Spatial Data Supply Chain Provenance Modelling for Next Generation Spatial Infrastructures Using Semantic Web Technologies

Muhammad Azeem Sadiq
Department of Spatial Sciences, Curtin University
Cooperative Research Center for Spatial Information
Perth, Australia
Email: Muhammad.sadiq@postgrad.curtin.edu.au

David McMeekin
Department of Spatial Sciences, Curtin University
Cooperative Research Center for Spatial Information
Perth, Australia
Email: d.mcmeekin@curtin.edu.au

Lesley Arnold
Department of Spatial Sciences, Curtin University
Cooperative Research Center for Spatial Information
Perth, Australia
Email: l.arnold@curtin.edu.au

Abstract—**This research addresses spatial data supply chain provenance issues using semantic Web technologies to resolve knowledge gaps when disseminating spatial data products. Two models from the World Wide Web Consortium (W3C) and the Open Provenance Group for general data on the Web do not satisfy geospatial end-user needs. The Open Geospatial Consortium (OGC) has investigated the W3C PROV model for spatial datasets. Issues identified are the lack of provenance captured at the feature and attribute level, and for time series, data set series, representation and presentation interfaces, and elements at different levels. In order to answer user queries comprehensively, a geospatial provenance model in conjunction with semantic technologies has been identified as a potential solution to increase a user's trust in datasets and processes. This is important as raster dataset provenance, time series conflation processes and incremental updates have not been addressed. This has created a critical gap between provenance currency and the believability of geospatial datasets.**

Keywords-spatial data supply chain; spatuial data provenance; semantic Web; ontology; trust; processess and services.

I. INTRODUCTION

This research focusses on the needs of next generation spatial infrastructures. It explores different aspects of spatial infrastructures with a view to improving our understanding and management of data provenance along the spatial data supply chain, including end-user trust and believability. This research aims to produce a geospatial data provenance model called GEOPROV. It will investigate and implement semantic Web techniques to aid the user in assessing the results of comprehensive queries by linking provenance features with other information available from spatial systems. The objective is to improve the accessibility and usability of spatial data for Australia and New Zealand, in the first instance, but the techniques created will be generic and applicable to global use.

The main objectives of this research are: (1) detailed exploration of the requirements for geospatial provenance models; (2) development of a comprehensive spatial data supply chain provenance model for spatial information that is applicable to all feature types of spatial datasets including vector and raster datasets; (3) exploration and development of techniques to present provenance information to users for their assessment in an understandable form, via geospatial interfaces; and (4) exploration, development and enhancement of the proposed model through real case studies from relevant industry collaborators.

This paper is organized as follows: Section II describes the purpose of work performed; Section III identifies and discusses different provenance models and current work on provenance. In Section IV, the importance of work is followed by a detailed research methodology in Section V. In the last section, current findings, open issues and future directions are discussed.

II. BACKGROUND

The Cooperative Research Centre for Spatial Information (CRCSI) Program 3, Spatial Infrastructures, seeks to improve the organization, access and use of spatial data in Australia and New Zealand [1]. The research program has embraced advanced Semantic Web Technologies and Artificial Intelligence as a means of improving spatial data supply chains [1].

III. CURRENT RESEARCH

A. Current provenance models

The Open Geospatial Consortium (OGC) and the World Wide Web Consortium (W3C) define the provenance of spatial data as "information on the place and time of origin or derivation or a resource or a record or proof of authenticity or of past ownership. The W3C PROV model is a generic provenance information standard" [2].

W3C PROV is a conceptual model for provenance that offers an elegant and flexible solution for linking provenance information to geospatial elements with the necessary semantics. It can be realized in RDF, XML, and text formats, giving alternative options for implementing the same model suggested by [2]. However, no dedicated geospatial provenance model currently exists. The OGC test bed 10 Cross Community Interoperability (CCI) thread has conducted provenance activities and provided guidelines to capture provenance information through examining PROV for geospatial data.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5

