Building Spatio-Temporal Knowledge Graphs from Vectorized Topographic Historical Maps

Oct 19th, 2022

Basel Shbita[1]

Craig A. Knoblock[1], Weiwei Duan[2],

Yao-Yi Chiang[2], Johannes H. Uhl[3], and Stefan Leyk[3]

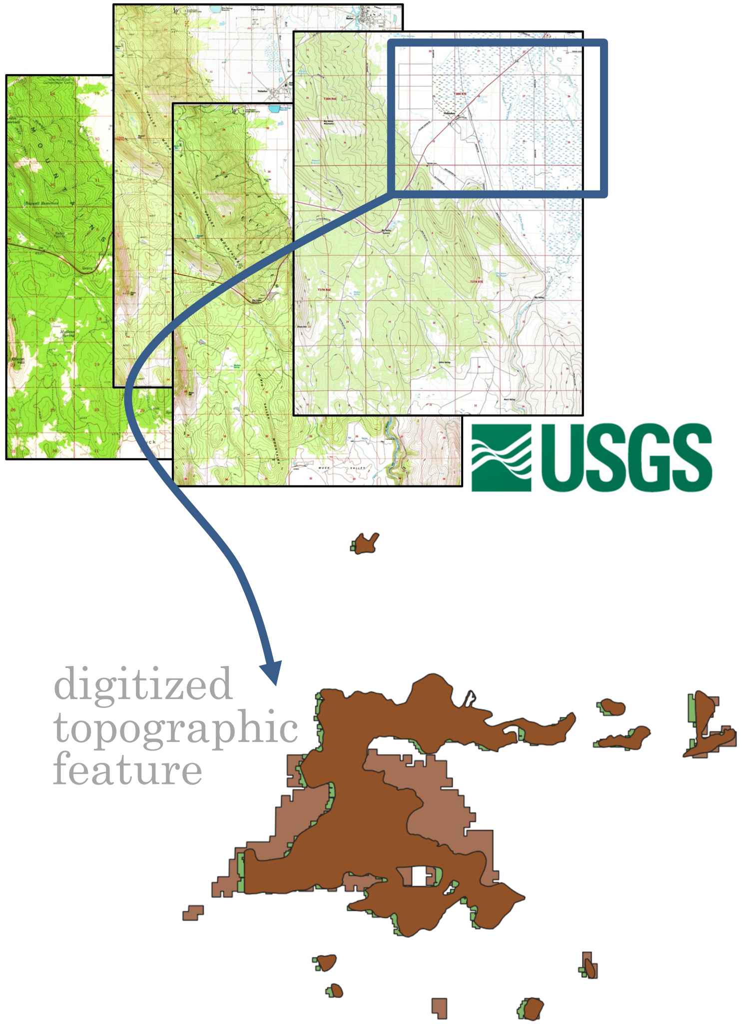
[1] Information Sciences Institute and Department of Computer Science, USC
[2] Spatial Sciences Institute and Department of Computer Science, USC
[3] Department of Geography, University of Colorado Boulder

Agenda

- Problem
- Approach
- Querying the Data
- Evaluation
- Discussion
- Related work
- Future work
- Conclusion

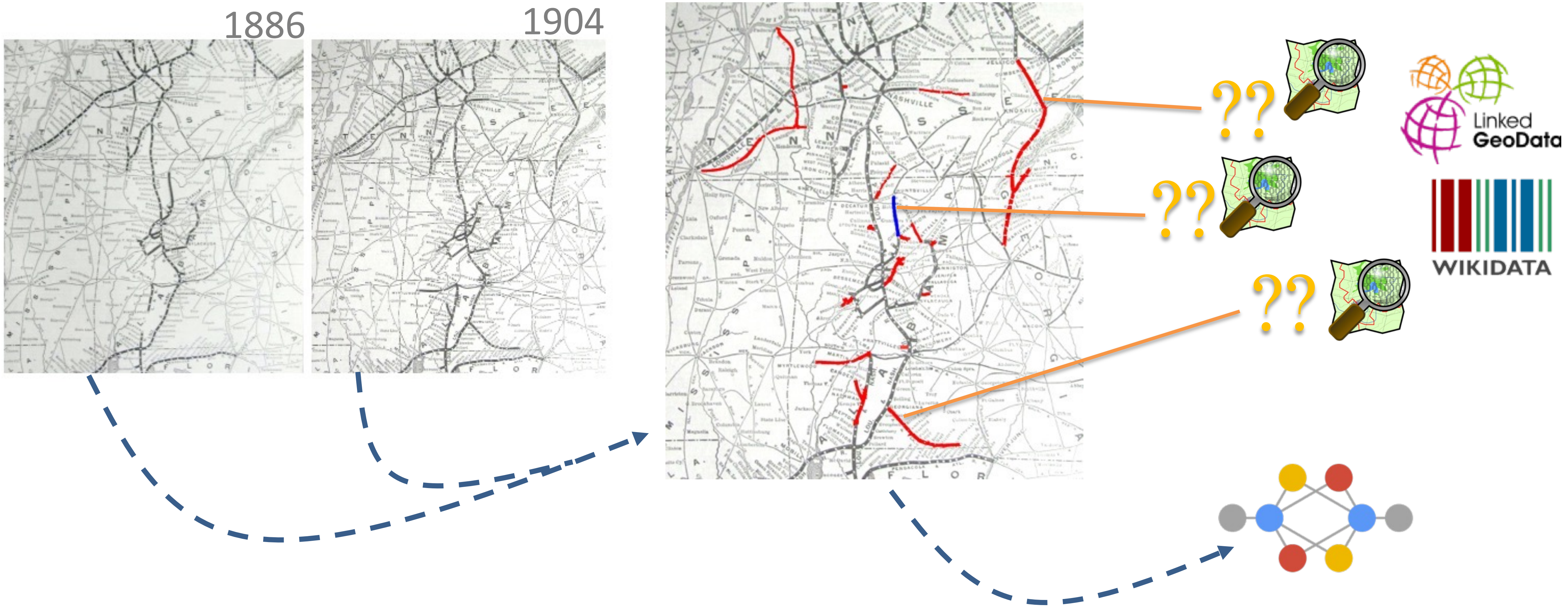
Intro – Geospatial Data

- Raster
- Vector
  - Compact way to represent real-world topographic features
    - Points (locations, addresses)
    - Lines (roads, rivers)
    - Polygons (waterbodies, islands)
  - Digitized **topographic historical maps**
    - Rich sources of information
      - Labor-intensive to analyze across time & space
        - » i.e., domain, format, sources, tools
      - Sometimes we need more contextual information
        - » i.e., geographic, demographic



