Building Spatio-Temporal Knowledge Graphs from Vectorized Topographic Historical Maps

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

\textsuperscript{a} Information Sciences Institute, University of Southern California, CA, USA
E-mails: shbita@isi.edu, knoblock@isi.edu

\textsuperscript{b} Spatial Sciences Institute, University of Southern California, CA, USA
E-mails: weiweidu@usc.edu, yaoyic@usc.edu

\textsuperscript{c} Department of Geography, University of Colorado Boulder, CO, USA
E-mails: johannes.uhl@colorado.edu, stefan.leyk@colorado.edu

Abstract. Historical maps provide rich information for researchers in many areas, including the social and natural sciences. These maps contain detailed documentation of a wide variety of natural and human-made features and their changes over time, such as changes in transportation networks or the decline of wetlands or forest areas. Analyzing changes over time in such maps can be labor-intensive for a scientist, even after the geographic features have been digitized and converted to a vector format. Knowledge Graphs (KGs) are the appropriate representations to store and link such data and support semantic and temporal querying to facilitate change analysis. KGs combine expressivity, interoperability, and standardization in the Semantic Web stack, thus providing a strong foundation for querying and analysis. In this paper, we present an automatic approach to convert vector geographic features extracted from multiple historical maps into contextualized spatio-temporal KGs. The resulting graphs can be easily queried and visualized to understand the changes in different regions over time. We evaluate our technique on railroad networks and wetland areas extracted from the United States Geological Survey (USGS) historical topographic maps for several regions over multiple map sheets and editions. We also demonstrate how the automatically constructed linked data (i.e., KGs) enable effective querying and visualization of changes over different points in time.

Keywords: Spatio-Temporal Knowledge Graphs, Knowledge Graphs, Linked Spatio-Temporal Data, Linked Data, Semantic Web, Data Integration, Historical Maps, Digital Humanities

1. Introduction

Historical map archives contain valuable geographic information on both natural and human-made features across time and space \cite{1, 2} and are increasingly available in systematically acquired digital data formats \cite{3, 4}. The spatio-temporal data extracted from these documents are important since they can convey when, where and what changes took place \cite{5}. For example, this type of data enables the detection of long-term changes in railroad networks or studying the evolution of wetlands within the same region over time and thus can support decision-making related to the development of transportation infrastructure or questions related to land conservation, landscape ecology, or long-term land development and human settlement. Many applications assessing geographic changes over time typically involve searching, discovering, and manually identifying, digitizing, and integrating relevant data. This is a difficult and laborious task that requires domain knowledge and familiarity with various data sources, data formats and working environments, and the task is often susceptible to human error \cite{6}.

Generally, there are two types of geospatial data, namely, raster data and vector data. Recent technolog-
Previous work on creating linked data from historical geospatial information has focused on the problem of transforming a single instance of a map sheet or a formatted geospatial data source into Resource Description Framework (RDF) graphs [14–16]. However, this line of work does not address the issue of entity interlinking that is essential for building linked geospatial data for the task of change analysis with a semantic relationship between geospatial entities across map editions of the same region, which could be part of a large collection of topographic map sheets. Similar work is concerned with only a specific type of geometry, such as points, as in [12, 17], or is limited to a specific physical feature (i.e., flooded areas [18] or wildfires [19]). Our work does not impose such limitations.

Our approach is not only helpful in making the RDF data widely available to researchers but also enables users to easily answer complex queries, such as investigating the interrelationships between human and environmental systems. Our approach also benefits from the open and connective nature of linked data. Compared to existing tools such as PostGIS\(^1\), which can only handle queries related to geospatial entities and relationships within local databases, linked data can utilize many widely available knowledge sources in the Semantic Web, such as OpenStreetMap (OSM) [20], GeoNames [21], and LinkedGeoData [22], to enable rich semantic queries.

Providing contextual knowledge can help explain or reveal interactions of topographic changes to further spatio-temporal processes. For example, the external-linking enables augmentation of the contained geographic information with data from external KBs, such as historical population estimates. Once we convert the map data into linked data, we can execute SPARQL queries to identify the changes in map features over time and thus accelerate and improve spatio-temporal analysis tasks. Using a semantic representation that includes geospatial attributes, we can support researchers to query and visualize changes in maps over time and allow the development of data analytics applications that could have great implications for environmental, economic, or societal purposes.

This paper is based on our earlier conference and workshop papers [23, 24]. The paper expands our methods to support polygon-based geographic features, elaborates on the details of the algorithms, and provides a more extensive evaluation of the methods. In addition, we use OpenStreetMap instead of LinkedGeoData as an external KB to obtain the most up-to-date information and provide a solution that does not depend on time-sensitive third-party data dumps.

### 1.1. Problem definition

The task we address here is as follows: Given geographic vector data extracted from multiple map edi-

---

1\https://postgis.net/
ties of the same region, we aim to automatically construct a knowledge graph depicting all the geographic features that represent the original data, their relations (interlinks), and their semantics across different points in time. Using the constructed knowledge graph, we enable tackling more specific downstream analysis tasks. These may include the visualization of feature changes over time and the exploration of supplementary information (e.g., population data, elevation data, etc.) related to the region originating from an external knowledge base.

As an example, consider the maps in Figure 1 where changes in the New Albany (OH) and Chicago (IL) railroad system have occurred between 1886 and 1904. Figure 2 shows the railroad line changes between the different map editions. Line segments that have been added are marked in red and line segments that have been removed are marked in blue. Assuming we have the data available as vector data (which can be generated from scanned maps using approaches such as in Duan et al. [7]), our task in such a setting would be to construct a knowledge graph describing the shared line segments that are shared across these maps or unique in individual maps with a conventional semantic representation for the railroad line segments in each map edition. This would include objects from a list of common geographic features (points, lines, or polygons), their geolocational details, and their temporal coverage to allow easy analysis and visualization.

