Knowledge Graph of Places

Zhenning Shangguan
Pitney Bowes Inc.,
27 Waterview Drive,
Shelton, CT 06484, USA
zhenning.shangguan@pb.com

Background

Pitney Bowes, as a market leader in the GIS industry, possesses large amounts of geospatial data, covering business points, land parcels, road networks, indoor maps, Points of Interests (POI), 3D building structures, and geographical boundaries. Recently, data sources that generate geospatial data in the form of continuous streams, such as real-time traffic, social activities and GPS navigation trajectories, are also incorporated into our data universe. In theory, these datasets should enable us to provide rich contextual views of places in the physical world and enable new business opportunities (in the form of new data products and services), once properly combined and presented. However, reality is that even within the company itself, typical daily operations, such as discovering, exploring, merging, augmenting, and making good use of these geospatial data assets, quite often involve cumbersome steps and processes. This is the major motivation that drives the need for a linked geospatial data platform, which we call the Knowledge Graph of Places.

The Knowledge Graph of Places

Scope. The Knowledge Graph of Places is an enterprise linked geospatial data platform that exposes various geospatial data, possessed or licensed by Pitney Bowes, as an accessible, unified semantic data graph. One special notice is that we use the term Place deliberately to reflect its philosophical vagueness, and thus the broad coverage of our system in terms of physical (such as buildings, landmarks) and logical (such as neighborhoods, business areas) geospatial entities. From a functionality perspective, it serves as the single source of ground truths for our data universe. Also, arbitrary parts of the knowledge graph should be easily accessible and interacted with using standard web technologies, enabling ad-hoc exploration and lightweight consumption. Finally, flexible export mechanisms should be supported in order to fit into existing geospatial data engineering pipeline.

Challenges. There are various challenges embodied in building a linked geospatial data platform that exposes resources in the knowledge graph and supports common geospatial operations. They can be roughly categorized as follows:

1) Vocabularies to semantically represent a place. Conceptually, an arbitrary place can be represented using its identity, geometric properties, surrounding environments, and internal structures. While initiatives like GeoSPARQL provide a solid basis to represent geometries and spatial relationships, the coverage of our knowledge graph goes way beyond that. A good example is the task of representing a particular building. Besides its unique identifier, parcel boundaries, and surrounding neighborhoods, its indoor layout should also be represented as part of the knowledge graph (given that we have data

about it). This covers not only geometric points (e.g., bulletin boards) and polygons (e.g., stores), but also internal building structures and functional units, such as different levels and staircases. Standardized vocabularies to represent dimensions are noteworthy as well, because they are critical for 3D building structures and layouts. One last note is how to represent qualitative counterparts, such as zip codes in the United States and postal codes in Canada. It is obvious that an *owl:sameAs* is semantically inaccurate in this case.

2) Publishing Knowledge Graph of Places as linked geospatial data platform. One outstanding challenge is that our geospatial data are stored in several different formats, such as Shapefile, TAB files, relational databases, and GeoJSON files stored in Hadoop/HDFS. Our initial experiment of converting and loading all of them into a single triple store was largely unsuccessful. Ideally, we would prefer an approach that can leave geospatial data distributed and untouched as is, while still be able to expose them as a semantically unified body of knowledge. Investigating techniques such as R2RML is currently underway.
3) Enabling novel ways of consuming data from Knowledge Graph of Places. Typical ways of consuming linked data is either to query against dedicated SPARQL endpoint(s) or to dereference URIs of geospatial entities to retrieve either a human readable format (e.g., HTML) or a machine understandable format (e.g., RDF/XML) via content negotiation. This, unfortunately, holds true for existing linked geospatial data too. Technically, both approaches have their own pitfalls: SAPRQL endpoints give you full expressive power of graph pattern matching but does not let you see the whole picture (or graph) easily; single resource-based URI dereferencing is intuitive and lightweight, but it soon becomes cumbersome when you want to navigate the knowledge graph. Unique characteristics of geospatial data actually impose the requirement for a proper mixture of both paradigms.

Presentation Proposal

We will present an overview of our endeavors to build the Knowledge Graph of Places within Pitney Bowes, with an emphasis on experiences and lessons learned when addressing those aforementioned challenges throughout the project. We will also highlight some use cases enabled by its functionalities as a linked geospatial data platform, such as place-based geocoding/reverse geocoding, place context resolution, ad-hoc geospatial traversal queries, and web/mobile application mashups. Our goal is to inspire and participate in various possible discussions around best practices, technical challenges, and standardization gaps in the process of consolidating, publishing, exchanging, and consuming geospatial data, both internally and externally facing for an enterprise.