SPATIAL DATA SUPPLY CHAINS

Premalatha Varadharajulu [a,b], Muhammad Azeem Saqiq [a,b], Feiyan Yu [a,b], David A. McMeekin [a,b,*], Geoff West [a,b], Lesley Arnold [a], Simon Moncrieff [a,b]

[a] Department of Spatial Sciences, Curtin University, GPO Box U1987, Perth 6845, Western Australia, Australia
[b] Cooperative Research Centre for Spatial Information, Australia -
d.mcmeekin@curtin.edu.au

Commission VI, WG VI/5

KEY WORDS: Spatial Data Supply Chains, Provenance, Conflation, Semantic Web, Ontologies, Artificial Intelligence

ABSTRACT:

This paper describes current research into the supply of spatial data to the end user in as close to real time as possible via the World Wide Web. The Spatial Data Infrastructure paradigm has been discussed since the early 1990s. The concept has evolved significantly since then but has almost always examined data from the perspective of the supplier. It has been a supplier driven focus rather than a user driven focus. The current research being conducted is making a paradigm shift and looking at the supply of spatial data as a supply chain, similar to a manufacturing supply chain in which users play a significant part. A comprehensive consultation process took place within Australia and New Zealand incorporating a large number of stakeholders. Three research projects that have arisen from this consultation process are examining Spatial Data Supply Chains within Australia and New Zealand and are discussed within this paper.

1. INTRODUCTION

The next generation spatial infrastructures must address multiple contemporaneous issues within the spatial data supply chains (SDSC). A SDSCs consists of numerous value add processes along the chain. At each value add point in the chain there may be heterogeneous geo-processes, methods, models and workflows combining to generate, modify and consume spatial data. The value add processes occurring in integrating and processing multiple datasets raises questions about data trust, quality, its fitness for purpose, currency and authoritative level. A reason for this is these datasets originated from different sources having had different geo-processes executed upon them to arrive at this final product. Knowing how data is collected and what level of accuracy was used gives understanding as to what purpose the data can be used for. The creation of a geo-spatial provenance model that captures these kinds of processes will enable an ability to measure how fit for purpose data may actually be.

Geospatial data sharing is extremely important in Spatial Infrastructures (SI) as huge amounts of data are supplied by a variety of different organisations, stored in different formats and managed at different user levels. In Australia, the increased dependency on timely spatial data has led to an identified need to consider a supply chains model for spatial data from local government authorities all the way through to the Commonwealth government.

A large quantity of the Australian spatial data is acquired at the local government level, it is then combined to form the State or Territory level datasets and then used to create national level datasets. Many processes used in spatial data generation are manual and undocumented as well as implicitly requiring human intervention. There is a lack of or no linking mechanisms at all between datasets. Multiple versions of data sets are also often being used which may lead to an inaccurate or out of date dataset being used. There are dependencies between the different data at different levels including differing formats and human intervention. These factors complicate dataset integration at different levels.

This research is examining technical solutions to the spatial industry supply chain problems through the application of semantic web and linked data technologies to address the boundaries and gaps identified that prevent seamless supply chain integration and operation. The research is concentrated on determining a universal approach that can be framed to deal with SI supply chain issues that allow for the understanding and automation of the supply chain process that incorporates multiple data sources including crowd sourced data.

2. SPATIAL DATA CONFLATION

In Australia many organisations at the local government level, within state government departments in different jurisdictions as well as Commonwealth agencies, acquire spatial data for specific areas or points of interest independent of each other. This leads to data duplication at multiple points along the SDSC. Lack of awareness or simply because no single dataset suits multiple agencies' needs, leads to this duplication.

To improve the SDSC process, part of this research is to examine data conflation as a means to reduce or even remove the data duplication within the SDSC. Through combining multiple, overlapping data sources into a single point of truth dataset while retaining accuracy, reducing redundancy, reconciling data conflicts and obtaining richer attributes is the aim of this research.

