Building **Linked Spatio-Temporal Data** from Vectorized Historical Maps

April 21\textsuperscript{st}, 2020

**Basel Shbita\textsuperscript{1}, Craig A. Knoblock\textsuperscript{1}, Weiwei Duan\textsuperscript{2}, Yao-Yi Chiang\textsuperscript{2}, Johannes H. Uhl\textsuperscript{3}, and Stefan Leyk\textsuperscript{3}**

\textsuperscript{1}Information Sciences Institute and Department of Computer Science, USC
\textsuperscript{2}Spatial Sciences Institute and Department of Computer Science, USC
\textsuperscript{3}Department of Geography, University of Colorado Boulder
Problem

Digitized Historical Maps = Rich sources of information

Natural- & Human-made features
- Wetlands
- Railroad Networks

Labor-intensive to analyze across time & space

Additional discovery
Idea

Decompose to building blocks

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

Fuel Discovery

Make data widely available

Query & visualize
Our Approach

Shapefiles (vector data) → Feature (Line) Segmentation

PostGIS service

Reverse Geocoding (map to LinkedGeoData)

Semantic Model

Generate RDF

RDF Triples → SPARQL endpoint

User
Automatic Feature Segmentation

• **Goal**: create *partitions* of geo features (segments)
  – Entity matching/linking & entity “partitioning” task
  – “Building Blocks”

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

• Allow *incremental* additions over time
segments of different map edition

current “building blocks”

\[
\text{foreach } i \in M \text{ do}
\]

\[
\text{foreach } k \in L \text{ do}
\]

\[
\mathcal{F}_\alpha = \mathcal{F}_i \cap \mathcal{F}_k;
\]

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

\[
\text{end}
\]

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

\[
\text{end}
\]

\[
A \quad B
\]

\[
A' \quad B'
\]

\[
AB
\]
Feature Segmentation – cont’d

segments of different map edition

current “building blocks”

\[
\begin{align*}
\text{foreach } i \in \mathcal{M} \text{ do} \\
\quad & \text{foreach } k \in \mathcal{L} \text{ do} \\
\quad & \quad \mathcal{F}_\alpha = \mathcal{F}_i \cap \mathcal{F}_k; \\
\quad & \quad \mathcal{F}_\gamma = \mathcal{F}_k \setminus \mathcal{F}_\alpha; \\
\quad & \text{end} \\
\quad & \mathcal{F}_\delta = \mathcal{F}_i \setminus \left( \bigcup_{j \in \mathcal{L}} F_j \right); \\
\text{end}
\end{align*}
\]
Geo-linking

- **Goal**: map segments to **Linked Open Vocabularies**
  - Entity matching
  - Enrich data to fuel **discovery**

- Use a **reverse geocoding** service (OpenStreetMap)
  - **LinkedGeoData** instances
Geo-linking – cont’d

\[ B_s = \text{bounding box wrapping } s; \]
\[ \mathcal{L} = \text{reverse-geocoding}(B_s, T); \]
\[ \text{for } 1 \ldots N \text{ do} \]
\[ e = \text{randomly sample a Point in segment } s; \]
\[ E = \text{reverse-geocoding}(e, T); \]
\[ \mathcal{L}.\text{add}(E); \]
\[ \text{end} \]

filter out instances with a single appearance in \( \mathcal{L} \);
return \( \mathcal{L} \);
Geo-linking – cont’d

Bounding Box

- Generate candidates
- Higher confidence
- Additional sampling
Geo-linking – cont’d

Bounding Box

generate candidates

additional sampling

Long Bell Lumber Company Railroad at LinkedGeoData
http://linkedgeo.data.org/triplify/way177559134
• **Goal**: construct a **KG** from the data we collected
  – Useful **semantic representation**
  – Support downstream **spatial reasoners**

• Construct a meaningful **semantic model**
  – OGC GeoSPARQL **standard**
  – **Universal** conventions
  – Easily queried
We constructed a KG

