Geospatial Data Integration with Linked Data and Provenance Tracking

Andreas Harth
Karlsruhe Institute of Technology
harth@kit.edu

Yolanda Gil
USC Information Sciences Institute
gil@isi.edu

Abstract

We report on our experiences with integrating geospatial datasets using Linked Data technologies. We describe NeoGeo, an integration vocabulary, and an integration scenario involving two geospatial datasets: the GADM database of Global Administrative Areas and NUTS, the Nomenclature of Territorial Units for Statistics. We identify the need for provenance to be able to correctly interpret query results over the integrated dataset.

1 Introduction

Geospatial data is an important domain in open data; 31 of 295 datasets (around 19 % of triples) in the Linking Open Data cloud are in the geospatial domain[1]. Given the push for open data, including in the area of geospatial data (e.g., the EU's INSPIRE directive), we expect more datasets to become available on the web. As these datasets cover complementary aspects but also overlap to some degree, there is a need for integration, to be able to query and analyze the various datasets in combination.

In the following we focus on the integration of multiple geospatial datasets and describe several open issues we have encountered when integrating geospatial datasets based on semantic technologies. To be able to integrate and query multiple geospatial datasets from the web, we need:

- An integration vocabulary; we explain how we use NeoGeo [3] to model two datasets: the GADM database of Global Administrative Areas and NUTS, the Nomenclature of Territorial Units for Statistics, used to identify geospatial regions in the EU. Both datasets are available as Linked Data[2].
- Means of evaluating queries over the integrated datasets; we use the Region Connection Calculus (RCC) [4] to model and query relations between regions, such as "region 1 is fully contained in region 2".
- A way of tracking the provenance of integrated data, to be able to interpret query results correctly. Provenance is especially important in an open environment such as the web, where sources may make arbitrary statements. We propose to use the W3C PROV vocabulary [2] for describing the provenance of individual datasets. We sketch how to provide provenance information in the query scenario where query results may be derived from the combination of multiple datasets.

[1] http://lod-cloud.net/state/, accessed 2014-01-19
[2] http://gadm.geovocab.org/ and http://nuts.geovocab.org/

Our contributions are as follows: i) we describe a scenario for geospatial data integration and querying; and ii) we highlight the need for provenance in the web setting where software integrates data from multiple providers without necessarily checking a priori the trustworthiness of the data providers.

2 NeoGeo Integration Vocabulary

To be able to bring together datasets from different sources, we need an integration vocabulary which brings the source datasets to roughly the same modelling abstractions. Publication of geospatial data should be as generic and independent of the final use as possible to ensure applicability of the data in a broad variety of scenarios. We use a layered approach, where at the bottom layer we use Linked Data as generic way to publish and access individual datasets, with an optional integration layer on top. Such an architecture provides flexibility in accessing and linking different sources. Before we cover the integration layer in Section 3, we describe our approach to modelling in more detail.

NeoGeo consists of two vocabularies, the `spatial` vocabulary covering geographic features and the `geometry` vocabulary covering geometries (see Figure 1). The distinction between geographic features and geospatial geometries is often used in recent geospatial data formats[3]. NeoGeo models that distinction, with `spatial:Feature` and `ngeo:Geometry` being separate classes, connected with a `spatial:geometry` property.

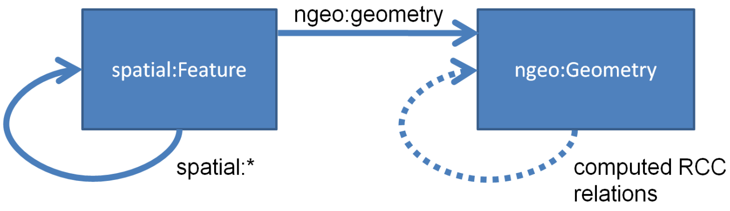

Figure 1: High-level modelling approach taken in NeoGeo. Relations are described on the level of `spatial:Feature`, whereas RCC calculations that form the basis of the relations are computed from instances of `ngeo:Geometry`.