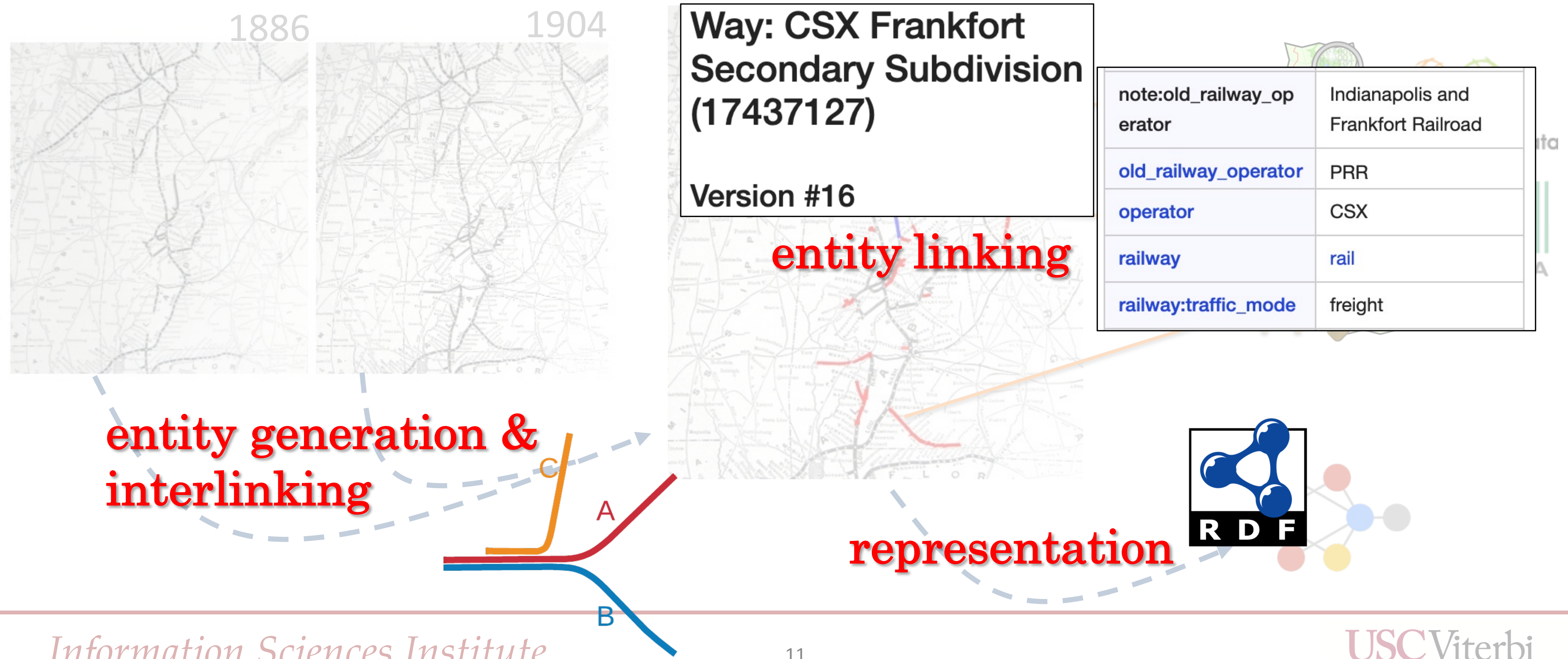
Problem Definition

How to transform & enrich **digitized** archival geo-data in an expressive & interoperable way to enable easy analysis over time & space?



Problem Definition

How to transform & enrich **digitized** archival geo-data in an expressive & interoperable way to enable easy analysis over time & space?



Agenda

- Problem
- Approach
  - Feature Partitioning
  - Geo-entity Linking
  - RDF Generation & Data Modeling
- Querying the data
- Evaluation
- Discussion
- Related work
- Future work
- Conclusion

Motivation

- Analyze
  - Railroad network change
  - Decline of wetlands
- Unsupervised & automatic
  - Answer complex queries re human & environmental systems
- Utilize & follow semantic web & linked open data
  - Make data widely available to researchers across domains
  - Structured & semantic
    - Easy query & visualization
    - Utilize available knowledge on the web

Motivation

- Knowledge on the web?

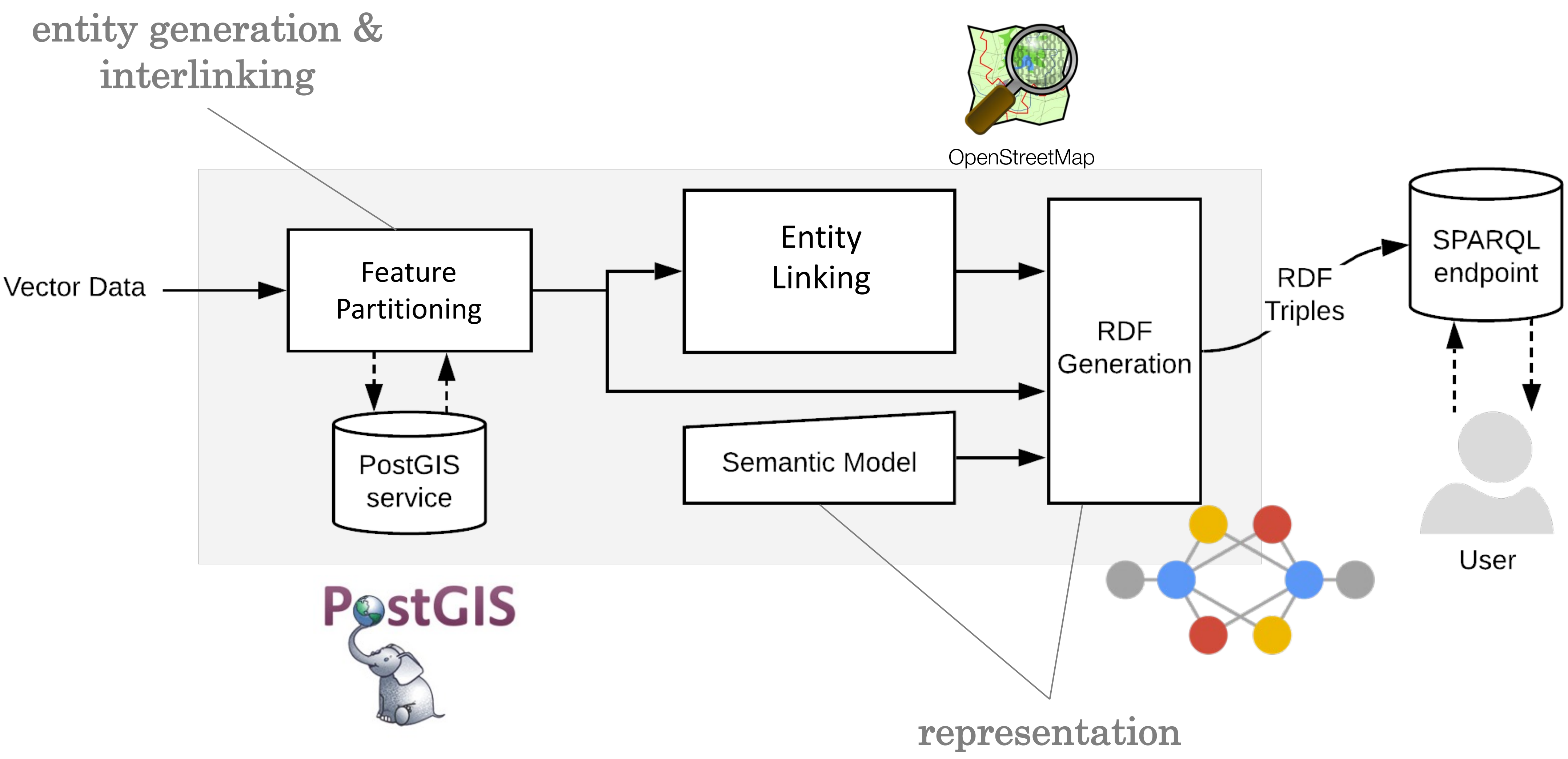
ele	172

Linked GeoData

population	8,765	
	point in time	2010

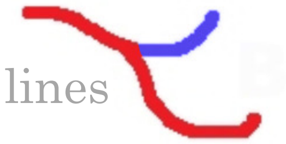
WIKIDATA

Our Approach

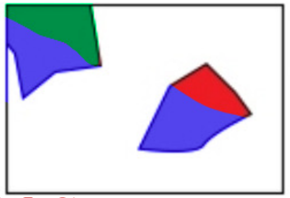


Feature Partitioning

- <u>Goal</u>: Generate building block geometries (i.e. geo entities) to represent the topographic features from different map sheets
  - Entity matching/linking & entity "partitioning" task
    - Represent common & distinct parts (changes) of the features
      - "granular building blocks"
      - e.g., railroad segments, wetland areas
    - Allow incremental additions over time



lines



areas

- <u>How</u>? Algorithm to create a DAG of building-block geometries (nodes) and their relations (edges)
  - Spatially-enabled DB service to manipulate & transform spatial data
    - buffer (hyperparameter)
  - Incremental