### 1.2. Contribution

The overall contribution of this paper is a fully automatic end-to-end approach for building a contextualized spatio-temporal knowledge graph from a set of vectorized geographic features extracted from topographic historical maps. We tackle the core challenges we mentioned earlier by presenting:

1. An algorithm to identify and partition the original vector data into interlinked geospatial entities (i.e., “building block” geometries) that constitute the desired geographic features across map editions (entity generation and interlinking task)

2. A method to identify and retrieve geospatial entities from a publicly available knowledge base (external geo-entity linking task)

3. A semantic model to describe the resulting spatio-temporal data in a structured and semantic output that can be easily interpreted by humans and machines in a form of a knowledge graph that adheres to linked data principles (representation task)

We also present a thorough evaluation for each of the above and apply our method to five relevant datasets that span two types of geographic features: railroads (line-based geometry) and wetlands (polygon-based geometry), each resulting in an integrated knowledge graph. Finally, we make the source code, the original datasets, and the resulting data publicly available. We have also published the resulting knowledge graph as linked data, along with a designated SPARQL endpoint with compliant IRI dereferencing. The resulting knowledge graph characteristics are described in Table 1.

### Table 1

<table>
<thead>
<tr>
<th>SPARQL endpoint</th>
<th><a href="https://linked-maps.isi.edu/sparql">https://linked-maps.isi.edu/sparql</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>Linked Data explorer</td>
<td><a href="https://linked-maps.isi.edu">https://linked-maps.isi.edu</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 classes</td>
</tr>
<tr>
<td>8 relations</td>
</tr>
<tr>
<td>630 nodes (340 non literals)</td>
</tr>
<tr>
<td>1514 edges</td>
</tr>
</tbody>
</table>

---

2 https://github.com/usc-isi-i2/linked-maps
1.3. Structure of paper

The rest of the paper is organized as follows. Section 2 presents our proposed approach and methods. Section 3 presents an evaluation of our approach by automatically building a knowledge graph for (i) a series of railroad networks from historical maps covering two different regions from different points in time, and (ii) a series of wetland data from maps covering three different regions from different periods. Section 4 reviews the related work. Section 5 concludes, discusses, and presents future work.

2. Approach to Building Spatio-temporal Knowledge Graphs

2.1. Preliminaries

Before we discuss our proposed method, it is essential to define certain preliminaries and the terminology used in this paper. A geographic feature refers to a collection of geometries (points, lines, or polygons), which together represent an entity or phenomenon on Earth. In this paper, a geographic feature is represented using a collection of “building block” geometries. Given a set of geographic features pertaining to the same region from different points in time, each generated “building block” is essentially the largest geographic feature part that is either shared across different points in time or unique for a specific point in time. These blocks can be either lines or areas in the case of linear geographic features or polygon geographic features, respectively. For example, in Figure 4a, A and B are line building blocks. Each building block may be decomposed into smaller building blocks. For example, if a part of A and B represents the same entity, two new building blocks are created: A’ and B’. Similarly, AB represents the shared geometry detected from A and B, and A’ and B’ represent the unique parts in A and B, respectively. Similarly, each color area in Figure 5 (A’ in red, B’ in blue, and AB in green) is an area building block (given a total of three building blocks). Each building block geometry is encoded as well-known text (WKT) representation (MULTILINE or MULTIPOLYGON textual format) and corresponds to a geospatial entity in our resulting KG.

2.2. Overview of the approach

The proposed pipeline for the construction of the knowledge graph consists of several major steps as illustrated in Figure 3. These steps can be summarized as follows:

1. Automatically partition the input geographic feature originating from the vector data into building block geometries (i.e., geospatial entities) using a spatially-enabled database service (e.g., PostGIS) (see Section 2.3).
2. Perform external entity linking by utilizing a reverse-geocoding service to map the geospatial entities to existing instances in an open knowledge base (e.g., OpenStreetMap) (see Section 2.4).
3. Construct the knowledge graph by generating RDF triples following a pre-defined semantic model using the data we generated in the previous steps (see Sections 2.5 and 2.6).

Once the RDF data is deployed, users can easily interact with the building block geometries (geospatial entities), the geographic features and metadata to perform queries (Section 2.7). These allow end-users to visualize the data and support the development of spatio-temporal downstream applications.

2.3. Generating building blocks and interlinking

The first task in our pipeline is the generation of building block geometries that can represent the var-
A and B have common and distinct parts

Buffer out and find the common parts

Partition line segments

Line segment C is added

Final line partitioning

Algorithm 1: The feature partitioning and interlinking algorithm

**Input:** a set $M$ of feature geometries for different map editions of the same region (vector data)

**Output:** a directed acyclic graph $G$ of building block geometries (nodes) and their relations (edges)

1. **foreach** $i \in M$ **do**
   2. $F_i = \text{set of geometries in } i$;
   3. $G.add(i \mapsto F_i)$;
   4. $L = \text{list of current leaf nodes in } G$;
   5. **foreach** $k \in L$ **do**
      6. $F_k = \text{set of geometries in } k$;
      7. $F_i = F_k \cap F_i$;
      8. $G.add(i \mapsto F_i)$; set $i,k$ as direct predecessors of $\alpha$;
      9. $F_i = F_i \setminus F_k$;
      10. $G.add(\gamma \mapsto F_k)$; set $k$ as direct predecessor of $\gamma$;
      11. $F_k = F_k \setminus (\bigcup \{F_j \mid j \in L \})$;
      12. $G.add(\delta \mapsto F_k)$; set $i$ as direct predecessor of $\delta$;

transform geographic objects in databases. To handle our task efficiently and enable an incremental addition of map sheets over time, we implemented Algorithm 1. The algorithm performs the partitioning task by employing several PostGIS Application Programming Interface (API) calls over the geometries of our lines or polygons in the database. In the case of line geometries, we buffer each line segment to create two-dimensional areas before applying any geospatial operation described below.