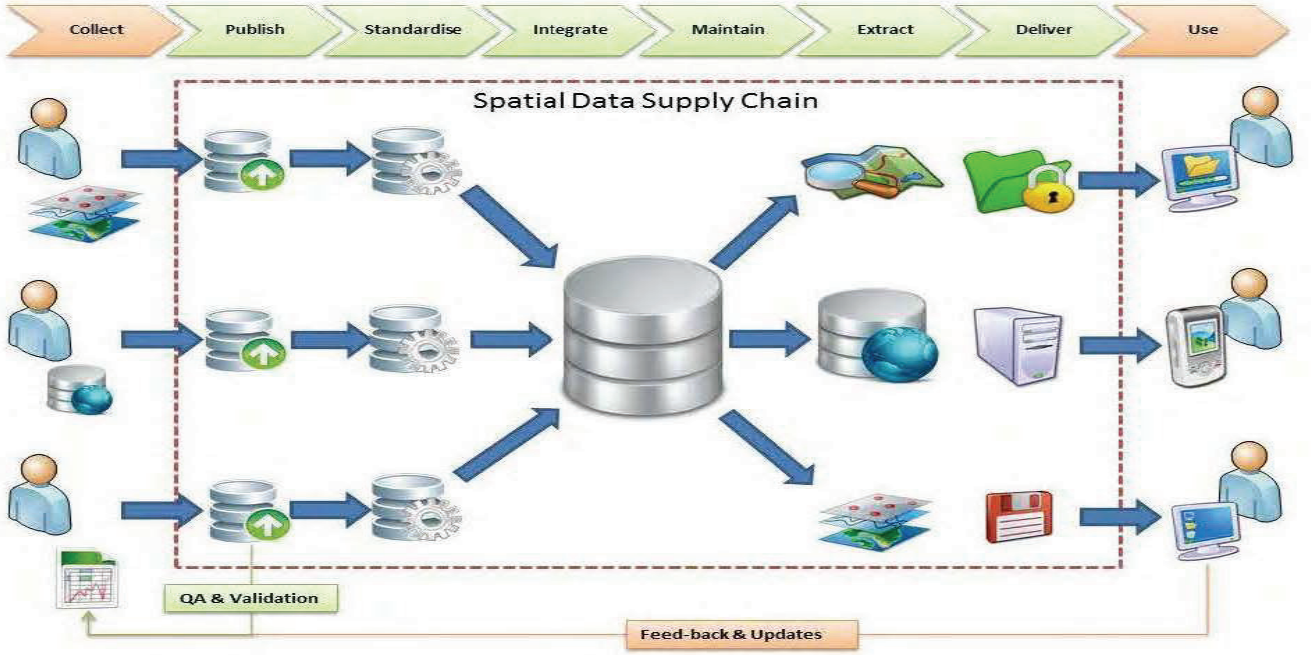


Figure 1. Spatial Data Supply Chain (van der Vlugt., 2012) [4]

B. Contemporary Research

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 in [3]. It is important for evaluating the quality of data products, tracing workflows, updating or reproducing scientific results, and in evaluating geospatial data products' reliability and quality. As a consequence, geospatial data provenance has become a fundamental issue in establishing Spatial Data Infrastructures (SDIs).

The exchange and sharing of geospatial data provenance in a distributed information environment requires an interoperable model for provenance in [3]. The rationale for designing the provenance model in this way is to combine the W3C PROV with the ISO 19115 metadata standards. This will enrich the model with domain specific details and allow domain specific representations to be translated into an interoperable form for exchange on the Web. He et al. in [3] argue that fine-grained provenance modelling for geospatial data could be achieved by borrowing from existing modelling approaches for geospatial data such as feature, coverage and observation.

One representation of a spatial infrastructure is a supply chain that involves processes from data collection to production. An alignment study of SDSCs proposed the model shown in Figure 1 in which a number of stages are identified. It shows the variability in data from suppliers and products used by consumers. Of importance are the feedback loops that are needed for quality assurance and performance monitoring [4].

An attempt has been made to develop standards for provenance tracking in spatial analytical workflows. A prototype using spatial weights has been developed by [5] for metadata and provenance for spatial analysis. They are currently in collaboration with other researchers to refine these standards and extend them as a broader set of spatial analytical services. In scientific workflows, the most valuable service is the automatic capture of sufficient provenance data to establish trust and potentially allow other researchers to reproduce a result. Scientific workflows have emerged as de facto models for researchers to process, transform and analyse scientific data. Workflow management systems provide researchers with many valuable and time saving features, from cataloguing, workflow activities and Web services, to visual authoring and monitoring [6].

A current CRCSI project is concerned with geocoded address optimisation. Each valid address in Australia is required to have one or more geocoded locations for emergency services, and efficient delivery of mail and other services. At a Geocoding Address Workshop (held in Canberra, July 2014), stakeholders highlighted the need for geospatial provenance for geocoded addressing spatial data supply chains.

As the products are published, managers and scientists will have easy access to consistent baseline information of Australia to holistically monitor and predict the impact of natural and man-made changes on the Australian environment. We presumed that temporal Provenance is an active research area that has generated complex findings.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5

This is because as the original data source is updated, the integrated dataset will also be updated. Similarly, an integrated dataset may be updated if a new version of the integration algorithm becomes available.

The integration process is usually re-executed or the updates may be done routinely or as required. To manage these scenarios, Harth et al. in [7] have developed different approaches to temporal provenance for geospatial data and derived integrated spatial products. Changes in spatial datasets with time are crucial as well as continuous capturing of provenance for each process.

C. Provenance in spatial infrastructure

Spatial data provenance is often difficult to trace in a spatial infrastructure. Regardless of the application domain, data are collected and manipulated by a wide range of users, with distinct interests and applications, using their own vocabularies, work methodologies, models, and sampling needs. We observed that in particular there is a huge effort to improve the means and methodologies to capture process and disseminate geospatial data. In real life situations, provenance information of geospatial data is used to decide pre-processing procedures, storage policies and even data cleaning strategies, with direct impact on data analysis and synthesis policies in [8].

The next generation of Spatial Infrastructures will need the capability to integrate and federate geospatial data that are highly heterogeneous. Adams et al. in [9] discuss new data that can come with variable, loosely defined, and sometimes unknown provenance, semantics and content. They further explain that the geospatial datasets that we might wish to combine could be highly heterogeneous. They will be represented in many forms, will have been generated by a variety of producers using different processes and may have originally been intended for purposes that are different from their present use. The explicit consideration of provenance into Spatial Infrastructures is needed because of massive datasets and complex functionality involved. Wang et al. in [10] state that Geographical Information Systems (GISs) are widely used for manipulating geographically referenced data and supporting spatial analysis and modelling.