A similar modelling approach has been taken by the recent ISA Programme Location Core Vocabulary[4] and GeoSPARQL[5]. Table 1 lists possible equivalence mappings[6]. Please note that we do not prescribe any specific format for encoding geometries in NeoGeo, due to the contentious issue of choosing one format over the other. Rather, we allow multiple formats via HTTP Content Negotiation[7].

Table 1: Proposed equivalence mappings between classes in NeoGeo, the ISA Programme Location Core Vocabulary and GeoSPARQL.

NeoGeo	Location Core	GeoSPARQL
`spatial:Feature`	`dcterms:Location`	`<http://www.opengis.net/ont/geosparql#Feature>`
`ngeo:Geometry`	`locn:Geometry`	`<http://www.opengis.net/ont/geosparql#Geometry>`

Given such a high-level model, we are able to cover many different established GIS data formats encoded in text or XML syntax. Please note that in many such datasets, there is a 1:1 connection

[3] e.g., GeoJSON, `http://geojson.org/`
[4] `http://www.w3.org/ns/locn`
[5] `http://www.opengeospatial.org/standards/geosparql`
[6] Namespaces via `http://prefix.cc/`
[7] `http://www.w3.org/Protocols/rfc2616/rfc2616-sec12.html`

between feature and geometry. A 1:n relation between feature and geometry may occur if a data provider publishes multiple geometries in different granularity (e.g., in cases where downloading or rendering complex geometries would take up too much time). Also, when integration takes place, the same feature (e.g., the country of Luxembourg) may be connected to many different geometries. Consider one geometry from NUTS with low resolution and one geometry from GADM with high resolution, connected to the same `spatial:Feature` instance representing Luxembourg, as illustrated in Figure 2. For one possible algorithm for detecting equivalences between features based on geometries calculated via the Hausdorff distance see [5].

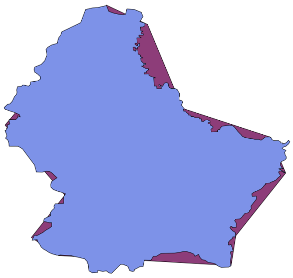

Figure 2: Incongruence of geometric data about Luxembourg from NUTS (low resolution, violet) and GADM (high resolution, blue).

In both cases, there is one "master" geometry and multiple derived geometries; we need provenance to capture the relations between the different geometries to be able to assess the correctness of query results.

A corner case arises with the widely-used W3C geo:Point[8] class, which conflates the notion of feature and geometry. Latitude and longitude of a `geo:Point` are directly used with the URI of a feature. When integrating data from, e.g., DBpedia and GeoNames, the resulting set of triples contains two `geo:lat` and `geo:long` values, with no way to discern which latitude/longitude pairs belong together without recording provenance of the individual triples.

3 Integrating and Querying Datasets

For the following discussion, we assume access to multiple geospatial datasets in NeoGeo, with access to their geometries via HTTP Content Negotiation. In our scenario, we want to enable queries over the combination of datasets, such as "return all geographic features that Luxembourg is contained in".

To encode and query for relations between geometries, we use the Region Connection Calculus (RCC) [4]. We provide HTTP access to RCC relations that are computed based on spatial indices using Linked Data Services [6].

Please note that RCC defines relations between regions (i.e., geometries), whereas users often want to pose queries that relate to the features[9]. In other words, users who want to figure out the relation between two features would need to write queries that check the relation of the connected geometries, making queries unwieldy to write.

To simplify writing queries, a system could derive the relations between features automatically from computing RCC relations between geometries. In such a case, however, the system should annotate the

[8] `http://www.w3.org/2003/01/geo/wgs84_pos#Point`

[9] We thank Sean Gilles for pointing out that use case during the 2011 GeoVoCamp in Southampton, UK.

query results with information about the method of deriving the results. Only then users have a means to inspect, interpret and potentially fix incorrect query results. We can use provenance to achieve such functionality.

4 Trusting Integrated Geospatial Data: The Need for Provenance

When users are presented with geospatial information that has been integrated from different sources, they need to understand its provenance in order to trust it. We define *trust* as a judgment that a user makes based on the context of the information they see [1]. A crucial part of this context is provenance, which aims to capture the who/what/when/how/why the information was generated.