In detail, the procedure works as follows. The **for** loop in line 1 iterates over each of the map editions to extract the feature geometry (as seen in line 2 and stored in $F_i$) to create the initial “building block” geometry (line 3, denoted as node $i$ and added to graph $G$, which eventually will hold our final set of building blocks and record their relations in a data structure). Line 4 retrieves the leaf nodes from graph $G$ to list $L$. 

Fig. 4. Illustration of the geometry partitioning to building blocks for a line geometry: spatial buffers are used to identify the same line segments considering potential positional offsets of the data.

Fig. 5. Illustration of a geometry partitioning result for a polygon geometry: each color represents a different building block. A single building block may contain disconnected areas.
In the first iteration list \( L \) is empty. In the next iterations it will include “leaf” nodes. These are nodes that represent the most fine-grained building blocks computed so far. \( A \) and \( B \) in Figure 4a correspond to \( k \) and \( i \) respectively (in iteration 2 of the algorithm when executed over the data in the Figure 4). We then perform the following over the newly added building block \( i \):

1. For each “leaf” in \( L \) we execute:
   
   (a) **Geometry intersection.** \( k \)'s geometry is stored in \( F_k \) (line 6) and then used to compute the matched geometry parts between \( i \) and \( k \) to generate the geometry \( F_i \) (line 7) and create the new building block \( \alpha \), a direct successor of nodes \( i \) and \( k \) (line 8). \( \alpha \) (iteration 2) corresponds to \( AB \) in Figure 4c.
   
   (b) **Geometry difference (local “subtraction”).** In line 9, we compute the geometry in \( k \) that is not in \( i \), resulting in the geometry \( F_y \) corresponding to the new building block \( \gamma \), now a direct successor of \( k \) (line 10). \( \gamma \) (iteration 2) corresponds to \( B' \) in Figure 4c.

2. **Geometry union-difference (global “subtraction”).** Once we finish going over the list of leaves, we compute the unique geometries that exist in \( i \) (the last added map edition in \( M \)) by subtracting the union of the geometries of the leaf node intersections (with previous processed maps) from the original map block \( i \) (as described in line 11), resulting in the geometry \( F_x \) corresponding to the new building block \( \delta \), now a direct successor of node \( i \) (line 12). \( \delta \) (iteration 2) corresponds to \( A' \) in Figure 4c.

In the worst-case scenario, graph \( G \) will grow as a balanced binary tree. For each added map edition, we increase \( G \)'s depth by one and split the geometry of each leaf node into two parts (shared and unique, with respect to the lastly added node). For \( M \) map editions, we get \( 2^M \) leaf nodes. Assuming that the average number of vectors per feature is \( V \), each leaf node computation will introduce at most \( V \) new vectors. With the assumption that the computation cost for a pair of vectors is constant, the expected time complexity of Algorithm 1 is \( O(2^M \cdot MV) \).

The relations between the nodes in graph \( G \) carry a semantic meaning between the different building blocks (a node is contained in its predecessors and contains its successors) and will play a critical role in the RDF generation and query mechanism since they represent the relations between the building blocks across different points in time of the same region.

### 2.4. Reverse-geocoding and geo-entity linking

Volunteered geographic information platforms [25] are used for collaborative mapping activities with users contributing geographic data. OpenStreetMap (OSM) is one of the most pervasive and representative examples of such a platform and operates with a certain, albeit somewhat implicit, ontological perspective of place and geography more broadly. OSM suggests a hierarchical set of tags\(^3\) that users can choose to attach to its basic data structures to organize their map data. These tags correspond to geographic feature types that we will query (i.e., wetland, railroad, etc.).

Additionally, a growing number of OSM entities are being linked to corresponding Wikipedia articles, Wikidata [26] entities and feature identifiers in the USGS Geographic Names Information Service (GNIS) database. GNIS is the U.S. federal government’s authoritative gazetteer. It contains millions of names of geographic features in the United States.

Our proposed method for the enrichment of the generated geospatial entities (i.e., building block geometries) with an external resource is built upon a simple geo-entity linking mechanism with OSM. This is again a task of entity matching/linking: this time it is with an entity in an external knowledge base.

The method is based on reverse-geocoding, which is the process of mapping the latitude and longitude measures of a point or a bounding box to an address or a geospatial entity. Examples of these services include the GeoNames reverse-geocoding Web service\(^4\) and OSM’s API.\(^5\) These services support the identification of nearby street addresses, places, areal subdivisions, etc., for a given location.

The geo-entity linking process is depicted in Algorithm 2 and illustrated in Figure 6. We start with individual building block geometries of known type (\( T \) in Algorithm 2). In the case of the data we present later in the evaluation in Section 3, we start with the building blocks of geometries of railroads or wetlands (seen in blue in Figure 6), so we know the feature type we are searching for. Each input building block geometry, \( s \), is an individual node in graph \( G \) from Section 2.3. We first generate a global bounding box for \( s \) by find-

---

\(^3\)https://wiki.openstreetmap.org/wiki/Map_Features

\(^4\)http://www.geonames.org/export/reverse-geocoding.html

\(^5\)https://wiki.openstreetmap.org/wiki/API
ing its northmost, southmost, eastmost, and westmost coordinates. The OSM service takes the resulting box as input to execute a reverse-geocoding API call that locates instances of type $T$ on the external knowledge base, as described in lines 1-2. Some of these instances do not share any geometry parts with our inspected building block. As a heuristic, we randomly sample a small number of coordinate pairs (Points), corresponding to the number of entities composing $s$ ($N$ ranges from 10 to 85 in our datasets, as presented in Section 3.1); thus, we gain more confidence in the detected instances, as seen in lines 4-6 in Algorithm 2 and in red in Figure 6. Finally, we reduce the list by removing the matching candidates in the external KB that have a single appearance, thus filtering out entities that are not likely part of the enclosed geometry of our geospatial entity. Each one of the resulting instances is used in later stages to enrich the knowledge graph we construct with additional semantics and metadata from the external knowledge base.