Gill in [11] researched intelligent semantic workflows for complex computations and data processing at a large scale, providing assistance in setting up parameters and data, validating workflows created by users, and automating the generation of workflows from high-level user guidance. Harth et al. in [7] report their experiences with integrating geospatial datasets using Linked Data technologies. They describe NeoGeo, an integration vocabulary, and an integration scenario involving two geospatial datasets.

Despite significant advances in computational infrastructure, many environmental scientists are hampered by the resource intensive task required to set up their analysis process because data comes in daily from their sensors [12]. Data preparation is time-consuming: scientists (1) gather data from multiple sources and sensors, (2) clean the data, (3) normalize it so that data from different sources is represented using the same units and formats, and (4) integrate it and configure it according to the requirements of their models and simulation software.

D. The semantic Web approach

Provenance is seen as an important aspect of the Web that becomes crucial in Semantic Web research. Research described in [13] addresses the many questions raised about the Semantic Web in the context of automated applications. “Modelling Provenance of DBpedia Resources Using Wikipedia Contributions” by Orlandi et al. in [14] presents an approach for adding provenance information about the statements in DBpedia by connecting these statements to the Wikipedia edits they are derived from. This provenance information is subsequently exposed as Linked Data using several existing provenance ontologies.

The use of provenance for information is recommended by Artz et al. in [15] to support trust decisions, as is the automated detection of opinions as distinct from objective information. They provide an overview of existing trust research in computer science and the Semantic Web and argue that trust has another important role in the Semantic Web.

IV. SIGNIFICANCE

A provenance model is needed for geospatial data as no model currently exists. This research will investigate the generation of a provenance model, called GEOPROV building on the work of the W3C and the Open Provenance Group. Based on GEOPROV a comprehensive provenance application will be developed which will extract, capture and store provenance information in an intelligent way that it can be queried semantically.

V. RESEARCH METHODS

Experts from industry will be engaged as industry supervisors to assist with aligning the research with the needs and requirements of stakeholders. Workshops have been conducted with the land survey commission from the Surveying and Spatial Sciences Institute of Western Australia and in conjunction with their comments the ontology details have been created. During the ontology design, trust, quality, lineage, history and authoritative attributes of datasets have been considered as the building blocks. On the basis of these elements of provenance, different decision metrics will be built to rank and further analyse datasets for decision making processes. Progress review workshops will be conducted quarterly with stakeholders. Regular visits will be arranged and close working relationships will be maintained. Below are the major activities which are on-going:

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5

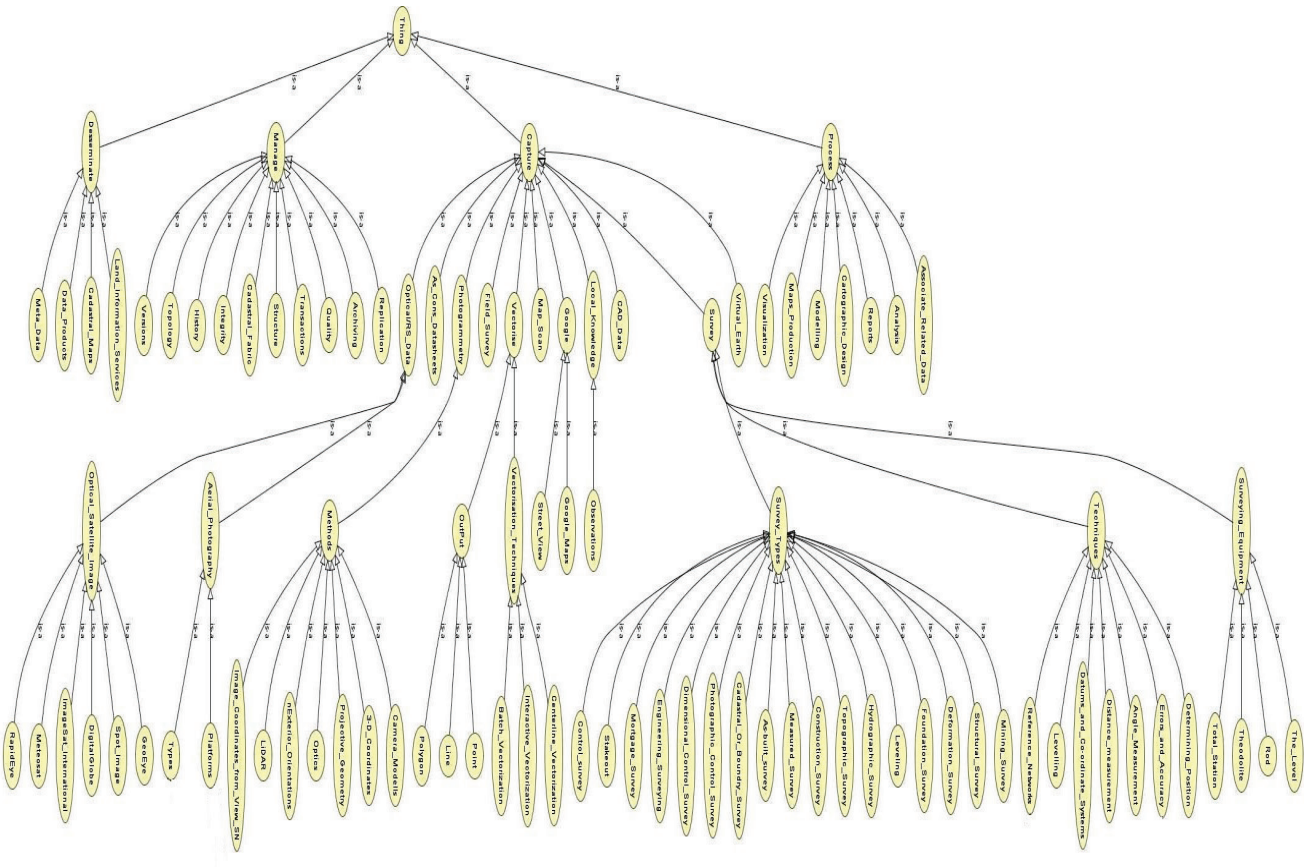


Figure 2. Land Administration Provenance Ontology Model

A. Use case development

With the close consultation of stakeholders, use cases will be developed and modelled in a comprehensive way to provide a common standard for the Australian and New Zealand geospatial industry.

Use cases are being explored with the Public Sector Mapping Agency (PSMA), Department of Land, Water and Planning (DELWP Victoria), Landgate (WA), as well as Land Information New Zealand (LINZ), and Geoscience Australia. As a result of this consultation process, the final design of GEOPROV will be developed and as a common standard.

Based on final design, GEOPROV tool will be developed to extract, store and visualize provenance of spatial datasets and will be tested by GIS teams of land administration departments across Australia and New Zealand.

B. Use case 1

A land administration subdomain provenance ontology structure has been developed. During the land administration process, data may be collected using several surveying techniques and methods. Different types of equipment are used and at various levels of sophistication. For example the popular Total Stations verses simple handheld GPSs are used to capture locations. The use of optical remote sensing techniques is also used to obtain location information that is produced by different organisations using different accuracy and modalities. For example, aerial photography is often a combination of many platforms and techniques. The nature and requirement of data capture is domain specific. Google Street View and Maps are handy sources of information for visual ground verification. All these methods, equipment and techniques are included in the ontology defined as type, resolution, calibrations, orientations, optics, geometry, combinations and principles. Capturing all these characteristics is important for determining feature accuracy and suitability for further use in the land administration life cycle (Figure 2).

C. Use case 2

Integrating road network data across State jurisdiction level may have result in anomalies due to the different standards currently used to collect the datasets. The conceptual design process is presented in Figure 3. When a linear feature road, river or any utility infrastructure are collected by different organisation and using a diverse range of standards, tools and methods, they may not be aligned at State borders. Automating such processes can provide several benefits as compared to the manual process. Matching source features with corresponding adjacent features quality may be questionable if errors or uncertainties will not be audited in post processing as in Figure 4.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5

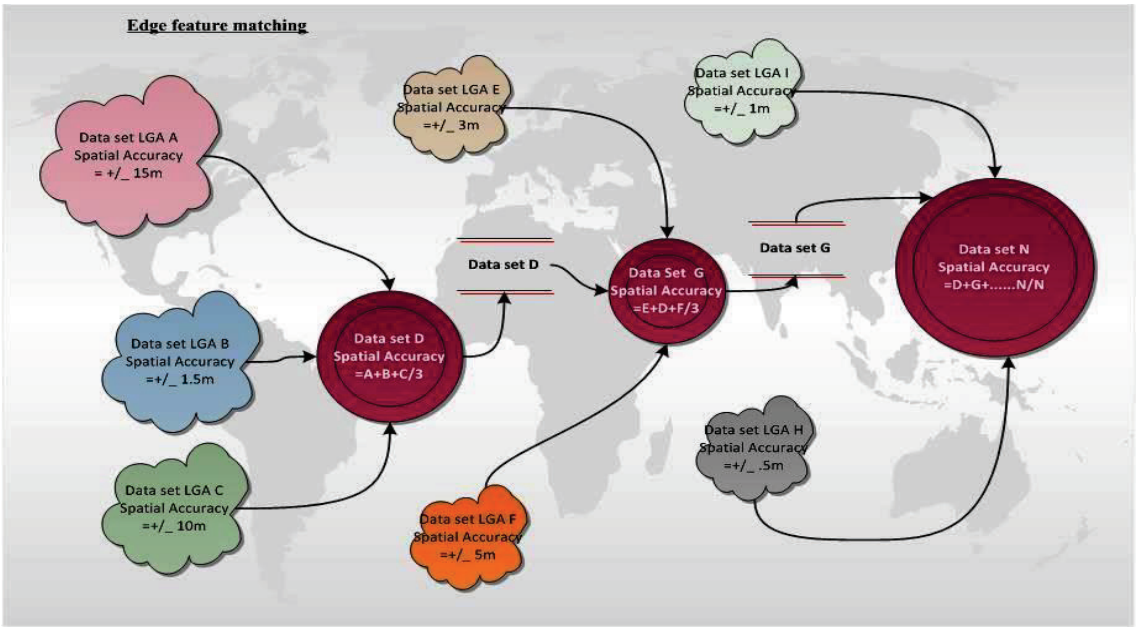


Figure 3. Edge feature matching

D. *Encoding and mapping GEOPROV*

A modular approach will be investigated initially and may be used depending upon the nature of the geospatial processes involved. Elements of the provenance model will be encoded, relationships will be defined, and geospatial data will be mapped. GEOPROV will make use of ontologies and rules. These will be explored and developed as part of this research. Open source tools such as Protégé, Pellet and others will be evaluated for usability.

