Provenance in the next-generation spatial knowledge infrastructure

I. Ivánová [a, b], K. Armstrong [b] and D. McMeekin [a, b]

[a] *Department of Spatial Sciences Curtin University, Perth, Australia*
[b] *Cooperative Research Centre for Spatial Information, Australia*
Email: ivana.ivanova@curtin.edu.au

Abstract: In quality evaluation of the piece of art its provenance (e.g. ownership over time) is often more important than the item itself. Considering that spatial data is rarely used and shared in its raw form, knowing its history (i.e. computation, transformation and other processes) may be decisive in evaluating the quality of spatial data (Buneman and Davidson, 2010). There are two concepts referring to the history of (spatial) resources on the web: lineage (defined by the International Organization for Standardization – ISO) and provenance (defined by the World Wide Web Consortium – W3C). In geospatial domain, these two concepts are widely understood and used as synonyms.

Lineage of a spatial resource (dataset or a service) is the standard term used in the spatial information domain, which is used to describe the history of a dataset and, in as much as is known, recount the life cycle of a dataset from collection and acquisition through processing, compilation and derivation to its current form (SA, 2015). Lineage metadata is only optional in the ISO-compliant standard metadata set, however, this element often appears in often sparsely populated spatial resources' metadata and, along with other metadata elements, serves as a potent vehicle for deciding on spatial resources' fitness for use. However, even in if lineage information might be present in its most exhaustive form, its main drawback for widespread automated use, is its standard data structure: more or less structured collection of the free-form text descriptions, a format unsuitable for the use in a geospatial (semantic) web.

Provenance is the standard term used in the context of the web and it is defined within a standard known as PROV as information about entities, activities, and people involved in producing a piece of data or a thing, which can be used to form assessments about its quality, reliability or trustworthiness (W3C, 2013a). PROV defines highly structured conceptual model for provenance encoding, enabling its interchange between systems and automated use on the web (W3C, 2013b).

With the aim of enabling spatial data on the web, there have been several attempts to align the standards for lineage and provenance, and these efforts demonstrate a strong convergence towards extending the current W3C provenance standard for geospatial resources. Enabling geospatial web services with provenance is paramount for reusability of spatial resources (data, information and services). In this context, there are two perspectives on provenance modelling:

- Modeling provenance of spatial resources for their discoverability and automated evaluation of spatial resources fitness for use, and
- Modeling provenance capture for automated data production and/or update.

In this paper, we review the state-of-the-art in using provenance and lineage in the spatial domain. We also demonstrate the importance and applicability of both perspectives with a use-case in which, at this point in time, we provide a partially complete solution, and look at what should be possible in current and next generation spatial knowledge infrastructure.

Keywords: *Lineage, provenance, spatial knowledge infrastructure, semantic web, fitness-for-use*

1. INTRODUCTION

The burgeoning nature of automated machine-to-machine processes within the geodomain has led to an increased emphasis on the ability to access the history of spatial information ensuring data and processes are suitable for use. There are two concepts referring to the history of (spatial) resources on the web: lineage (defined by ISO) and provenance (defined by W3C). In spatial information science and applications, lineage and provenance are often understood and used as synonyms.

In this paper we discuss the definitions of provenance and lineage (Section 2) used in the geodomain, we highlight the most important current efforts related to provenance usage in geospatial web (Section 3) and conclude with our suggestion for further use of provenance illustrated with how a real life example could manifest itself in the next generation spatial knowledge infrastructure, which is based on Linked Data and services. (Section 4).

2. LINEAGE & PROVENANCE IN THE GEODOMAIN

2.1. Definitions

Lineage of a spatial resource (object, dataset or a service) is the standard term used in the spatial information domain, and, according to the international standard ISO 19115-1 Metadata Fundamentals, it is defined as "*provenance, source and production processes used in producing a spatial information resource*", in which provenance, in line with its definition in ISO 5127 (AS, 2004) is "*information about organization or individual that created, accumulated, maintained and used records*" (SA, 2015).The ISO standard defines a models for documenting lineage for geographic vector and raster datasets (SA, 2011; 2015) and implicitly also for geospatial services (SA, 2006).

PROV, the information model defined by the World Wide Web Consortium (W3C) [1] for lineage, offers a more detailed description for resources parentage, in which *provenance* is defined as "*information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness*" (W3C, 2013). The standard defines a structure that enables provenances for automated information interchange among systems on the web.

2.2. Evolution of the terminology

W3C and ISO definitions have developed in parallel and independently, however, as some of the projects from geospatial discussed below confirm, there is a strong convergence in adopting the wider and more detailed W3C PROV model for describing the lineage of geospatial data and services. Moreover, recent web developments favor the full utilization of PROV and there is a strong inclination and need to replace web documents about data and services with data and services themselves. For the geospatial domain this means publishing data and processes with its full semantics instead of publishing and loosely coupled data and processes (web services) with their meaning (metadata services). This technological trend is confirmed and supported by one of the youngest W3C groups: Spatial Data on the Web[2], which is a partnership between Open Geospatial Consortium (OGC)[3], the main driving organization for anything 'geo-' on the internet (e.g. web map services, web feature services and web processing services) and W3C, one of the main driving organizations for anything on the web (e.g. HTML, XML and PROV).