Figure 7 shows an example of a scanned topographic map (seen in the background), which we used to extract its corresponding wetland vector data, alongside two OSM instances detected using our geo-entity linking method. The labels in Figure 7 (i.e., Four Mile Cove Ecological Preserve and Six Mile Cypress Slough Preserve) correspond to the name attribute of each entity. These labels are part of a set of attributes we use to augment our resulting data with information that did not exist in the original dataset.

The time complexity of Algorithm 2 depends on the number of samples $N$ we choose and on the bounding box calculation. Each API call has a constant cost. As well, the bounding box calculation has a constant cost. Thus, the expected time complexity of Algorithm 2 is $O(N)$. As we mentioned in Section 2.3, in the worst-case scenario, graph $G$ will contain $2^M$ leaf nodes (for $M$ map editions); thus, the total expected time complexity of this step in the pipeline is $O(2^M N)$.

2.5. Semantic model

As a generic model or framework, RDF can be used to publish geographic information. Its strengths include its structural flexibility, particularly suited for rich and varied forms for metadata required for different purposes. However, it has no specific features for encoding geometry, which is central to geographic information. The OGC GeoSPARQL [27] standard defines a vocabulary for representing geospatial data on the Web and is designed to accommodate systems based on qualitative spatial reasoning and systems based on quantitative spatial computations. To provide a representation with useful semantic meaning and universal conventions for our resulting data, we define a semantic model that builds on GeoSPARQL.

Our approach towards a robust semantic model is motivated by the OSM data model, where each feature is described as one or more geometries with at-

---

**Algorithm 2: The geo-entity linking algorithm**

**Input:** building block geometry $s$, number of samples $N$, feature type $T$

**Output:** list $\mathcal{L}$ of OpenStreetMap instances in $s$

1. $B_s = \text{bounding box wrapping } s$;
2. $\mathcal{L} = \text{reverse-geocoding}(B_s, T)$; // returns OpenStreetMap instances of type $T$ in $B_s$;
3. for $1 \ldots N$ do
4.   $e = \text{randomly sample a Point in } s$;
5.   $E = \text{reverse-geocoding}(e, T)$;
6.   $\mathcal{L}.\text{add}(E)$;
7. remove instances with a single appearance in $\mathcal{L}$;
8. return $\mathcal{L}$;
tached attribute data. In OSM, relations are used to organize multiple nodes or ways into a single entity. For example, an instance of a bus route running through three different ways would be defined as a relation.

Figure 8 shows the semantic model we describe in this section. In GeoSPARQL, the class type \texttt{geo:Feature} represents the top-level feature type. It is a superclass of all feature types. In our model, each instance of this class represents a single building block extracted from the original vector data.

By aggregating a collection of instances of the class \texttt{geo:Feature} with a property of type \texttt{geo:sfWithin} we can construct a full representation for the geometry of a specific geographic feature in a given point in time. Similarly, we can denote the decomposition to smaller elements using the property \texttt{geo:sfContains}. The use of these properties enables application-specific queries with a backward-chaining spatial “reasoner” to transform the query into a geometry-based query that can be evaluated with computational geometry. Additionally, we use the property \texttt{geo:sfOverlaps} with subjects that are instances from OSM to employ the Web as a medium for data and spatial information integration following linked data principles. Furthermore, each instance has at least one property of type \texttt{dcterms:date} to denote the different points in time in which the building block exists. Each of the aforementioned properties has a cardinality of 1:n, meaning that multiple predicates (relations) of this type can exist for the same building block. The property \texttt{dcterms:created} is used to denote the time in which this building block was generated. \texttt{dcterms} stands for the Dublin Core Metadata Initiative\textsuperscript{6} metadata model, as recommended by the World Wide Web Consortium (W3C).\textsuperscript{7}

Complex geometries are not human-readable as they consist of hundreds or thousands of coordinate pairs. Therefore, we use dereferenceable URIs to represent the geospatial entity instead. Using a named node in this capacity means that each entity has its own URI as opposed to the common blank-node approach often used with linked geospatial entities. Each URI is generated using a hash function (the MD5 message-digest algorithm, arbitrarily chosen) on a plain-text concatenation of the feature type, geometry, and its temporal extent, thus providing a unique URI given its attributes. The geometry contains absolute coordinates, thus rules out the possibility of hash clashes. Each building block instance (geospatial entity) holds a property of type \texttt{geo:hasGeometry} with a subject that is an instance of the class \texttt{geo:Geometry}. This property refers to the spatial representation of a given feature. The class \texttt{geo:Geometry} represents the top-level geometry type and is a superclass of all geometry types.

In order to describe the geometries in a compact and human-readable way we use the WKT format for further pre-processing. The \texttt{geo:asWKT} property is defined to link a geometry with its WKT serialization and enable downstream applications to use SPARQL graph patterns.