Feature Partitioning

vector data from various map editions (“initial” building blocks)

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

current “building blocks”

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

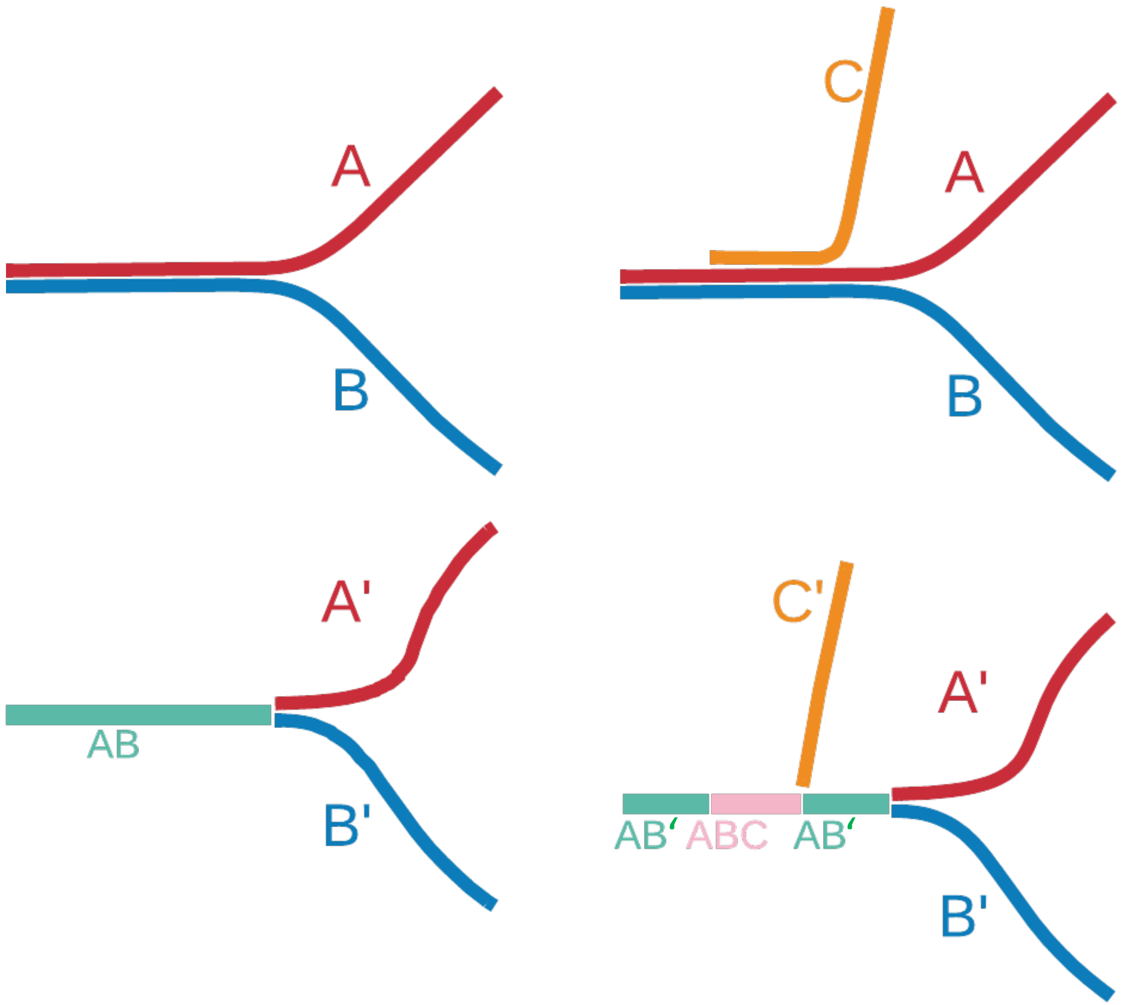
$$\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$$

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

end

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

end



Feature Partitioning

vector data from various map editions (“initial” building blocks)

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

current “building blocks”

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

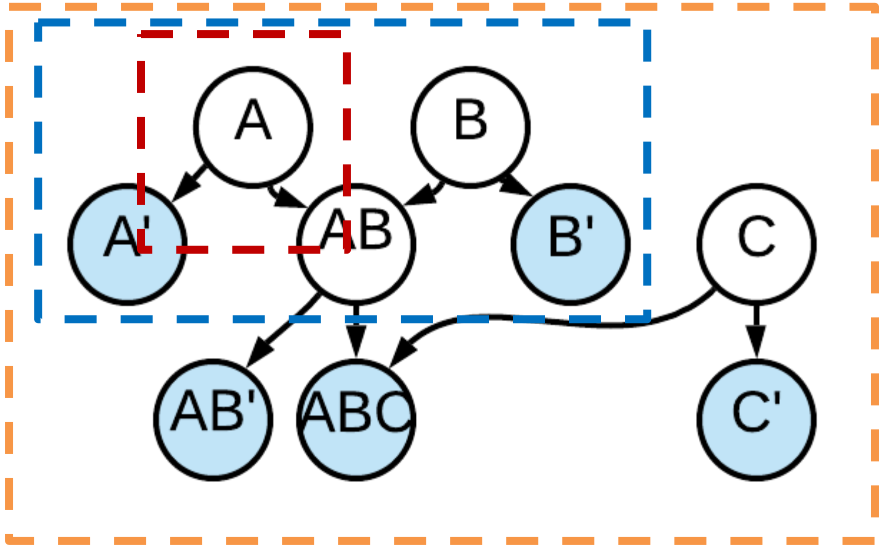
    $\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$

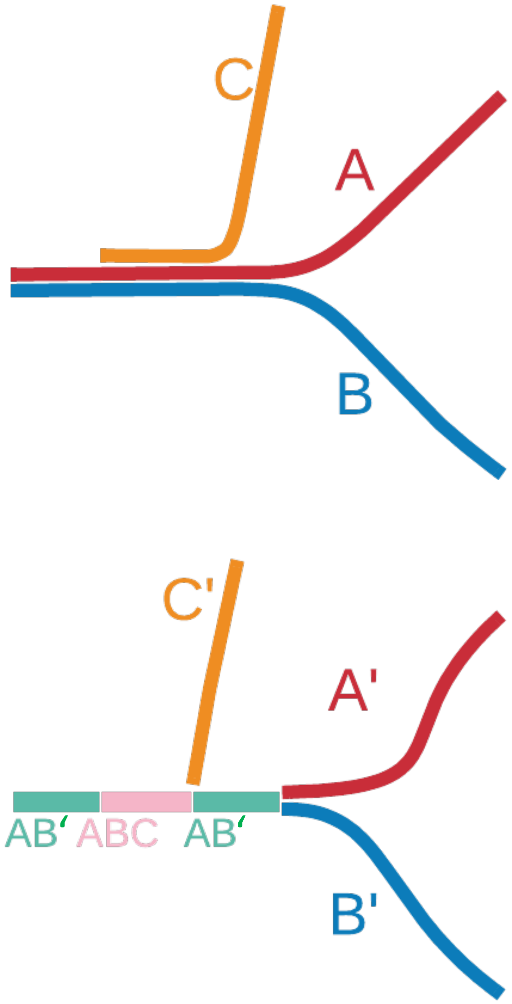
    $\mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha;$

  end

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

end





Feature Partitioning

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

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

current "building blocks"