W3C PROV ontology classes, properties, and constraints will be used to represent and allow the interchange of provenance information. Using this ontology, provenance records can be encoded in RDF triples. The OGC defined geospatial terms will be mapped to GEOPROV in RDF. Relationships between features, their geometric and non-geometric attributes will be defined through the latest version of an ontology Web language, namely OWL-2 and RDF.

E. *Engineering Design Experiments*

All use cases will be tested with the developed solution combined with a linked data approach. Provenance will be linked with other information available in the system to make query results more comprehensive. Research will also develop a weighted matrix approach to enable a user to determine the fitness for purpose of datasets for selected use cases. Besides this, the performance of queries will be studied as well as issues with storage, redundancy and application architectures. Geospatial provenance model requirements will be explored and defined and validated through stakeholder consultation and use cases investigated. There may be different query requirements based on the business need of each organization and governance level. These requirements will be input for testing provenance model effectiveness.

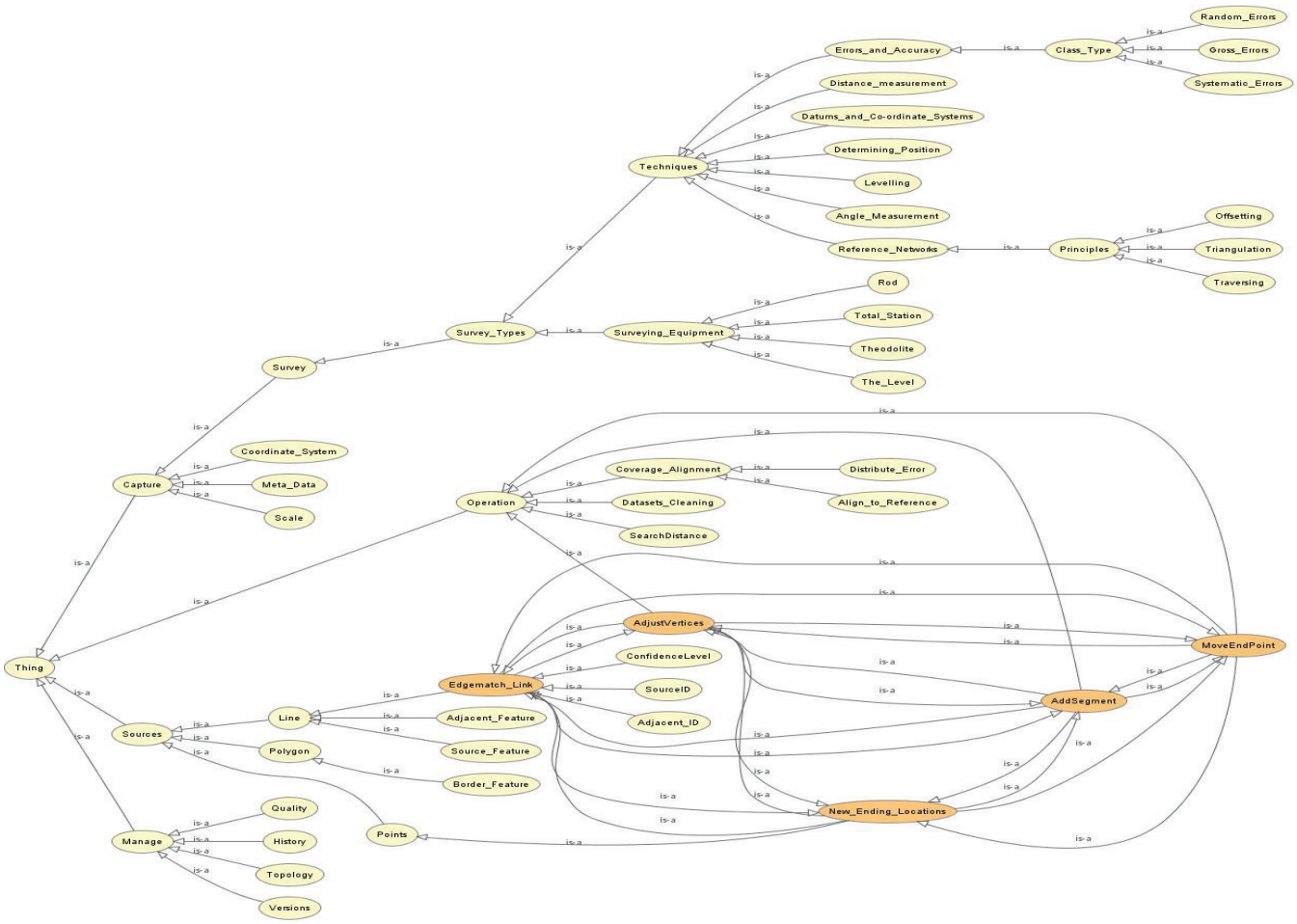
GEOPROV will be applicable to all feature types of spatial datasets including vector and raster datasets. The GEOPROV physical geospatial data provenance model will be developed in UML and Protégé along with a conceptual and functional business model applicable to all feature types of spatial datasets and different levels of granularity for geometric and non-geometric attributes including vector and raster.