Figure 9 shows how the spatio-temporal data, resulting from the previous steps, is mapped into the semantic model (from Figure 8) to generate the final RDF graph. The first column, titled \texttt{gid}, corresponds to the local URI of a specific node (building block geometry). The columns titled \texttt{predecessor_id} and \texttt{successor_id} correspond to the local URIs of the nodes composed-of and composing the specified \texttt{gid} node, respectively. All the three node entities are of type \texttt{geo:Feature}. The data in the \texttt{wkt} column contains the geometry WKT representation. It is linked to the building block node via an entity of type \texttt{geo:Geometry}, as we described above. The rest of the attributes (\texttt{year}, \texttt{time_generated}, and \texttt{OSM_uri}) are stored as literals, following the semantic model we presented in Figure 8.

2.6. Incremental linked data

Linked Data technologies can effectively maximize the value extracted from open, crowdsourced, and pro-

---

\textsuperscript{6}https://www.dublincore.org/specifications/dublin-core/
dcmi-terms/

\textsuperscript{7}https://www.w3.org/
B. Shbita et al. / Building Spatio-Temporal KGs

Following the data extraction and acquisition tasks described in Sections 2.3 and 2.4, and the semantic model described in Section 2.5, we can now produce a structured standard ontologized output in a form of a knowledge graph that can be easily interpreted by humans and machines, as linked data. In order to encourage reuse and application of our data in a manageable manner, we need to make sure that the linked data publication process is robust and maintainable.

This hierarchical structure of our directed acyclic graph $G$ (introduced in Algorithm 1) and its metadata management allows us to avoid an update across all the existing published geographic vector data (in linked data) and instead handle the computations incrementally once a new representation of the feature from a subsequent map edition is introduced.

The approach we present is complete and follows the principles of Linked Open Data by:

1. Generating URIs as names for things, without the need to modify any of the previously published URIs once further vector data from the same region is available and processed.
2. Maintaining existing relations (predicates) between instances (additional relations may be added, but they do not break older ones).
3. Generating machine-readable structured data.
4. Using standard namespaces and semantics (e.g., GeoSPARQL)
5. Linking to additional resources on the Web (i.e., OpenStreetMap)

2.7. Querying

The semantic model presented in Section 2.5 and its structure provide a robust solution enabling a coherent query mechanism to allow a user-friendly interaction with the linked data.

In order to elaborate on the query construction idea, we describe the elements that are needed for a general query “skeleton” from which we can establish more complicated queries to achieve different outcomes as required. Listing 1 shows a query (i.e., the “skeleton” query) that retrieves all the leaf node building blocks (i.e., the most granular building blocks). As shown in Listing 1, we first denote that we are interested in a geoFeature that has a geometry in WKT format which gets stored in the variable ?wkt as shown in lines 3-4 (the variable we visualize in Figures 14, 15, 16, and 17). Line 5 restricts the queried building blocks (geoFeatures) to leaf nodes only (in graph $G$), thus retrieving the most granular building blocks. This is done by discarding all the nodes that hold a predicate of type geo:sfContains, which means that we retrieve only leaf nodes.

```sparql
SELECT ?f ?wkt
WHERE {
  ?f a geo:Feature ;
  FILTER NOT EXISTS { ?f geo:sfContains _:_ }
}
```

Listing 1: Our SPARQL query “skeleton”

This is important due to the way graph $G$ “grows”: as we mentioned previously, every time we add a new representation of the feature from a subsequent map edition, we decompose the existing leaf nodes (most granular building blocks) to a new layer of leaf blocks (newer, smaller and more granular building blocks, if subject to decomposition) and its metadata migrates to the lowest level of nodes (new leaves). This property makes our solution robust and suggests an efficient way of querying, avoiding the need to “climb up” the graph for more complicated (“composed”) blocks.

If, for example, we are interested to see the entire geographic feature in a specific point in time, we can add the clause \{?f dcterms:date <...> .\} inside the WHERE block (lines 2-5). If we are interested
to see the changes from a different time, we can add an additional clause \( \text{\{MINUS \{ ?f \text{dcterms:date }<\ldots> \} \}} \) as well. The syntax and structure of the query allows an easy adaptation for additional tasks such as finding the distinct feature parts from a specific time or finding the feature parts that are shared over three, four or even more points in time or map editions. The nature of our knowledge graph provides an intuitive approach towards writing simple and complex queries.

### 3. Evaluation

We evaluate and analyze our methods using qualitative and quantitative methods over two types of geographic features: railroads (line-based geometry) and wetlands (polygon-based geometry). In this section, we present the results, measures, and outcomes of our pipeline when executed on the following datasets:

**Railroad data** We tested two datasets of vector railroad data (encoded as MULTILINEs) extracted from the USGS historical topographic map archive,\(^{89}\) using the extraction methods of Duan et al.\(^{[7]}\). Each dataset covers a different region and includes map sheets for different points in time. The railroad data originates from a collection of historical maps for:

1. Bray, California (denoted as CA) from the years 1950, 1954, 1958, 1962, 1984, 1988, and 2001 (the original raster maps are shown in Figure 10).
2. Louisville, Colorado (denoted as CO) from the years 1942, 1950, 1957, and 1965.

**Wetland data** We tested three datasets of vector wetland data (encoded as MULTIPOLYGONs) that were similarly extracted from the USGS historical topographic map. Again, each of these map sheets covers a different region and spans different points in time. The wetland data originates from a collection of historical maps for:

1. Bieber, California (denoted as CA) from the years 1961, 1990, 1993, and 2018 (the original raster maps are shown in Figure 11).
2. Palm Beach, Florida (denoted as FL) from the years 1956, 1987, and 2020.
3. Duncanville, Texas (denoted as TX) from the years 1959, 1995, and 2020.