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

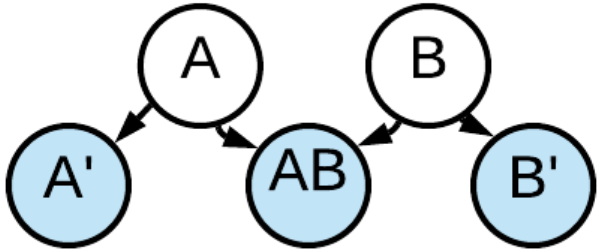
$$\mathcal{F}_\alpha = \mathcal{F}_i \bigcap \mathcal{F}_k;$$

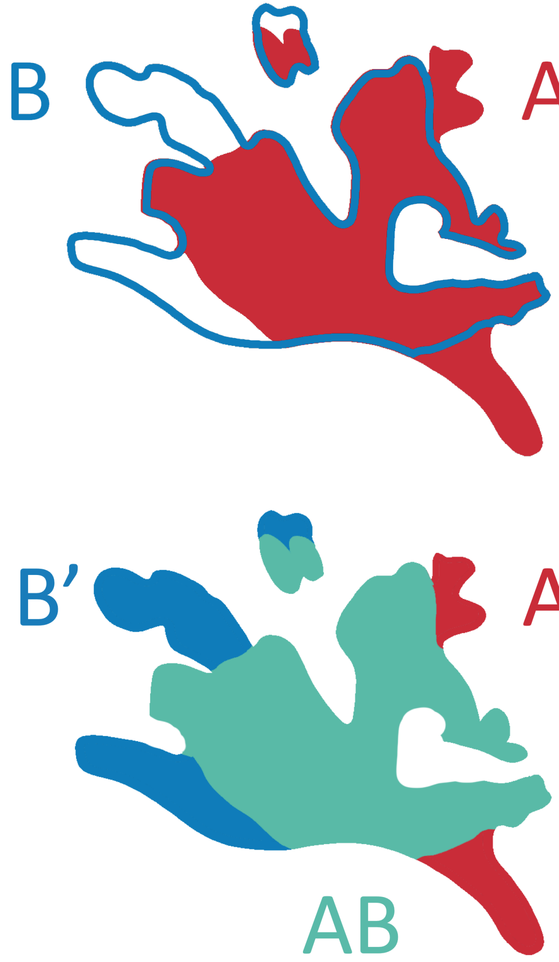
$$\mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha;$$

end

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

end





Geo-entity Linking

- Goal: Link the generated entities to open KBs
  - Entity matching with OSM (external entities)
    - Enrich to fuel discovery
      - geoNames, LinkedGeoData, Wikidata
      - Wikipedia, USGS GNIS

- How? Sampling
  - Reverse geocoding for initial filtering
    - OSM feature type is known (i.e., railroad)
  - Determine confidence by frequency of (random) samples that match
    - N (# samples, hyperparameter) determined by coverage area

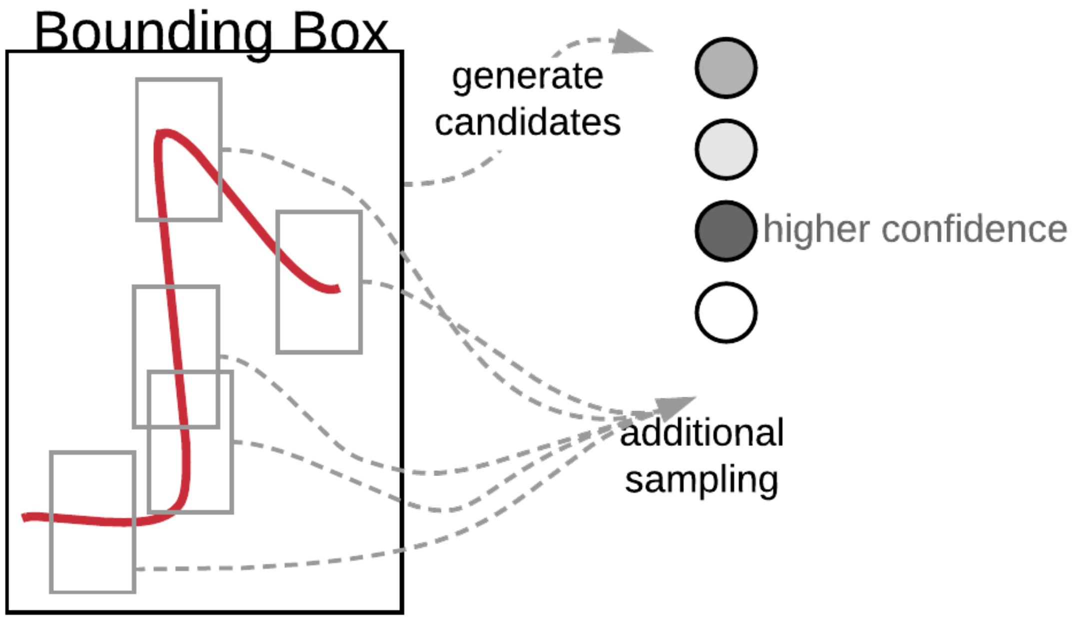
Geo-entity Linking

segment or area
"building block"

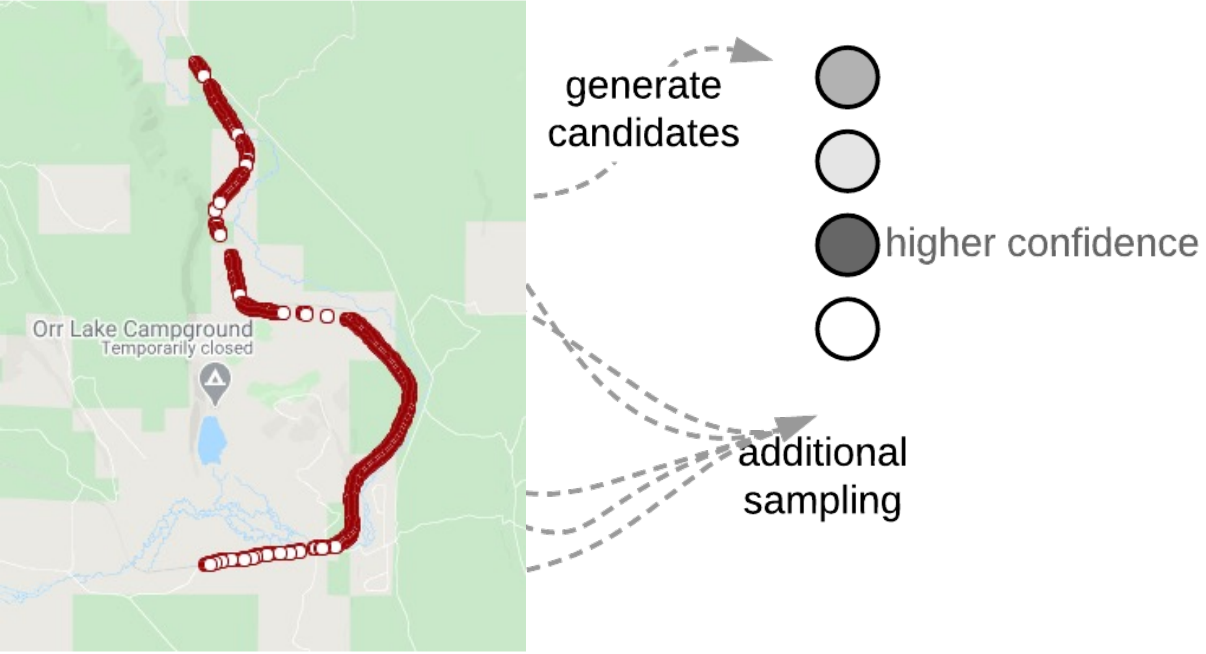
B_s = bounding box wrapping s;
$\mathcal{L}$ = reverse-geocoding(B_s, T);