This research is applying semantic web technologies to automate this conflation process. The focus is on creating ontologies using OWL-2 for spatial datasets and coding relevant geometry, topology, and policy rules that can be mined from

* Corresponding author

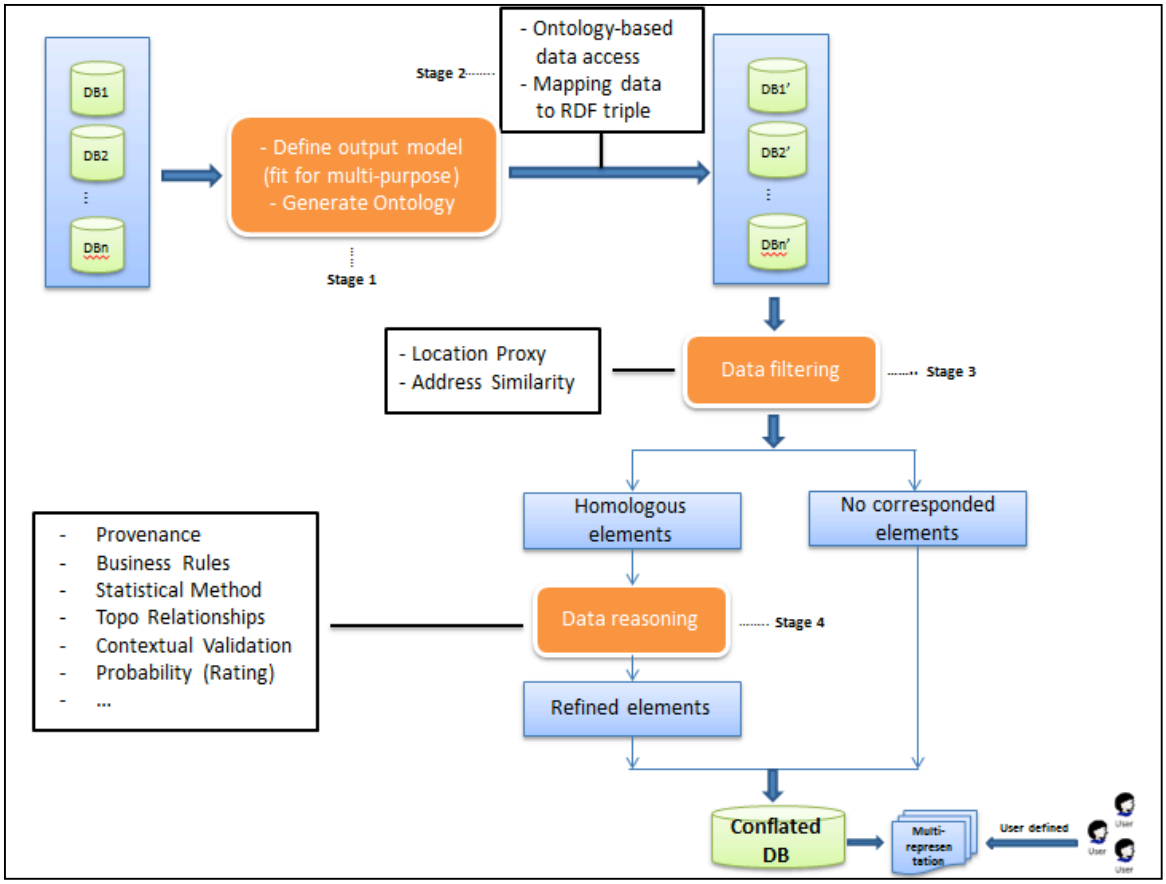

Figure 1. Data Conflation Conceptual Model

either the data or the supporting documents. The satisfying of these rules (Description Logic or DL) through computer reasoning, relevant datasets may be intelligently linked and integrated.

Based on this idea, a Data Conflation Conceptual Model has been designed and is presented in Figure 1. It includes the following:

- **Stage 1**: Preliminary analysis of heterogeneous source datasets with different user needs taken into account to formulate the output data model which needs to meet multiple purposes. Ontologies are then generated accordingly.
- **Stage 2**: Datasets are accessed and data instances are mapped to ontologies and stored in a RDF triple format. In this way all data are in a common format and ready for initial filtering in Stage 3, and the reasoning process in Stage 4.
- **Stage 3**: An initial filter based on location proxy and address similarity is run to determine which elements are homologous elements and which elements are not. No corresponded elements are stored at this stage, as they will be conflated at a later point.
- **Stage 4**: A comprehensive reasoning process is run among homologous elements in order to identify the best location (spatial accuracy) and richest attributes (feature characteristics). The reasoning results, together with those elements that do not correspond are then exported as the single conflated dataset.