Our primary goal in this section is to show that our proposal provides a complete, robust, tractable, and efficient solution for the production of linked data from vectorized historical maps.

#### 3.1. Evaluation on the feature partitioning

In order to evaluate the performance of this task, we look into the runtime and the number of generated nodes (in graph \(G\)) for each region and feature type (executed on a 16 GB RAM machine @ 2.9 GHz Quad-Core Intel Core i7). The number of vector features in the geographic feature geometry (column ‘\# vecs’), resulting runtimes (column ‘Runtime’, measured in seconds) and total number of nodes following each sub-step of an addition of another map sheet feature geometry (column ‘\# nodes’) are depicted in Tables 2 (CA) and 3 (CO) for the railroad data, and in Tables 4 (CA), 5 (FL) and 6 (TX) for the wetland data.

<table>
<thead>
<tr>
<th>Year</th>
<th># vecs</th>
<th>Runtime (s)</th>
<th># nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1954</td>
<td>2382</td>
<td>&lt;1</td>
<td>1</td>
</tr>
<tr>
<td>1962</td>
<td>2322</td>
<td>36</td>
<td>5</td>
</tr>
<tr>
<td>1988</td>
<td>11134</td>
<td>1047</td>
<td>11</td>
</tr>
<tr>
<td>1984</td>
<td>11868</td>
<td>581</td>
<td>24</td>
</tr>
<tr>
<td>1950</td>
<td>11076</td>
<td>1332</td>
<td>43</td>
</tr>
<tr>
<td>2001</td>
<td>497</td>
<td>145</td>
<td>57</td>
</tr>
<tr>
<td>1958</td>
<td>1860</td>
<td>222</td>
<td>85</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Year</th>
<th># vecs</th>
<th>Runtime (s)</th>
<th># nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1965</td>
<td>838</td>
<td>&lt;1</td>
<td>1</td>
</tr>
<tr>
<td>1950</td>
<td>418</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>1942</td>
<td>513</td>
<td>5</td>
<td>8</td>
</tr>
<tr>
<td>1957</td>
<td>353</td>
<td>4</td>
<td>10</td>
</tr>
</tbody>
</table>

As seen in Tables 2, 3, 4, 5 and 6, we observe that for both types of geographic features, the building block geometries extracted from these maps vary in terms of “quality”. That is, they have a different number of vector lines that describe the geographic feature and each one has a different areal coverage (the bounding box area for each feature geometry is reported in Table 7). This is caused by the vector extraction process and is not within the scope of this paper. We also acknowledge that the quality and scale of the original images used for the extraction affects these parameters, but we do not focus on such issues. We treat these values and attributes as a ground truth for our process.

---

8https://viewer.nationalmap.gov
9http://historicalmaps.arcgis.com/usgs/
First, we note that the growth of the number of nodes in graph $G$ is not exponential in practice due to the way the given geographic features actually change over time. Furthermore, the runtime of each sub-step is also tractable and runs only once when a new set of geometries is inserted from a new map edition. As expected, the first two map editions (for all areas) generate results within less than a minute for railroads and less than three minutes for wetlands, requiring at most three computations: one geometry intersection between two building block geometries and two additional subtractions: a local and a global one (as explained in Section 2.3). By inspecting Tables 2, 3, 4, and 6, we observe that the partitioning runtime depends mostly on two factors: the number of vectors in the geometries and the number of processed maps, as we expected. The more geometry elements we have and the more geometries exist, the more operations we need to run.

These results are not surprising because "leaves" in the graph will only be partitioned in case it is "required," that is, they will be partitioned to smaller unique parts to represent the geospatial data they need to compose. With the addition of new map sheet feature geometries, we do not necessarily add unique parts since changes do not occur between all map editions. This shows that the data processing is not necessarily becoming more complex in terms of space and time, thus, providing a solution that is feasible and systematically tractable.

Additionally, by comparing the first three rows in Tables 2 and 5, we notice that the computation time over a polygon geometry is significantly slower than that of a buffer-padded line geometry. This is despite the railroads having a larger number of feature vec-
3.2. Evaluation on geo-entity linking

In the process of linking our data to OpenStreetMap, we are interested in the evaluation of the running time and correctness (precision, recall, and F1, which is the harmonic mean of precision and recall) of this task.

As we expected, the running time is strongly dependent on the number of nodes in graph $\mathcal{G}$ (exponentially dependent on the number of the processed maps) and the block’s areal coverage, which affects the number of samples using the OpenStreetMap API. The API response time averages 3 seconds for each sample. For the railroad data, the execution time for the set of maps from the CA region took approximately an hour (85 nodes) and only a few minutes for CO (10 nodes). This is not surprising as the CA region covers a bigger area and a larger number of nodes. We observe similar behavior in the wetland data. The FL execution time took approximately 2 hours (13 nodes), as it has the largest areal coverage (as seen in Table 7), while the other regions (CA, TX) took approximately 30 minutes to finish. This provides a feasible solution to a process that runs only once for a given set of geometries from different map editions.

Due to the absence of a gold standard for this mapping task, we had to manually inspect and label the sets of instances found in each bounding box that we query for each building block geometry. The measure we present here is in terms of entity (instance) coverage. Precision and recall are calculated according to the labeled (type) instances that are available on OpenStreetMap and make up the inspected geographic feature (i.e., railroad or wetland). Figures 12 and 13 show an example of such instances, with their corresponding tags in OSM’s graphic interface. Figure 12 shows a wetland instance on OSM marked as a SwampMarsh (with its corresponding GNIS code) and matching an active wetland area in our data (a common building block from all the editions of the CA wetland data). Figure 13 shows a railroad instance on OSM marked as abandoned and matching an abandoned railroad segment in our data (a unique building block from the 1950 edition of the CA railroad data). This shows our ability to enrich and link our graph to external resources on the Web.

