Challenges for Provenance Analytics Over Geospatial Data

Daniel Garijo[1](✉), Yolanda Gil[2], and Andreas Harth[3]

[1] Ontology Engineering Group,
Universidad Politécnica de Madrid, Madrid, Spain
dgarijo@fi.upm.es
[2] Information Sciences Institute,
University of Southern California, Los Angeles, USA
gil@isi.edu
[3] Institute AIFB, Karlsruhe Institute of Technology,
Karlsruhe, Germany
harth@kit.edu

Abstract. The growing availability of geospatial data online, the increased use of crowdsourced maps and the advent of geospatial mashups have led to systems that deliver data to users after integration from many sources. In such systems, understanding the provenance of geospatial data is crucial for assessing the quality of the data and deciding on whether to rely on the data for decision making. To be able to use and analyze provenance in geospatial integration systems in a principled manner, we identify different levels of provenance in the geospatial domain, provide a set of provenance questions from the point of view of end users, and relate our geospatial provenance model to the W3C PROV recommendation.

1 Introduction

The Open Geospatial Consortium and the World Wide Web Consortium are working jointly towards standards for linking and integrating geospatial data [1]. As geospatial data is often used in decision making (e.g., navigation), the accuracy of integrated data is important. While we specifically cover provenance for geospatial information, some of these challenges are present in many other domains as well. The area of geospatial data integration is a prime scenario for provenance management, as the involved data and systems are complex and exhibit many challenging characteristics:

- External sources: when integrating two geospatial datasets, an algorithm might consult other sources.
- Human-in-the-loop processes: in some cases, the integration might involve manual intervention, to check particular values by seeking additional confirmation or even perhaps with eyes on target.
- Crowdsourcing: datasets may have been collected from many small contributions, which should attacj provenance too.

B. Ludäscher and B. Plale (Eds.): IPAW 2014, LNCS 8628, pp. 261–263, 2015.
DOI: 10.1007/978-3-319-16462-5_28

- Granularity: geospatial information may be represented at different levels of granularity in space; a geographical feature can be a point in space (e.g., a road intersection), a one-dimensional segment (e.g., a bridge that connects two points) or a two-dimensional region (e.g., a parking lot).
- Computation: spatial reasoning may be needed to compute relationships between features; the integration system may have to integrate computed relations from different sources.
- Versioning: maps are updated as the original data sources are updated. The objects in a map themselves can have multiple revisions.

We present an initial study on the requirements and challenges of tracking geospatial provenance, based on discussions with researchers and practitioners at several meetings and workshops on geospatial data.

2 Geospatial Provenance Model

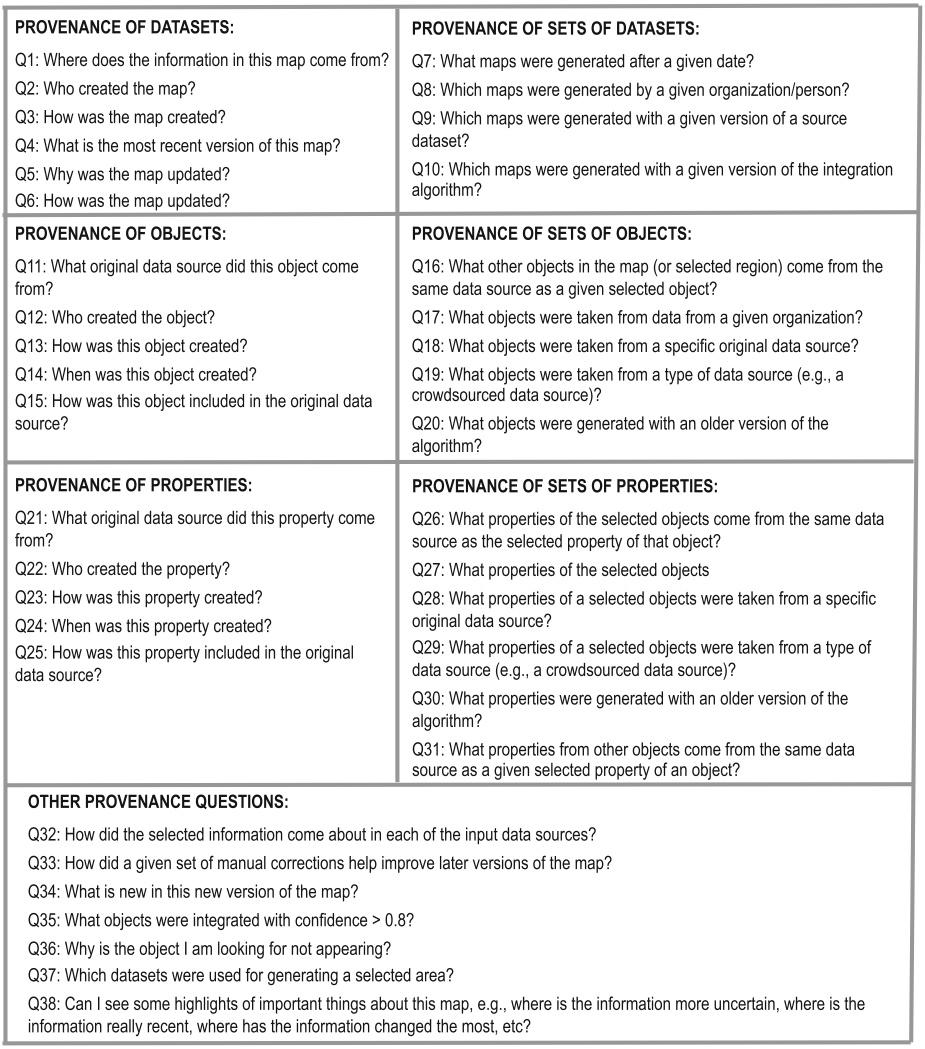
Before we explain how to apply the W3C PROV standard model [2] to the geospatial domain, we present a classification of provenance levels on geospatial data:

- Dataset-level provenance: provenance assertions about a map as a single entity. The map contains objects, and these objects contain properties and values, but provenance is associated with the map as a whole.
- Object-level provenance: how different objects were created in the map.
- Property-level provenance: enables us to answer questions about attributes and attribute values of objects shown in the map.

Modeling detailed provenance across all levels presents a challenge of scale. Maps can have millions of objects, and if we represented each of the integration processes for each object, the amount of information could become larger than the map itself, especially if we assume updates at regular intervals. Property-level provenance aggravates the scale issues of object-level provenance.

In Fig. 1, we list user questions concerning geospatial provenance, grouped according to our provenance model for geospatial data.

Applying PROV to the geospatial domain is straightforward for dataset-level and object-level provenance, as we can use dataset and object identifiers as handle for attaching provenance records to. Property-level provenance requires a more involved approach, as properties are typically accessed through the object and cannot be referenced as a separate entity. Therefore, we would either need to create new identifiers for each property assertion, or to repeat the property assertion itself to be able to attach the provenance record to. Tracking appearing and disappearing objects or values across versions would require to store the entire history of all datasets, including provenance records.



PROVENANCE OF DATASETS: Q1: Where does the information in this map come from? Q2: Who created the map? Q3: How was the map created? Q4: What is the most recent version of this map? Q5: Why was the map updated? Q6: How was the map updated?	**PROVENANCE OF SETS OF DATASETS:** Q7: What maps were generated after a given date? Q8: Which maps were generated by a given organization/person? Q9: Which maps were generated with a given version of a source dataset? Q10: Which maps were generated with a given version of the integration algorithm?
PROVENANCE OF OBJECTS: Q11: What original data source did this object come from? Q12: Who created the object? Q13: How was this object created? Q14: When was this object created? Q15: How was this object included in the original data source?	**PROVENANCE OF SETS OF OBJECTS:** Q16: What other objects in the map (or selected region) come from the same data source as a given selected object? Q17: What objects were taken from data from a given organization? Q18: What objects were taken from a specific original data source? Q19: What objects were taken from a type of data source (e.g., a crowdsourced data source)? Q20: What objects were generated with an older version of the algorithm?
PROVENANCE OF PROPERTIES: Q21: What original data source did this property come from? Q22: Who created the property? Q23: How was this property created? Q24: When was this property created? Q25: How was this property included in the original data source?	**PROVENANCE OF SETS OF PROPERTIES:** Q26: What properties of the selected objects come from the same data source as the selected property of that object? Q27: What properties of the selected objects Q28: What properties of a selected objects were taken from a specific original data source? Q29: What properties of a selected objects were taken from a type of data source (e.g., a crowdsourced data source)? Q30: What properties were generated with an older version of the algorithm? Q31: What properties from other objects come from the same data source as a given selected property of an object?
OTHER PROVENANCE QUESTIONS: Q32: How did the selected information come about in each of the input data sources? Q33: How did a given set of manual corrections help improve later versions of the map? Q34: What is new in this new version of the map? Q35: What objects were integrated with confidence > 0.8? Q36: Why is the object I am looking for not appearing? Q37: Which datasets were used for generating a selected area? Q38: Can I see some highlights of important things about this map, e.g., where is the information more uncertain, where is the information really recent, where has the information changed the most, etc?	

Fig. 1. User questions concerning geospatial provenance.

References

1. Archer, P.: Joint W3C/OGC Workshop on Linking Geospatial Data, March 2014. http://www.w3.org/2014/03/lgd/
2. Moreau, L., Missier, P.: PROV-DM: The PROV Data Model (2012). http://www.w3.org/TR/prov-dm/