VI. CONCLUSION

Spatial data supply chains (SDSC) for next generation spatial infrastructures require extensive investigation to address several contemporary issues and challenges that are hampering innovation and the use of spatial information across 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 as shown in Figure 5.

The integration and processing of multiple datasets gives rise to questions about trust, quality, fitness for purpose, currency and the authoritativeness of data. 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 data provenance and none that adequately cater for spatial information management and the dissemination of data to users.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5



Edge matching provenance ontology model

A comprehensive provenance model for the spatial domain in Australia and New Zealand is an industry imperative. 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. In addition, geospatial provenance models related to spatial data storage, scalability, robustness and query performance are yet to be examined.

Currently, GEOPROV is under development. Use cases have been produced. As result of which a Land Administration subdomain model in Figure 6 has been developed and ontology produced. The sub domain provenance model is still to be tested.

Besides this, a model for temporal and spatial provenance at feature and attribute level is under development (Figure 7). Ontologies and relationships between classes and subclasses have been defined. To achieve feature and instance level spatial provenance an edge matching line feature Web processing service is being modeled when two or more line features from heterogeneous source aligned and merged together to produce as new feature or manipulation of existing features.

This may result in the addition of new vertices, and shift vectors that may change the position of existing edges. This is a typical use case for survey data that is merged to form multiple sources to form a single cadastral dataset. The question is, what are the best techniques to enable a user to query, understand and analyse provenance information to determine trust in the data and whether it is fit for purpose? For example, a weighted matrix of provenance values may be appropriate, similar to hotel star rating. Alternatively, a user may want data at a specific accuracy and use other provenance information, such as a Web service having graphical charts for different levels of accuracy and thus trust. One representation can be the retrieval of provenance information by selecting a specific feature on the screen by querying the triples stored. The model developed will answer provenance information at feature and attribute level. For example if two features are merged, information will be captured and can be retrieved to answer queries about how the information was generated in the first instance. It will support requests about data, type of source, entities, processes, characteristics and agents.