The conflated dataset then becomes a single authoritative, trusted data source, fit for multiple purposes. The data can be co-maintained, providing the different agencies, departments and jurisdictions with an understanding that they maintain their own data and yet be used by multiple organisations eliminating the need for "siloed" duplicated systems.

3. GEOSPATIAL TRANSACTIONS

Currently in Australia, local government authorities transact with a State Government Agency, such as Landgate in Western Australia and DELWP in Victoria, for administrative boundary changes, as well as road and place name changes. However, currently implemented techniques are far from automatic in nature and in fact are highly manual requiring significant human involvement. This research uses semantic web technologies and Artificial Intelligence to enable automated spatial transactions on a central database. The case study programs state government agency business policies in a knowledge-based system. Thus it provides a rule-based decision support system to the user, resulting in accurate spatial transactions.

Currently developers approach a land agency through local government for any spatial transitions that may cause a change to spatial datasets. These datasets can then be used as a data source for other process (Schmitz, Scheepers, De Wit, & De la Rey, 2007). However, currently implemented techniques are far from automatic in nature and are highly manual requiring significant human involvement. The manual nature of the current situation means speed and adaptability to new situations is intolerable because of the need for comprehensive sets of tools and techniques that are demanded to produce such a large amount of different spatial data and to automatically link the different type of features present in an Open Distributed Marketplace (ODM) (West, 2014).

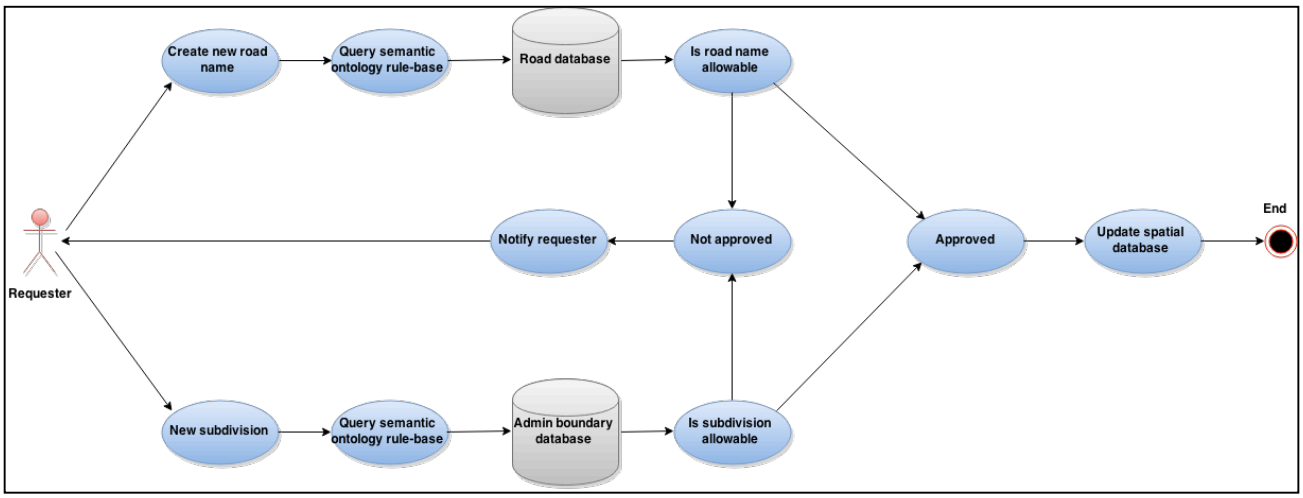

Figure 2. Geospatial Transaction Model.

Recent research (Yu & Liu, 2013; Zhao, Zhang, Wei, & Peng, 2008) has applied the concepts of the semantic web to geospatial data sharing. The semantic web (Berners-Lee, Hendler, & Lassila, 2001) promotes the use of formats such as the Resource Description Framework (RDF) with Universal Resource Identifier (URI) and the Web Ontology Language (OWL) to represent relationships between data and concepts and provide semantics for data. This research aims to provide a self-service mechanism for local government spatial transactions with their state government's spatial data through the use of the semantic web combined with Linked Data technologies (Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods (Heath, 2009)). Technically, Linked Data refers to data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external data sets, and can in turn be linked to from other external data sets (Bizer, Heath, & Berners-Lee, 2009). This addresses the need for more seamless spatial data supply chain operations.