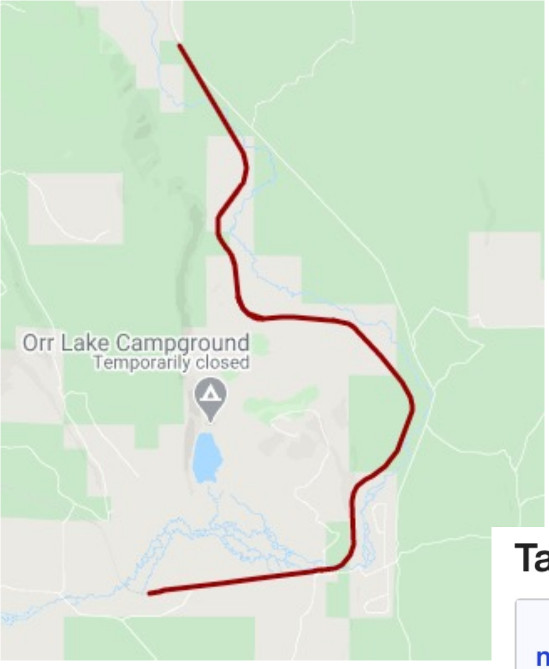
topographic feature type

for $1 \ldots N$ **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





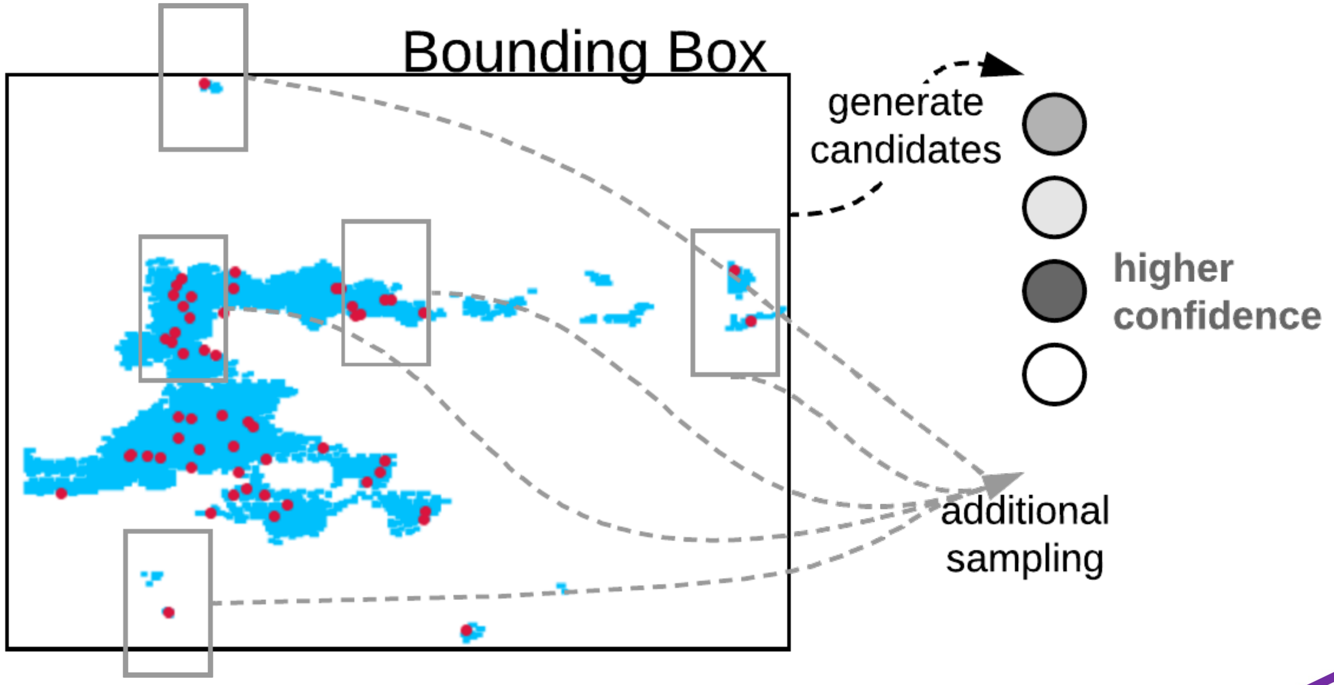
Way: Black Butte Subdivision (322131253)