3. USE OF PROVENANCE IN GEODOMAIN

Lineage has had a prominent role in the core metadata element set since the early versions of the ISO metadata standard and, until today, despite being defined as optional in the standard core metadata set, it frequently appears in otherwise scarcely populated metadata sets offering an information rich description of datasets' origin. The two main elements of lineage are the source and the process steps descriptions and the format of this information is less ('LI_Source' and 'LI_ProcessStep' elements in ISO 19115-1) or more ('LE_Source' and 'LE_ProcessStep' elements in ISO10115-2) structured collection of a character strings of unlimited length (SA, 2015). Altogether, this model offers an unstructured narrative to the history of the spatial resource, thereby it is unsuitable for its automated use on the web. The need for improvements of the lineage model inspired several authors in exploring alternative options to the ISO lineage format for the web. Below we provide an

[1] https://www.w3.org
[2] https://www.w3.org/2015/spatial/wiki/Main_Page
[3] http://www.opengeospatial.org

overview of most important efforts of enabling spatial data and its metadata on the web to date, focusing on attempts to align the standards for lineage (ISO) and provenance (W3C):

- OWS 9 Testbed: Cross community interoperability conflation with provenance – in this testbed authors define an architecture for a dataset conflation web processing service with capability for capturing provenance for the result at dataset and feature level. The model for the provenance is defined by the lineage model in ISO 19115, leaving the structure of both source and process step description as a free-form text hindering the automated use of dataset's provenance information for future processes (OGC, 2013).
- OWS 10 Testbed: Provenance – with the aim of adding more detailed provenance (at attribute level of geospatial objects) and looking for a more compact information structure for provenance (as opposed to ISO 19115), authors investigate and propose W3C's PROV standard for capturing provenance of automatically generated geospatial information in the web (OGC, 2014). Authors identify as an important future work a requirements analysis on provenance queries for geospatial applications; although there are several interesting ongoing research projects (see for example work of (Scheider and Ballatore, 2017 or Closa et al., 2017) on geospatial provenance, a thorough requirement analysis of geospatial provenance still remains an open problem in the geospatial domain.
- OWS 11 Testbed: Aviation data broker (OGC, 2015a) – in this testbed the authors investigate the capabilities of ISO 19115 lineage metadata element extending data brokers at dataset and the feature level. Authors recommend modeling lineage at feature level as this approach not only provides finer level of detail to lineage information, but also couples the lineage information with the feature allowing full automated exploration of features metadata and hence facilitates decisions on features fitness for use. However, as authors warn, such extension comes with significant performance costs, therefor, depending on the application context a balanced dataset and feature level lineage model needs to be considered.
- OWS 12 Testbed: Semantic portrayal, registry and mediation (OGC, 2016) – in this testbed the aim was to investigate the options for extending the capabilities of automated web services. Authors argue that current ISO 19115 document centric metadata model lacks sufficient flexibility for automated access on the web and instead recommend linked data and directed labeled graphs for modeling registries and their metadata allowing access to geospatial resources with semantic query languages (SPARQL and GeoSPARQL). Attention is drawn to the potential of semantic portrayal, mediation and registry services. Semantic registries are based on W3C's DCAT, PROV and Provenance authoring and versioning (PAV) ontology (Ciccarese et al., 2013).
- Yue et al. (2011) – OPM (Moreau et al, 2011) for provenance modelling and CSW and ebRIM registry service (OGC, 2009) for metadata catalogues for geospatial data web profile for capturing service provenance
- GeoSPARQL(2012) – an implementation standard for storing, accessing, querying and processing spatial data on the web. GeoSPARQL is robust enough to be used for 'serious' geospatial data and simple enough for linked open data (Battle and Kolas, 2012)
- He et al. (2015) – maps ISO 19115 metadata elements to PROV model, and captures provenance at dataset (for registries in a catalog web service for geospatial metadata) and feature (light-weight lineage information entity for a web feature service) level.
- GeoDCAT application profile (EU, 2016) – in this report, authors propose a geospatial extension of DCAT (2014) for data portals in Europe through mapping of the ISO 19115 'lineage' into DCAT and extending datasets' core ISO metadata elements with conformity information using elements of PROV (W3C, 2013). In this specification the ISO 'lineage statement' becomes 'dcat:provenanceStatement' expressed in free-form text, which is still unsatisfactory for automated use of provenance information on the web.