Provenance needs to be recorded at different granularity. At a very coarse level, a user may want to examine the provenance of the integrated dataset. For example, users may trust the information if they know what integration algorithm was used and what data sources were integrated. Finer-grained provenance would concern particular features in the dataset. For example, a user may want to see what features in the original datasets were integrated to generate a feature that appears on the integrated dataset. Even finer-grained provenance could be needed at the level of particular attributes of a feature. For example, in seeing the values of latitude/longitude attributes, a user may want to see from what original data sources those values were taken from.

A major challenge to provenance is scale. Maps have millions of objects and properties, and storing detailed provenance for each one of them can result in very inefficient reasoning to answer queries.

Provenance for geospatial information has an important temporal component. As the original data sources are updated, so is the integrated dataset. Another reason to update the integrated dataset is if new version of the integration algorithm becomes available and the integration process is re-executed. Or the updates may be done routinely at some set periodicity.

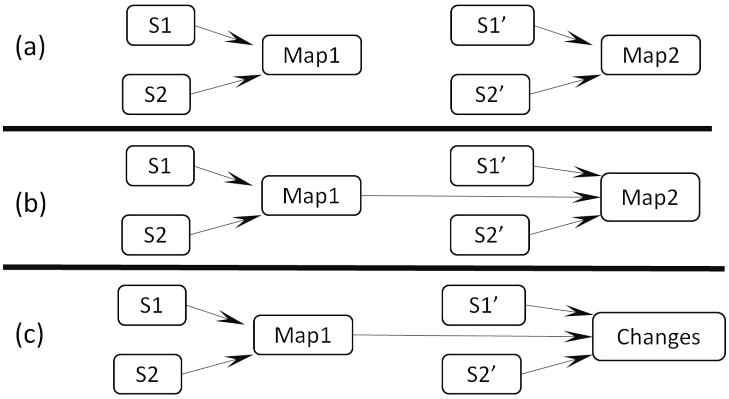

Figure 3: Alternative approaches to updating an integrated dataset.

Figure 3 illustrates alternative approaches to creating new versions of the integrated dataset: 1) the new version of the integrated dataset is generated anew, 2) the new version of the integrated dataset is generated taking into account the previous version of the dataset, and 3) only the delta of the changes are generated.

5 Conclusion

We have described the publication of two geospatial datasets, and analyzed the challenges that arise when integrating multiple geospatial datasets. Standardization activities could include the specification of preferred syntaxes for encoding geometries, means for accessing and computing RCC relations, the relation of feature/geometry vocabularies with the W3C WGS84 Geo Positioning vocabulary and the definition of best practices of using the W3C Provenance Vocabulary in the geospatial domain.

Acknowledgements

This work has been supported by the European Commissions Seventh Framework Programme FP7/2007-2013 (PlanetData Network of Excellence, grant 257641) and by the National Science Foundation with grant IIS-1117281.

References

[1] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2):58–71, 2007.

[2] P. Groth and L. Moreau. PROV-Overview, 2013. `http://www.w3.org/TR/prov-overview/`.

[3] B. Norton, L. M. Vilches, A. D. Len, J. Goodwin, C. Stadler, S. Anand, D. Harries, B. Villazn-Terrazas, and G. A. Atemezing. NeoGeo Vocabulary Specification - Madrid Edition. Juan Martin Salas and Andreas Harth (editors), `http://geovocab.org/doc/neogeo/`.

[4] D. Randell, Z. Cui, and A. Cohn. A Spatial Logic Based on Regions and Connection. *KR*, 92:165–176, 1992.

[5] J. M. Salas and A. Harth. Finding Spatial Equivalences Across Multiple RDF Datasets. In *Proceedings of the Terra Cognita Workshop*, 2009.

[6] S. Speiser and A. Harth. Integrating Linked Data and Services with Linked Data Services. In *Proceedings of the 8th Extended Semantic Web Conference (ESWC 2011)*, pages 170–184, 2011.