We have set up a baseline for comparison with our geo-entity linking method. The baseline approach returns the set of all instances found in the bounding box.
This is the list of candidates we generate in the first step of our method, without the additional sampling and removal steps we have described in Section 2.4.

The precision, recall, and $F_1$ scores of each method over each dataset are shown in Table 7. For each geographic feature and each region, we report the baseline result and our method’s result. We also present the bounding box area for each dataset (in square kilometers), as it is an important factor in the geo-entity linking evaluation. The bigger the area, the more sampling points we require. The first 4 rows correspond to the railroad data (in CA and CO). The rest corresponds to the wetland data. Note, that results for the TX wetlands have been omitted in this part of the evaluation due to a complete absence of labeled data in OSM covering that area. We will briefly discuss this in Section 5.

Due to the varying geometries, areal coverage, and available data in the external knowledge base for each region, and as expected, our measure shows different scores for each dataset. In 3 out of 4 datasets, our method achieves much higher $F_1$ scores than the baseline (0.774 and 0.909 compared to 0.323 and 0.625 respectively in the railroad data; and 1.000 compared to 0.714 in the CA wetland data) and achieves an acceptable score for this task. In the FL wetland dataset, we achieve lower $F_1$ scores for both methods (baseline and ours). This is not surprising as the area of coverage is significantly bigger than in all other datasets, requiring us to generate a bigger number of samples in order to capture all the relevant instances on OSM. We will expand on this issue in Section 5. Nonetheless, further examination of the FL wetland results shows that the low $F_1$ score of the baseline is due to the fact that it only considers the global bounding box (thus the high recall, but low precision). On the other hand, our method achieves a higher precision score but a much lower recall, compared to the baseline. This is a crucial point in geographic applications, as many systems consider the precision to be more important than recall due to a low false-positive tolerance [28].

3.3. Evaluation on querying the resulting data

We execute several query examples over the knowledge graph we constructed in order to measure our model in terms of query time, validity, and effectiveness. For the generated railroad data, we had a total of 914 triples for the CA dataset and 96 triples for the CO dataset. For the wetland data, we had a total of 270 triples for the CA dataset, 149 for the FL dataset, and 85 for the TX dataset.

Larger areas do not necessarily mean that we will have more triples. The number of triples depends on the geometry similarity and difference in the given data. Despite the FL wetland data covering a significantly larger area than other datasets, it did not cause notable triples growth or query performance degradation, as we show in Table 8.

The generated RDF triples would be appropriate to use with any Triplestore. We hosted our triples in Apache Jena.10 Jena is relatively lightweight, easy to use, and provides a programmatic environment.

Table 8 shows the query-time performance results (average, minimum and maximum). In the first type of query we want to identify the feature parts that remain unchanged in two different map editions (different time periods) for each region (e.g., Listing 2). Each row with a label starting with SIM− in Table 8 corresponds to this type of query (the label suffix determines the tested region). We executed a hundred identical queries for each feature type and each area across different points in time to measure the robustness of this type of query.

We repeated the process for a second type of query to identify the parts of the feature that were removed or abandoned between two different map editions for each region (i.e., Listing 3). Each row with a label starting with DIFF− in Table 8 corresponds to this type of query.

The third type of query retrieves the parts of the feature that are unique to a specific edition of the map (i.e., Listing 4). Each row with a label starting with UNIQ− in Table 8 corresponds to this type of query.

10https://jena.apache.org/

Listing 2: Query similar feature geometries in both 1962 and 2001

Listing 3: Query feature geometries present in 1962 but not in 2001
Fig. 14. Example of railroad system changes over time: the parts of the railroad that are similar in 1962 and 2001 are marked in red.


```
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)
```

Looking at Table 8, we notice that the average query times are all in the range of 10-48(ms) and do not seem to change significantly with respect to the number of map editions we process or the complexity of the query we compose. The query time results corresponding to the wetland data are slightly slower, but not significantly, comparing to the railroad data. This may be explained by the longer literal encoding of the WKT geometry for polygons, thus slower retrieval time comparing the the line encoding.

In order to evaluate the validity of our graph we observe the visualized results of the query presented in Listing 2 when executed over the CA railroad data, which are shown in Figure 14. The figure shows in red the unchanged building block geometries between the years 1962 and 2001. We notice that the geometries we retrieve qualitatively match what we observe in the original vector data (the line marked in black over the maps in Figures 14 and 15 represents the current railway, which has not changed since 2001). The results of the query presented in Listing 3 are shown in Figure 15, again, when executed over the CA railroad data. Figure 15 shows in blue the parts of the railroad that were abandoned between 1962 to 2001. Comparably, we perform a similar type of query and visualization for the CA wetland data, between the years 1961 and 2018. Figure 16 shows the similar parts in both editions (in red). Figure 17 shows the parts of the wetland (swamp) that were present in 1961 but are not present in 2018 (in dark blue). Again, this matches our qualitative evaluation based on the original vector files (the light blue marks that are part of the map’s background depict the current swamp, thus validating our results qualitatively).

The query results above establish high confidence in our model, showing that we can easily and effectively answer complex queries in a robust manner. Overall, we demonstrated that our approach and the proposed pipeline can be effectively used to automatically
construct effective and contextualized open KGs and linked data from historical and contemporary geospatial data.

4. Related Work

Much work has been done on mapping geospatial data into RDF graphs. Kyzirakos et al. [15] developed a semi-automated tool for transforming geospatial data from their original formats into RDF using R2RML mapping. Usery et al. [16] presented a method for converting point and other vector data types to RDF for supporting queries and analyses of geographic data. The transformation process presented in these papers does not address linking the data across multiple sources or linking the source data with additional knowledge bases on the Semantic Web as described in this paper.