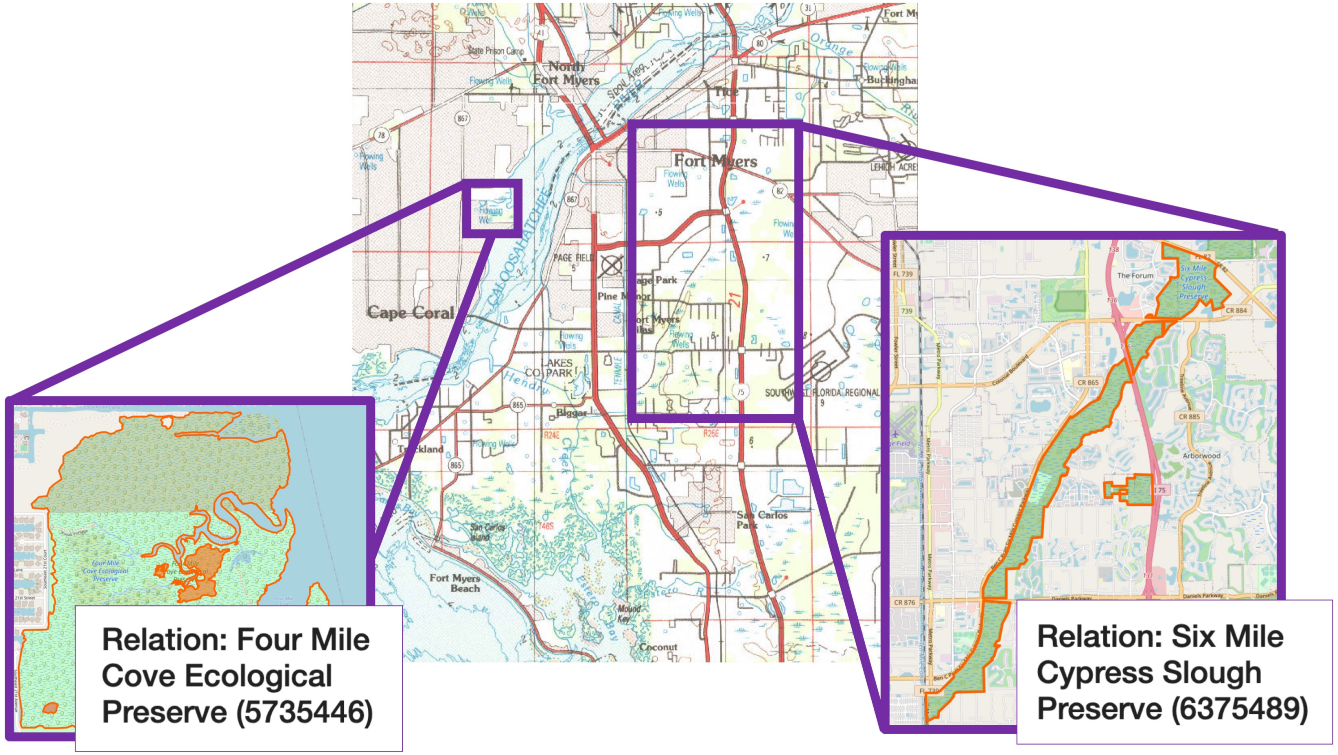
Version #6

Tags

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

Geo-entity Linking

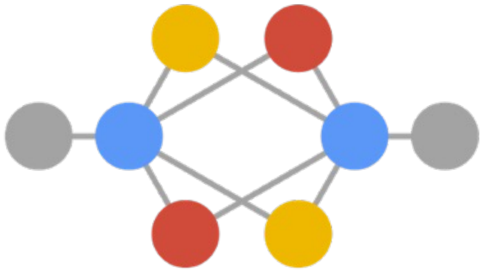




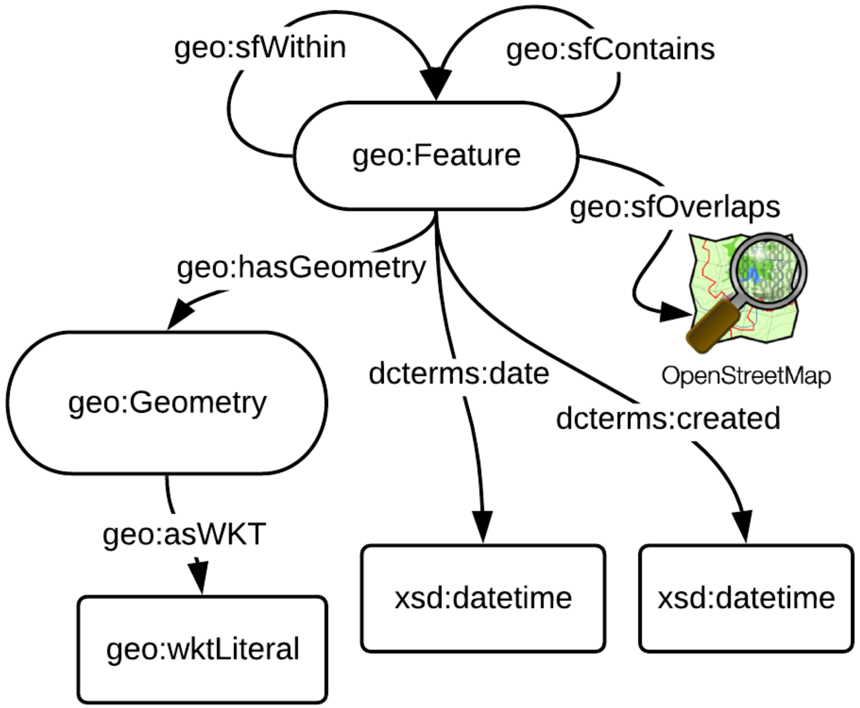
RDF Generation

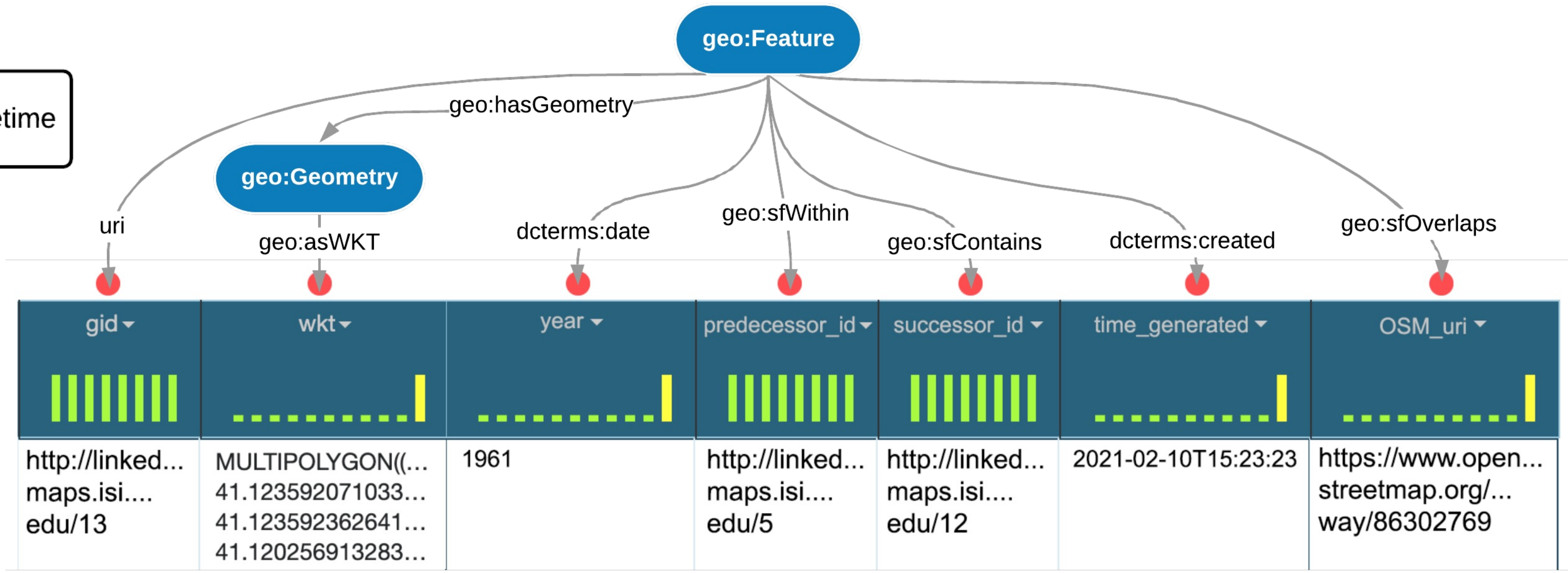
- <u>Goal</u>: Transform & publish the data (construct KG)
  - Semantic data modeling task
    - Follows linked data principles
    - Useful semantic representation
    - Support downstream tasks by accommodating
      - qualitative spatial reasoning systems
      - quantitative spatial computation systems

- <u>How</u>? Construct a meaningful semantic model
  - OGC GeoSPARQL standard
  - Universal conventions
  - Hierarchically-driven queries



Semantic Model & Data Modeling





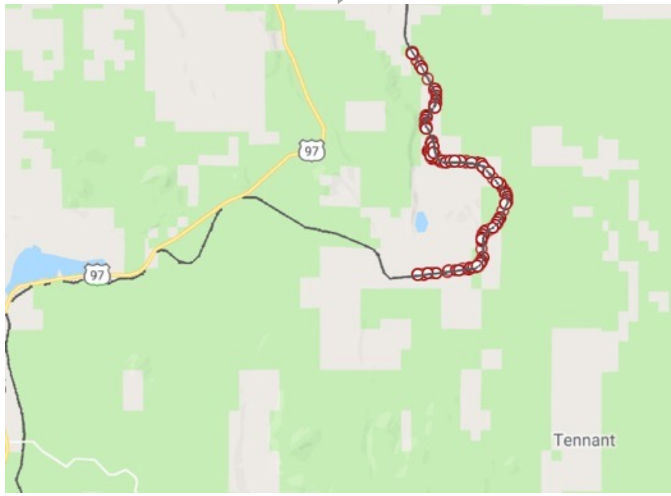
Agenda

- Problem
- Approach
- Querying the Data
  - ex. 1: Railroads
  - ex. 2: Wetlands
- Evaluation
- Discussion
- Related work
- Future work
- Conclusion