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5

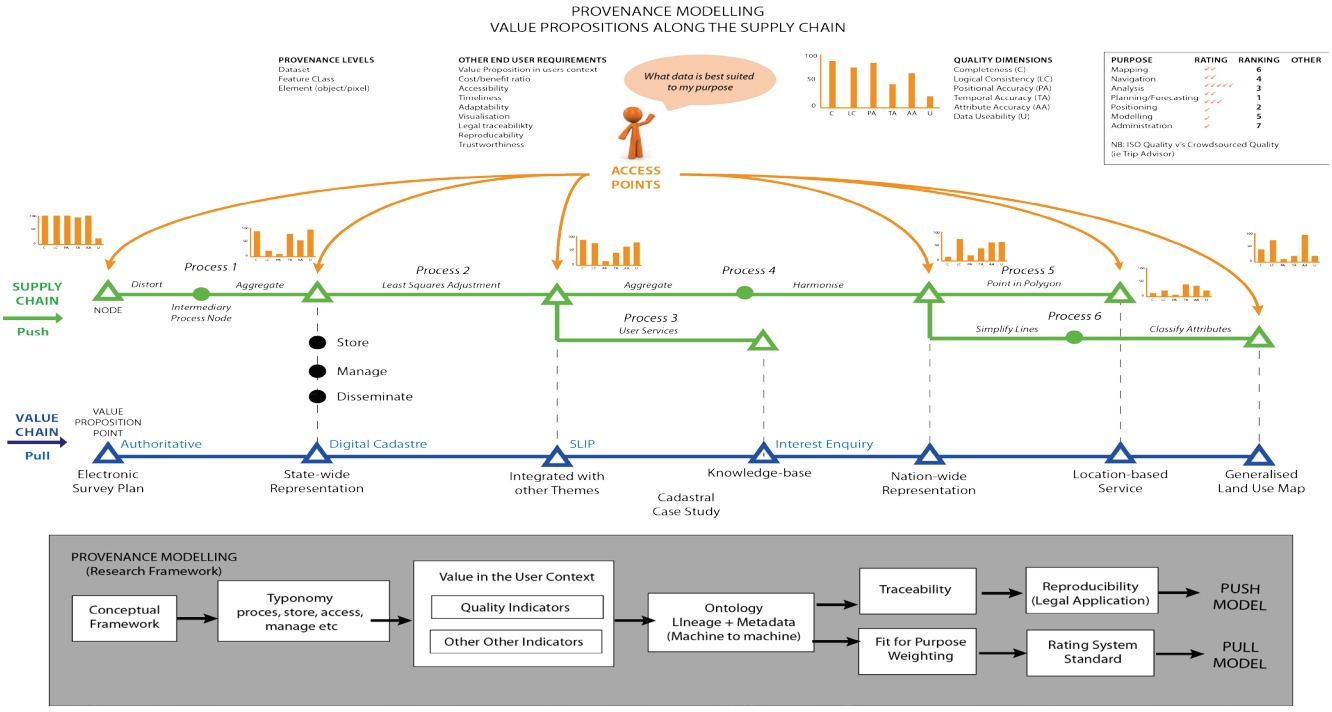
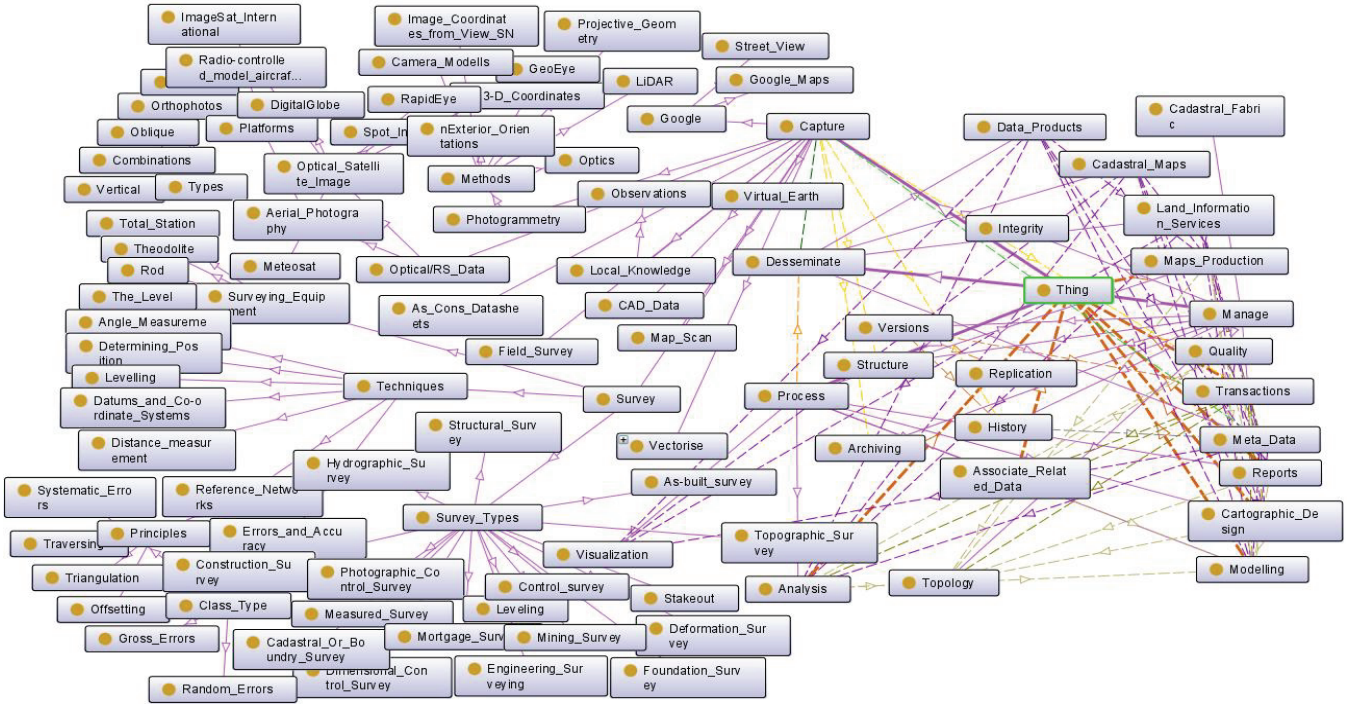