Annotating geospatial data with external data on the Web is used for contextualization and the retrieval of relevant information that cannot be found in the source data. This line of research has been addressed in different studies. Vaisman et al. [17] studied the problem of capturing spatio-temporal data from different data sources, integrating these data and storing them in a geospatial RDF data store. Eventually, these data were enriched with external data from LinkedGeoData [22], GeoNames [21], and DBpedia [29]. Smeros et al. [30] focus on the problem of finding semantically related entities lying in different knowledge bases. According to them, most approaches on geo-entity linking focus on the search for equivalence between entities (same labels, same names, or same types), leaving other types of relationships (e.g., spatial, topological, or temporal relations) unexploited. They propose to use spatio-temporal and geospatial topological links to improve the process.

The discovery of topological relations among a simplified version of vector representations in geospatial resources has been studied as well. Sherif et al. [31] presented a survey of 10 point-set distance measures for geospatial link discovery. SILK [30] computes topological relations according to the DE-9IM [32] standard. In contrast, RADON [33] is a method that combines space tiling, minimum bounding box approximation, and a sparse index to calculate topological relations between geospatial data efficiently. These approaches work on matching geometry to integrate vector data from two sources in a single matching process. In contrast, our approach can systematically match multiple vector datasets iteratively without performing duplicate work.

Current work on geospatial change analysis spans the construction of geospatial semantic graphs to enable easier search, monitoring and information retrieval mechanisms. Perez et al. [34] computed vegetation indexes from satellite image processing and exposed the data as RDF triples using GeoSPARQL [27]. The changes in these indexes are used to support forest monitoring. Similar to that approach, Kauppinen et al. [35] collected statistical data from open data sources to monitor the deforestation of the Amazon rainforest by building temporal data series translated into RDF. Kyzirakos et al. [19] used open knowledge bases to identify hot spots threatened by wildfire. However, this line of work does not address an incremental process of geospatial change over time. In this paper, we incorporate a temporal dimension to the geospatial se-
mantic graphs and present a pipeline for an automatic incremental geographic feature analysis over time.

5. Discussion and Future Work

With the increasing availability of digitized geospatial data from historical map archives, better techniques are required to enable end-users and non-experts to easily understand and analyze geographic information across time and space. Existing techniques rely on human interaction and expert domain knowledge. In this paper, we addressed this issue and presented an automatic approach to integrate, relate and interlink geospatial data from digitized resources and publish it as semantic-rich, structured linked spatial data in a form of a knowledge graph that follows the Linked Open Data principles.

The evaluation we presented in Section 3 shows that our approach is feasible and effective in terms of processing time, completeness and robustness. The partitioning process runs only once for newly added resources, and does not require re-generation of "old" data since our approach is incremental. In case a new map edition emerges for the same region, we only need to process the newly added geometry. Thus, data that has been previously published will continue to exist with a proper URI and will be preserved over time.

In a scenario that includes contemporary maps that change very quickly, we expect our method to require longer computation time, due to an increased number of computations, but would still be tractable with respect to the changes happening in the map geometries. As we mentioned in Section 2.3, the breakdown of the building block geometries depends on the complexity of the actual changes in the original topographic map. Further, the quality and level of detail of the original vector data play a significant role in the final RDF model as we have observed in Section 3.1.

Further, one may consider cases where a feature changes its very type over time. For example, a large body of water drying out to become represented as a line, or a river causing a flood and generating a polygonal feature. In such a case, we would determine that the feature has either disappeared or came into existence.

Our approach has several limitations, one of them is in a form of a hyper-parameter that governs the buffer size we use in the process of the partitioning of the geometries to smaller building block geometries in the case of line geometry. We currently set this parameter manually but we believe such parameter can be learned from the data or estimated using some heuristics with respect to additional attributes such as the area of the geometric object and the quality of the original vector data extraction.

As we have observed in the geo-entity linking evaluation for the wetland data in TX, in Section 3.2, it is not always possible to provide complete solutions to some problems. Given the large volume, and openness of the OSM schema, there are peculiarities about geographic information that present particular challenges with respect to the Semantic Web. One challenge is the vagueness that exists in geographic categories. For example, what makes a wetland a wetland and not mud? Again, it is worth emphasizing that these peculiarities are not unique to geography. OSM is open to non-expert users of geographic data and thus, the tagging attitude is rather intuitive than based on scientific methodologies and knowledge. We believe that these types of problems can be solved by the downstream application by adjusting the pipeline according to the user's needs.

Moreover, and as we have seen in Section 3.2, performing geo-entity linking over very large areas (such as the case of the FL wetland data), can cause poor performance. In order to achieve better recall, we need to increase the number of the reverse-geocoding calls and perform adjustable bulk query calls instead of an unbounded number of single API calls (which we avoided due to the policies in using the OSM API).

An additional solution for this issue can come in the form of a caching mechanism and by using multiple machines for faster parallel access.

We are currently looking into expanding the ability to utilize additional knowledge bases such as Yago2Geo [36]. This knowledge base gathers geospatial entities with their spatial data from different sources (including OSM). In future work, we also plan to investigate the possibility of using multiple machines for faster processing. This is possible since there are computations in Algorithm 1 that are independent of each other and can be executed in parallel in the same iteration over a single map edition. This will enable a faster processing time and strengthen the effectiveness of our solution.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant Nos. IIS 1564164 (to the University of Southern California) and IIS 1563933 (to the University of Colorado at Boulder) and Nvidia Corporation.
References