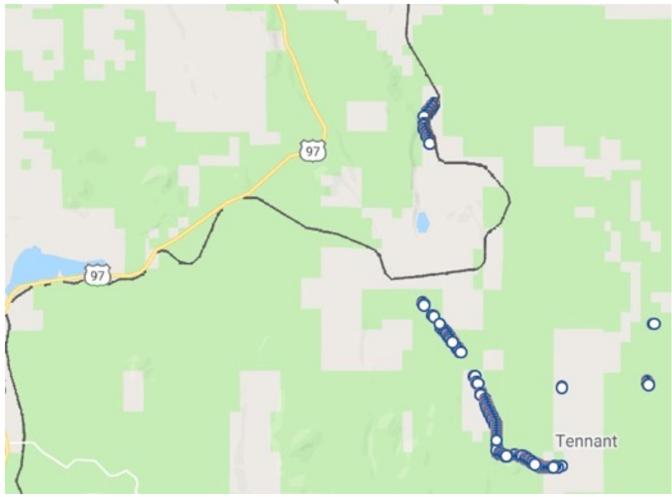
Ex. 1: Querying Railroads (CA)

SPARQL endpoint

Railroad segments that are similar in 1962 and 2001



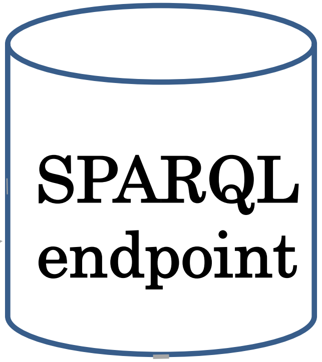
Railroad segments that are present in 1962 but are not present in 2001



Ex. 1: Querying Railroads (CA)

Linked GeoData

OpenStreetMap



Can you show me a subset that is abandoned?

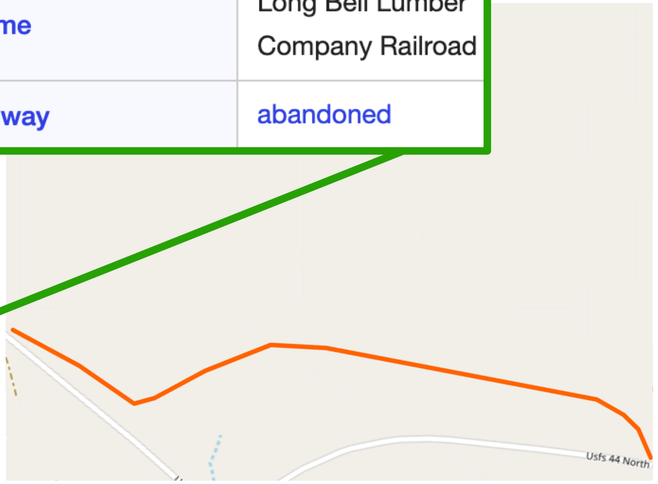
name	Long Bell Lumber Company Railroad
railway	abandoned

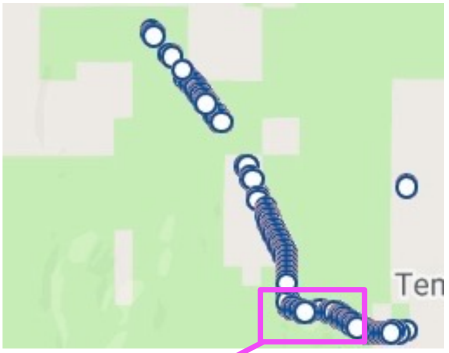
Way: Long Bell Lumber Company Railroad (177559134)

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



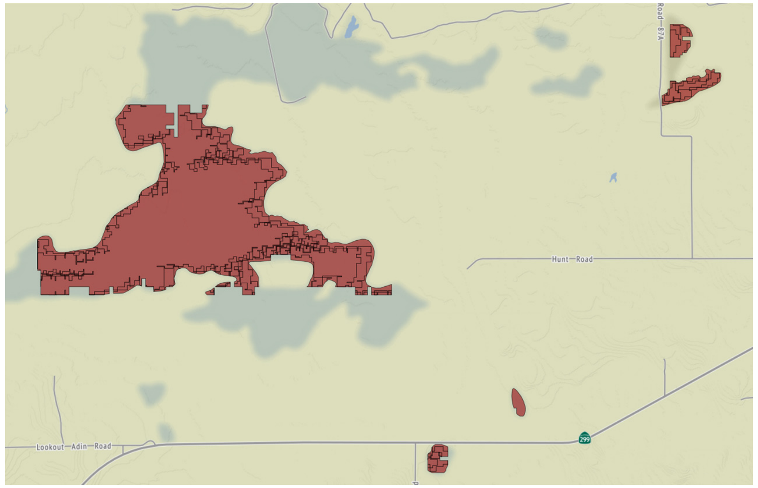


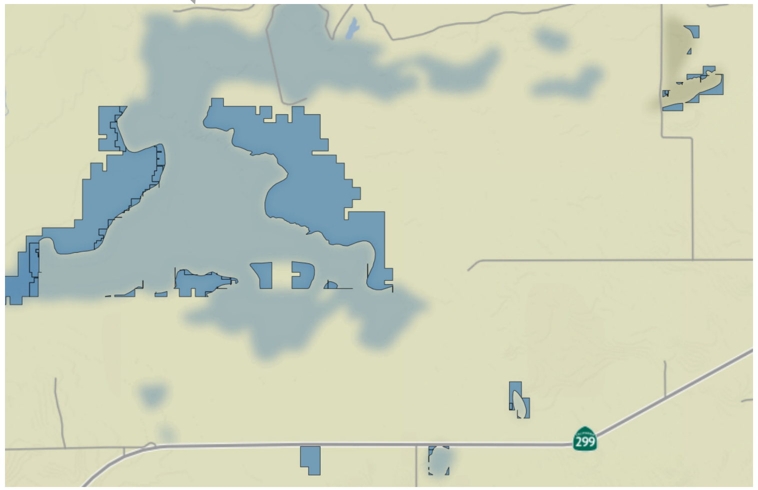
Ex. 2: Querying Wetlands (CA)