The case study develops a methodology to enable local government authorities to transact with a state government agency in an online environment for administrative boundary changes, and road and place name changes. The state government agency business rules are encoded using written in OWL-2. The code, and thus the evidenced-based decision making process, is transparent to the user.

Figure 2 shows an example of the geospatial transaction model with the request to create a new road name or subdivision. The top section of the figure depicts the steps that occur in the process of creating a new road name. The bottom half of Figure 2 depicts the process when a new sub-division is requested and the steps involved in this. An example of a rule that would be executed in the case of creating a new road would be that it must connect to an existing road. Hence, within a new subdivision the road it connects to may not yet exist. The rule would continue to be executed checking if the road the new road connects to connects to an existing road. Recursively executing this rule would inform the user whether or not that proposed new road is permitted.

The concept is based on an Automated Transaction Management (ATM) approach; where the result of a transaction, such as a boundary change, results in an accurate and allowable spatial database transaction. The approach can be universally applied to other spatial transactions within the spatial data supply chain in a Spatial Infrastructure.

4. SPATIAL DATA PROVENANCE

Multiple sources of spatial datasets, an increased number of data collection authorities and owners have lead to problems of integrity, quality and trust in these spatial datasets. Data collected from several stakeholders along the spatial data supply chain has encountered several issues: schema dissimilarity, semantic heterogeneity, multiple formats, various levels of quality as well as different data collection methods and techniques.

In the spatial data supply chain for a spatial data product, the data moves through different states before being finalised in the end product. In each state, multiple geo-process may be executed upon the data according to the product requirement and business need. These processes are often exclusive to each organisation as well as the type of product being produced. The use of several tools and human involvement is imperative in delivering the agreed outcome. To establish a trust with each spatial product and to understand the spatial data's quality provenance of the data is required which facilitates the extraction of the recorded history of the processes and transformation that applied to the spatial data at the various points in the value chain.

The World Wide Web Consortium (W3C) has published a provenance model (PROV) through its provenance working group. The model encompasses Entities, Activities and Agents. It also defines that provenance is the information about the entities, activities and agents that take part in any of the process and the information about the creator and generator involved in producing a piece of information that can be used in the evaluation of the data's quality, usefulness and reproducibility.

The W3C has produced a conceptual data provenance model that encapsulates the attribution, derivation, generation and associations of all the information that occurs during any heterogeneous integration of data on a common platform, such as the Web. They have also developed ontologies for this model. The model also has different constraints on it while processing integration processes (W3C, 2013).

The Open Provenance group initiated work on an Open Provenance Model (OPM) in 2006 for scientific work flows in scientific experiments to record the lineage of results and the validation of process that have been executed as a data product is derived. In 2010 the OPM model specifications was published allowing provenance information exchange between compatible systems, giving developers guidelines for developing tools and techniques and digital representations of provenance information and its different levels (Moreau et al., 2011).

The prioritised areas of this research are the search and integration of datasets identified by the Australian and New Zealand Land Information Council (ANZLIC), as well as many issues around supply chain generation and administration. The research program has embraced advanced Semantic Web and Artificial Intelligence technologies as a means of improving spatial data supply chains (West, 2014).

However, no dedicated geospatial provenance model currently exists. The W3C PROV standard provides a model to record provenance information in a generic way. The W3C initiated an incubator process in 2009, which collected many use cases from the community, and articulated technical and usage requirements based on those use cases. The process analysed state-of-the-art provenance research and contemporary implementations, and reviewed existing provenance vocabularies. A core set of terms was recommended to represent provenance. These recommendations provided a starting point for a provenance standard for the Web, and the W3C released the PROV in 2013 for generic use (Maso et al., 2014).

In a Geospatial Web Service environment data are often disseminated and processed widely and frequently, and often in an unpredictable way. This means that it is important to have a mechanism for identifying original data sources. Geospatial data provenance records the derivation history of a geospatial data product (He et al., 2014).

