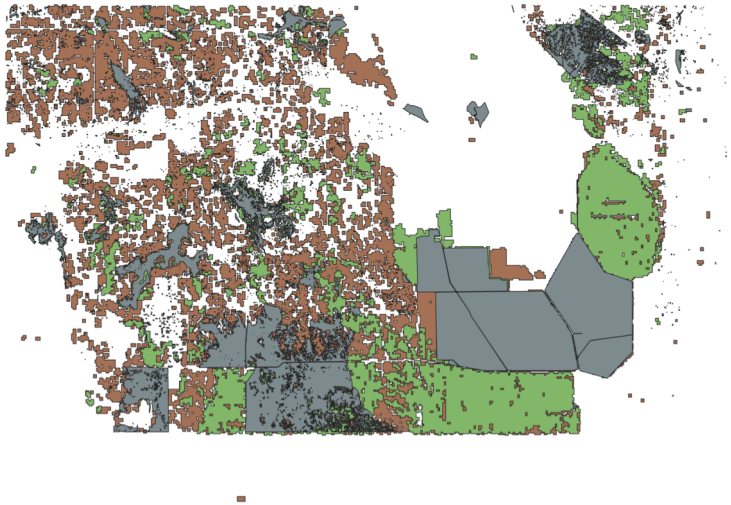
Results: RDF

Query time statistics (in milliseconds)

		avg	min	max
Railroads	SIM-CA	12	10	18
	SIM-CO	11	9	20
	DIFF-CA	10	8	20
	DIFF-CO	10	9	14
	UNIQ-CA	14	8	28
	UNIQ-CO	15	9	17
Wetlands	SIM-CA	22	18	34
	SIM-FL	35	18	55
	SIM-TX	21	12	44
	DIFF-CA	25	16	43
	DIFF-FL	32	18	60
	DIFF-TX	21	11	30
	UNIQ-CA	24	18	44
	UNIQ-FL	48	38	73
	UNIQ-TX	14	12	40

Discussion

- Complexity of changes in original topographic maps
- Quality & level of detail
- Crowdsourcing
  - OpenStreetMap

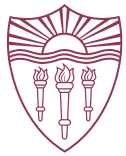


- How can we do better?
  - Partitioning:
    - Optimize buffer size hyperparameter (heuristics/learning)
    - Normalize & denoise the original data
    - Parallel processing
  - Geo EL:
    - Expand to additional KBs (Wikidata)

Related Work

- Transforming geospatial vector data into RDF
  - Kyzirakos 2014, Usery 2012
  - Do not address:
    - Geospatial entity inter-linking or intra-(distant) linking
    - Semantics
- Annotating geospatial data with external information
  - Vaisman 2019, Smeros 2016
  - Do not address:
    - linking unlabeled geographic “entities”
- Geospatial change analysis
  - Perez 2015, Kauppinen 2014
  - Do not address:
    - incremental process of change over time

Conclusions



- Unsupervised approach to integrate, relate, & interlink geospatial data from digitized resources
- Publishable structured semantic-rich linked spatio-temporal data
- Enables users to easily understand & analyze geographic information across time & space
- Fuel discovery

- Source code available at:

https://github.com/usc-isi-i2/linked-maps