Now what?
Use Case

SPARQL endpoint

Railroad segments that are similar in 1962 and 2001

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

Can you show me a subset of what’s abandoned?
Evaluation

**Railroad data** from a collection of historical maps:

- Bray, California (7)
- Louisville, Colorado (4)

- **Segmentation**
  - Runtime
  - Number of nodes

- **Geo-linking**
  - Runtime
  - Correctness (Precision, Recall & F1)

- **RDF**
  - Query time
  - Query complexity
  - Query robustness
Results

- Segmentation

**Table 1. Segmentation Statistics for Bray**

| 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      |

**Table 2. Segmentation Statistics for Louisville**

| Year | # vecs | Runtime (s) | # nodes |
|------|--------|-------------|---------|
| 1965 | 838    | <1          | 1       |
| 1950 | 418    | 8           | 5       |
| 1942 | 513    | 5           | 8       |
| 1957 | 353    | 4           | 10      |
Results – cont’d

- Geo-linking

### Table 3. “Geo-linking” Results

|               | Precision | Recall | F1    |
|---------------|-----------|--------|-------|
| BRA-baseline  | 0.193     | 1.000  | 0.323 |
| BRA           | 0.800     | 0.750  | 0.774 |
| LOU-baseline  | 0.455     | 1.000  | 0.625 |
| LOU           | 0.833     | 1.000  | 0.909 |
Results – cont’d

• RDF

```sql
SELECT ?f ?wkt WHERE {
  ?f a geo:Feature ;
  geo:hasGeometry [ geo:asWKT ?wkt ] ;
  dcterms:date "1962-01-01T00:00:00"^^xsd:dateTime ;
  dcterms:date "2001-01-01T00:00:00"^^xsd:dateTime .
  FILTER NOT EXISTS { ?f geo:sfContains _: _ } }
```

**Table 4. Query Time Statistics (in milliseconds)**

|       | avg | min | max |
|-------|-----|-----|-----|
| SIM-BRA | 12  | 10  | 18  |
| SIM-LOU | 11  | 9   | 20  |
| DIFF-BRA | 10  | 8   | 20  |
| DIFF-LOU | 10  | 9   | 14  |
| UNIQ-BRA | 14  | 8   | 28  |
| UNIQ-LOU | 15  | 9   | 17  |
Discussion

• Complexity of changes in original topographic maps
• Quality & level of detail
• Crowdsourcing
  – LinkedGeoData

• How can we do better?
  – Segmentation:
    • Optimize buffer size hyperparameter (heuristics/learning)
    • Normalize & denoise the original data
    • Parallel processing
  – Geo-linking:
    • Expand to additional KBs (Wikidata)
Related Work

• Transforming geospatial vector data into RDF
  – Kyzirakos et al. [1], Usery et al. [2]
  – Do not address:
    • Geospatial entity intra-linking or distant linking
    • Semantics

• Geographical data conflation
  – Li et al. [3], Ruiz et al. [4]
  – Do not address:
    • Linked Data or Semantics

• Geospatial data integration in the web
  – Prudhomme et al. [5]
  – Do not address:
    • Geospatial entity intra-linking or distant linking
Future (present) Work

- Extend
  - Wetlands
  - Forests
  - Highways
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-is-i2/linked-maps
References

[1] Kyzirakos, K., Vlachopoulos, I., Savva, D., Manegold, S., Koubarakis, M.: Geotriples: a tool for publishing geospatial data as rdf graphs using r2rml map- pings (2014)

[2] Usery, E.L., Varanka, D.: Design and development of linked data from the national map (2012)

[3] Li, L., Goodchild, M.F.: An optimisation model for linear feature matching in geographical data conflation (2011)

[4] Ruiz, J.J., Ariza, F.J., Urena, M.A., Blázquez, E.B.: Digital map conflation: a review of the process and a proposal for classification (2011)

[5] Prudhomme, C., Homburg, T., Ponciano, J.J., Boochs, F., Cruz, C., Roxin, A.M.: Interpretation and automatic integration of geospatial data into the semantic web (2019)