Figure 5. Supply chain provenance modeling



Use case 1 ontologies and relationships

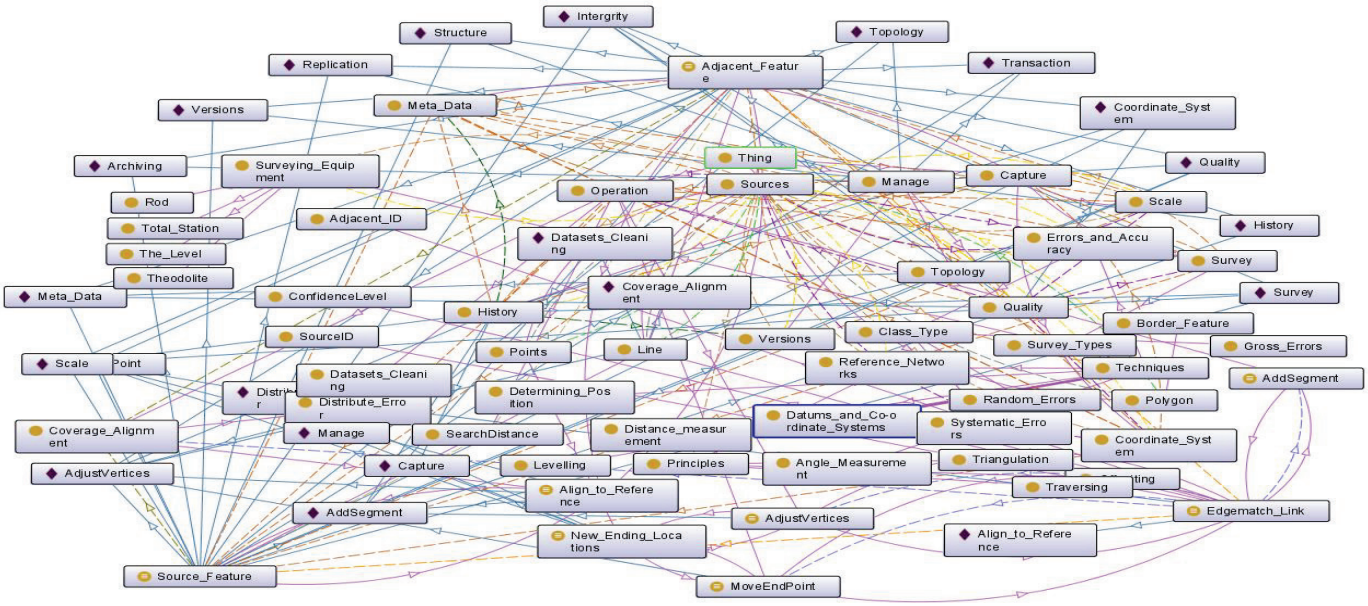


Figure7. Use case 2 ontologies and relationships

The implementation resulted from this study is a working prototype based on Python. This tool will extract metadata in XML format, it will extract properties of data sets and save them to a text files on physical location. Currently, it is extracting and collecting spatial properties of features and feature classes, their grouping schemes, data paths, counts, location, scale, data frame properties information. The next step is to capture spatial data workflow level provenance.

REFERENCES

[1] G. West, **"**Research Strategy Spatial Infrastructure, (Program 3)". Updated. Retrieved: March 2016, from www.crcsi.com.au/Resources/ Research/ P3-final-Research-Strategy.aspx

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

[3] L. He, P. Yue, L. Di, M. Zhang, and L. Hu, "Adding Geospatial Data Provenance into SDI-A Service-Oriented Approach". IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. Feburary 2015, vol. 2, pp. 926-936. doi: 10.1109/JST ARS.2014. 2340737

[4] D. McMeekin, and M. Van der Vlugt, "CRC-SI – Alignment Study of Spatial Data Supply Chains. Version 2.0". Retrieved: March 2016, from http: //crcsi.com.au/getattachment/dabdca57-d7e7-4080-97cc-36db9a89173a/. aspx

[5] L. Anselin, S. J. Rey, and W. Li,. "Metadata and provenance for spatial analysis: the case of spatial weights". International Journal of Geographical Information Science. 15 May 2014, pp.1-20. doi: 10.1080/13658816.2014.917313

[6] R. Barga, et al. "Provenance for Scientific Workflows Towards Reproducible Research". IEEE Data Eng. Bull., 2010, vol. 33(3), pp. 50-58.

[7] A. Harth, and Y. Gil, "Geospatial Data Integration with Linked Data and Provenance Tracking". Paper presented at the Linking Geospatial Data, London. 2014, Iss. 54, pp. 1-5.

[8] J. Malaverri, E. Medeiros, C. B., and R. C. Lamparelli, "A provenance approach to assess the quality of geospatial data". Paper presented at the Proceedings of the 27th Annual ACM Symposium on Applied Computing. 2012, pp. 2043-2049.

[9] B. Adams, and M. Gahegan, "Emerging data challenges for next-generation spatial data infrastructure". S. Winter and C. Rizos (Eds.), Research@Locate'14, Canberra, Australia, Vol. 1142, April 7-9, 2014, pp. 118-129. Retrieved: Feburary 2016 from http://ceur-ws.org/Vol-1142/paper13.pdf

[10] S. Wang, A. Padmanabhan, J. D. Myers, W. Tang, and Y. Liu,. "Towards provenance-aware geographic information systems". Paper presented at the Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems, Irvine, California. 2008 Nov 5, pp. 70.

[11] Y. Gill, "Intelligent Workflow Systems and Provenance-Aware Software". Paper presented at the Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.), Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California, USA. 2014, p. 91.

[12] Y. Gil, et al. "Mind Your Metadata: Exploiting Semantics for Configuration, Adaptation, and Provenance in Scientific Workflows". In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. Noy and E. Blomqvist (Eds.), The Semantic Web – ISWC 2011, Vol. 7032, pp. 65-80.

[13] Y. Gil, and P. Groth, "Using provenance in the Semantic Web". Web Semantics: Science, Services and Agents on the World Wide Web, 31 July 2011, Vol. 9(2), pp. 147-148.

[14] F. Orlandi, and A. Passant, "Modelling provenance of DBpedia resources using Wikipedia contributions". Web Semantics: Science, Services and Agents on the World Wide Web, 21 April 2011, Vol. 9(2), pp. 149-164.

[15] D. Artz, and Y. Gil, "A survey of trust in computer science and the Semantic Web". Web Semant., 15 March 2007, Vol. 5(2), pp. 58-71. doi: 10.1016/j.Web sem.2007.03.002

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-469-5