Feature and attribute level provenance for spatial data supply chain using semantic web technologies

M.A. Sadiq [a], G. West [a], D.A. Mcmeekin [a], L. Arnold [a] and S. Moncrief [a]
[a] *Department of Spatial Science, Curtin University, Perth, WA, Australia*
Cooperative Research Centre for Spatial Information, Australia
Email: Muhammad.sadiq@postgrad.curtin.edu.au

Abstract: Spatial data supply chains (SDSC) for next generation spatial infrastructures require extensive investigation to address several contemporary issues and challenges that are hampering the innovation and use of spatial information across different industry sectors. SDSCs consist of multiple value chains. Each value chain has heterogeneous geo-processes, methods, models and workflows that combine to generate, modify and consume spatial data.

The integration and processing of multiple datasets gives rise to questions about trust, quality, and fitness for purpose, currency and data authoritativeness. This is because multiple datasets originate from heterogeneous sources, and different geo processes have been executed to reach the final product. Users have different data requirements and therefore knowing how data is collected and at what level of accuracy, provides knowledge about what it can be used for leading to increased user confidence.

With the advent of semantic web technologies, new methods for exploring and understanding the provenance of spatial data have become possible. However, there are few models that address spatial data provenance and none that adequately cater for spatial information management and the dissemination of data to users. A comprehensive provenance model for the spatial domain in Australia and New Zealand is an industry imperative to establish trust, fit for purpose and quality. Understanding provenance is crucial to capturing information about spatial features such as who/what/when/how/why it has been generated. This information is needed to support well informed and reliable evidence-based decision making.

This research is addressing spatial data provenance issues using semantic web technologies to resolve the gaps in our understanding of data provenance when disseminating spatial products. Two generic models from the World Wide Web Consortium (W3C) and the Open Provenance Group are available for general data on the Web. However, both models do not satisfy geospatial needs. The Open Geospatial Consortium (OGC) has investigated the W3C PROV model for spatial datasets. Issues identified are the need for provenance to be captured at various levels including at the spatial feature and attribute levels, for time series data sets, for representation and in presentation interfaces and elements, and for different levels of provenance.

In this research we are focusing on a specific example, namely capturing provenance at feature and attribute level from capture through process such as edge matching to dissemination. It is very important that structural steps of each geo process model and in most of the cases groups of geo processes in a complex analysis data model including overlay, proximity and table analysis should be captured. In this paper we suggest a feature and attribute level provenance and develop an ontology model for an edge matching geo process. We chose feature and attribute level provenance because it has geometric and non-geometric attributes derived from different techniques and originated from multiple features and sources. In order to allow a user to determine the suitability of a dataset for their purposes, provenance information at the single feature level including its history and several other attributes are required. Four phases of edge matching process ontologies have been developed and relationships between classes and sub classes have been defined with object and data properties.

This model named as GeoPROV captures information of the geo process “generate edge matching feature links” and workflow history. This geo process examines features such as segments of a road network acquired by different people using possibly different techniques with the result that the features don’t always align. Information about the data sources and how adjacent features are matched and linked to form new dataset is stored. This is an incremental process and at each stage all provenance information is captured.
The history of each function and changes that occurs are recorded. Information captured according to ISO 19139 geographic information metadata specifications is stored as an RDF file, and SPARQL queries are used to search the RDF provenance store. Having access to real-time provenance information supports data access and end-user confidence in data products.

Keywords: *Spatial data supply chain, data provenance, modelling, ontologies, semantic web technologies*