SPARQL endpoint

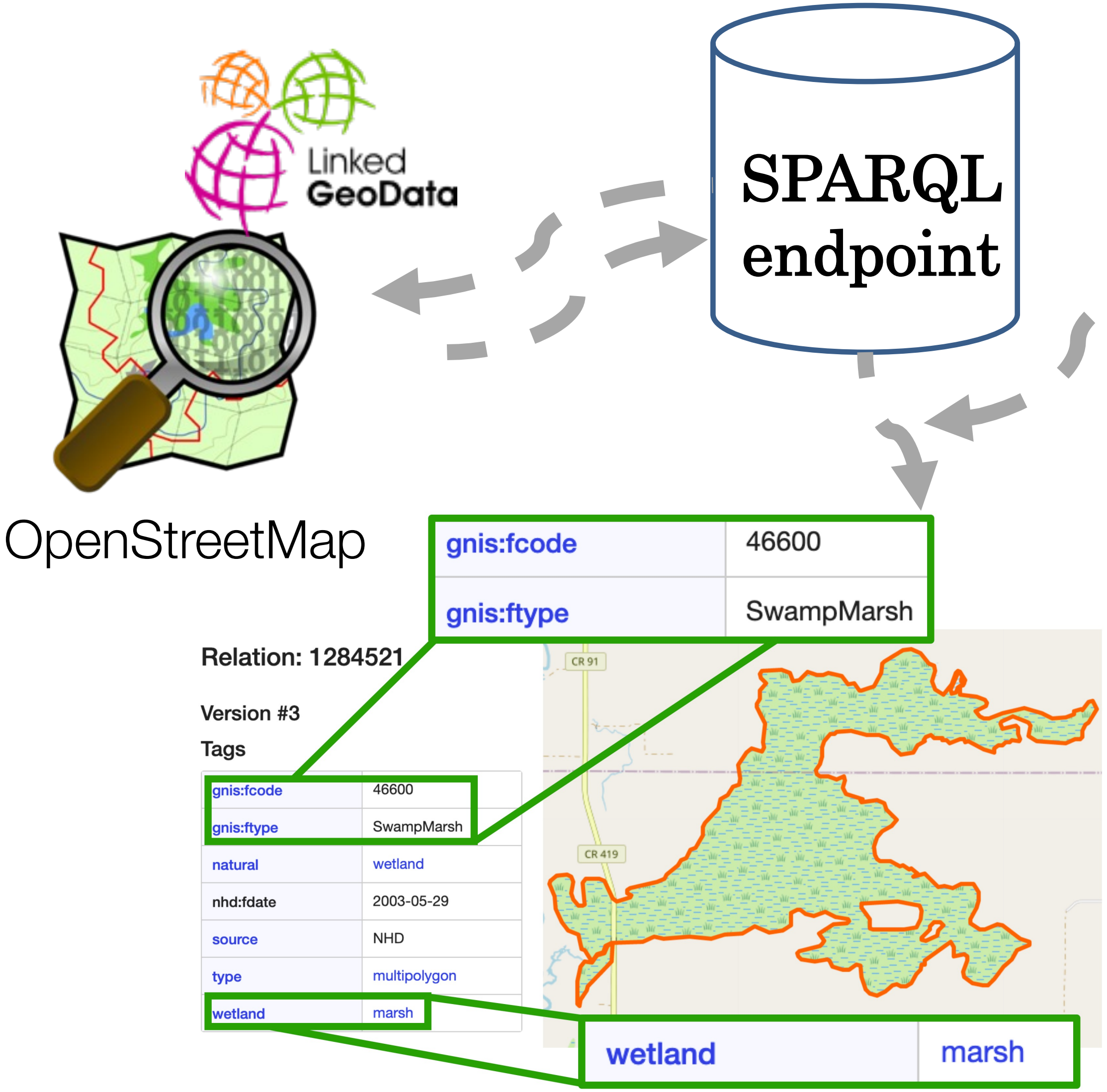
wetland blocks that are similar in 1961 and 2018

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





Ex. 2: Querying Wetlands (CA)



Can you provide additional attributes from that subset?

Agenda

- Problem
- Approach
- Querying the Data
- Evaluation
- Discussion
- Related work
- Future work
- Conclusion

Evaluation

- Feature Partitioning
  - Runtime
  - # of nodes
- Geo EL
  - Runtime
  - Correctness
    - Precision, Recall & F1
- RDF
  - Query complexity
  - Query time
  - Query robustness

Datasets:

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

Results: Feature 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 N = 20	0.193	1.000	0.323
	CA		0.800	0.750	**0.774**
	CO-baseline	132.01 N = 20	0.455	1.000	0.625
	CO		0.833	1.000	**0.909**
Wetlands	CA-baseline	224.05 N = 20	0.556	1.000	0.714
	CA		1.000	1.000	**1.000**
	FL-baseline	27493.98 N = 200	0.263	1.000	**0.417**
	FL		0.758	0.272	0.400
	TX*	16.62	-	-	-

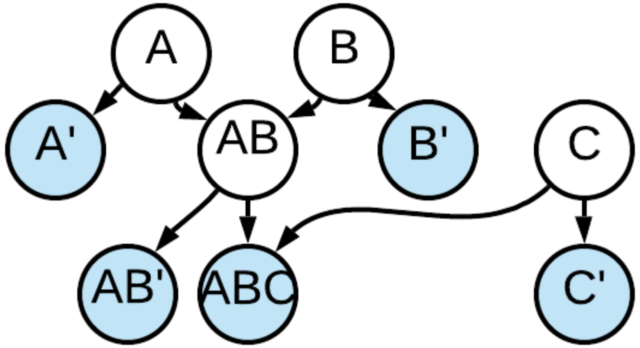
RDF: Query Complexity

```
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 Performance

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

Query time ranges **10-48 [ms]**

No significant change with respect to

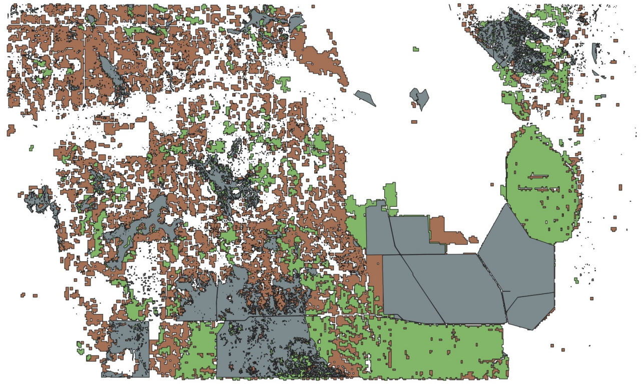
- # of map editions we process
- Complexity of the query we compose

Agenda

- Problem
- Approach
- Querying the Data
- Evaluation
- Discussion
- Related work
- Future work
- Conclusion

Discussion

- Feasible & effective in terms of processing time, completeness & robustness
  - Runs only once for newly added resources
  - Follows LD principles
    - Does not require re-generation of data
    - URIs are preserved



- Still, many challenges exist
  - Complexity of changes in original topographic maps
  - Quality & level of detail
  - Crowdsourcing: availability, granularity (e.g., mud vs. wetland)

Related work

- Transforming geospatial vector data into RDF (Kyzirakos 2014, Usery 2012)
  - Do not address:
    - Geo entity inter-linking or intra-(distant) linking
    - Semantics
- Contextualizing geospatial data (Vaisman 2019, Smeros 2016)
  - Do not address:
    - Linking unlabeled geo entities
- Geospatial change analysis (Perez 2015, Kauppinen 2014)
  - Do not address:
    - Incremental process of change over time

Future work

- How can we do better?
  - Feature Partitioning:
    - Optimize buffer size hyperparameter (heuristics/learning)
    - Normalize & denoise the original data
    - Parallel processing
  - Geo EL:
    - OSM bulk-read & joint matching (optimize hyperparameter N)
    - Expand to additional KBs (Wikidata, Yago2Geo)
    - Embed geometry features for type inference

Conclusion

- Unsupervised & automatic approach building contextualized spatio-temporal KGs from digitized topographic map archives
- Contributions
  - An incremental approach to compute topographic feature changes
  - A paradigm for geospatial data integration on the web
  - A hierarchy-driven semantic model for simple & efficient querying

- Source code & data available at:
  - https://github.com/usc-isi-i2/linked-maps

Thank you!