A generic land administration model has been crafted for creating test ontologies. In this work different classes and sub classes have defined a workflow in a typical land administration data flow. Four major classes have been identified as super classes: capture, process, manage and disseminate. Then further sub classes have been defined with child classes and profiling performed based on categories and the nature of work as per major classes. At first instance four levels of classes have been defined and nested with each other with Entities and Class hierarchies having been developed. Disjoint classes scenarios have been captured in an attempt to define relationships between process and functions. Relationships, domains and ranges still need to be clearly implemented. During this process the data and object properties will be defined with annotation properties. This model will be added and inherited by a specialised cadastral workflow. It will also incorporate different spatial data supply chain models, which will be further discussed and normalised with different stakeholders in the coming months.

5. CONCLUSIONS

In this paper three major research projects currently occurring in Australia and related to the spatial data supply chains of Australia and New Zealand have been discussed: spatial data conflation, geo-spatial transactions and spatial data provenance. It is seen that semantic web technologies will play a key role in these projects and also in the delivery, application and usage of spatial data to assist in problem solving real world issues currently and in the future.

ACKNOWLEDGEMENTS

The work has been supported by the Cooperative Research Centre for Spatial Information (CRCSI), whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme.

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web - A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, 284(5), 34-+.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems, 5(3), 1-22. doi: DOI 10.4018/jswis.2009081901

Buneman, Peter, Sanjeev Khanna, and Wang-Chiew Tan. 2001. "Why and Where: A Characterization of Data Provenance." In Database Theory — Icdt 2001, eds Jan Van den Bussche and Victor Vianu, pp 316-330. Springer Berlin Heidelberg.

Cockcroft, Sophie. 1998. "User Defined Spatial Business Rules: Storage, Management and Implementation—a Pipe Network Case Study."

Gill, Y. Miles, S. (2013, April 30). PROV Overview, PROV Model Premier, . Retrieved 29 November, 2014, from http://www.w3.org/TR/prov-overview/

He, L., Yue, P., Di, L., Zhang, M., & Hu, L. (2014). Adding Geospatial Data Provenance into SDI-A Service-Oriented Approach. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. doi: 10.1109/JSTARS.2014.2340737

Heath, T. (2009). Linked Data - Connect Distributed Data across the Web. Retrieved 15 sep 2013, 2013, from http://linkeddata.org/home

Madden, Brian A, Ian F Adams, Mark W Storer, Ethan L Miller, Darrell DE Long, and Thomas M Kroeger. 2011. Provenance Based Rebuild: Using Data Provenance to Improve Reliability. W3C. 2015. What Is Provenance. Accessed Jan, 2015, http://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance.

Maso J., C. G., Gil Y., Prob B. (2014). OGC® Testbed 10 Provenance Engineering Report OGC Public Engineering Report (pp. 1-87): Open Geospatial Consortium.

Moreau, Luc, Clifford, Ben, Freire, Juliana, Futrelle, Joe, Gil, Yolanda, Groth, Paul, Kwasnikowska, Natalia, Miles, Simon, Missier, Paolo, Myers, Jim, Plale, Beth, Simmhan, Yogesh, Stephan, Eric and Van den Bussche, Jan (2011)The Open Provenance Model core specification (v1.1). Future Generation Computer Systems, 27, (6), 743-756, (doi:10.1016/j.future.2010.07.005).

Schmitz, P., Scheepers, L., De Wit, P., & De la Rey, A. (2007). Understanding data supply chains by using the Supply-Chain Operations Reference (SCOR) model.

West, G. (2014). Research Strategy Spatial Infrastructure, (Program 3). Updated. Retrieved 14/10/2014, from www.crcsi.com.au/Resources/Research/ P3-final-Research-Strategy.aspx

Yu, L., & Liu, Y. (2013). Using Linked Data in a heterogeneous Sensor Web: challenges, experiments and lessons learned. International Journal of Digital Earth, 1-21. doi: 10.1080/17538947.2013.839007

Zhao, T., Zhang, C., Wei, M., & Peng, Z.-R. (2008). Ontology-based geospatial data query and integration Geographic Information Science (pp. 370-392): Springer.