4. USE-CASE: MODELING PROVENANCE FOR NEXT GENERATION SKI

To assist people in everyday decision-making and problem solving, the next generation Spatial Knowledge Infrastructure (SKI) expands the current spatial data infrastructure model to a model that creates an autonomous network of data, analytics, expertise and policies that assists end users to integrate spatial knowledge in real-time (CRCSI, 2017). In such an SKI, data and knowledge must be exposed for the semantic web by which the data and processes become discoverable by search engines and can be queried by dedicated machines. An example use of the next generation SKI is a spatial search, which inasmuch as it seems to be trivial is currently not fully functional – current search engines allow only limited search against bespoke datasets. For instance, openstreetmap.org allows searching for spatial objects by their location expressed by geographic coordinates or an address, but does not allow search by any other (spatial) characteristics defined for the object. Current

trends in spatial search are towards enabling natural language queries on spatial objects (see for example (Ivánová et al., 2013 or Reed et al., 2016).

We illustrate the utility of the PROV model for next generation SKI spatial search targeting the rich definition of spatial objects with example query:

'*Where are the properties for sale with no water restrictions in South Perth?'*

Response to such a question should be based on the users' needs and in this example will be a list of relevant spatial objects (`lg:RESULT` entity with trust score: `4.97` on Figure 2). In this type of search the response is often not readily available (e.g. as map, coordinates or address of the property). In Australia, public data, including spatial data, should be available for web access as open linked data, therefore, constructing a response for such a complex spatial query should be possible without pre-defining the queries and datasets.

Our example was inspired by the recent discussion on provenance based assessment of data for reuse presented by Car (2016). There are several fitness-for-use related aspects illustrated in the example, which are out of the scope of this paper, such as the natural-language query processing and the trust computation. For more detail on these aspects we refer to the work of (Ivánová et al., 2013 or Reed et al., 2016) and (de Nies et al, 2013) respectively. In our example we use the PROV diagramming style (W3C, 2013) for relevant provenance elements together with additional elements and namespaces as displayed on Figure 1.

Figure 1. Ontology classes and namespaces used in Figure 2 and Figure 3

To keep the example simple and illustrative, we made the following assumptions and idealizations:

- Data and processes exist in RDF format.
- The user has a trust model to determine their list of trusted geo providers, real estate agencies and license types, with trust scores from 1(worst) to 5 (best). Several options may exist on how to model trust information in our example on Figure 2 and Figure 3 trust is only illustrative and the resulting trust information is calculated as simple average from trust values of all involved workflow elements. It is noted that all elements of the metadata can be used to determine trust.

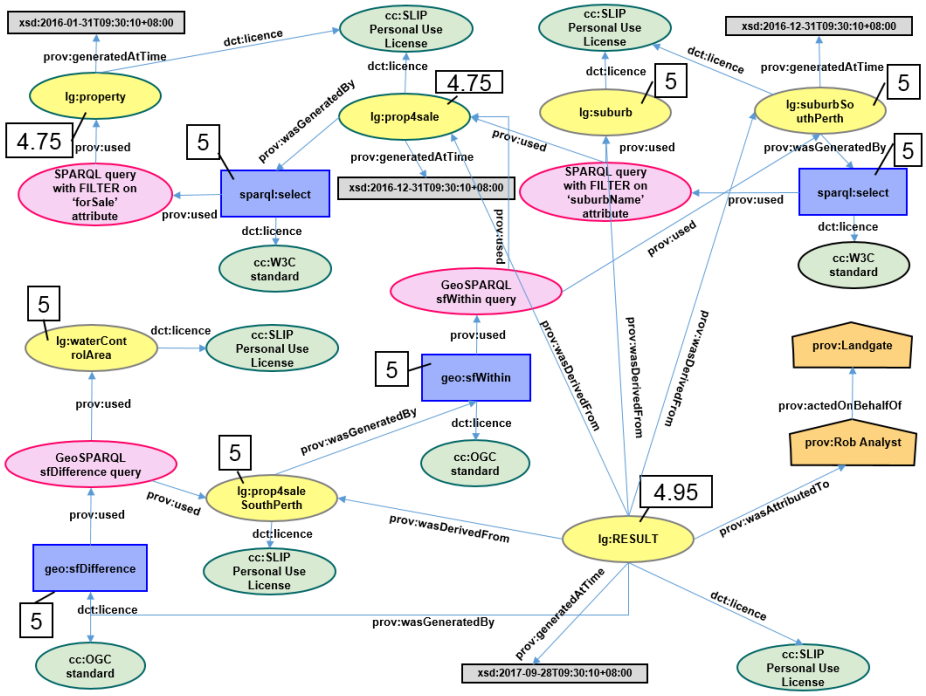

Figure 2. Producing a result for spatial search request – provenance at workflow and dataset level

Provenance and trust can be modelled at the dataset level (see Figure 2), as well as at the feature level (see Figure 3). The different trust scores for source data (`lg:property` and `lg:suburb`) used in the example on Figure 2 are explained in Figure 3 – `lg:property` is an entity that has attributes with values coming from different organizations – values for '`ID`' and '`geometry`' are provided by agent '`Landgate`' and value for '`forSale`' attribute is provided by the agent '`Reiwa`' (a real estate aggregation company), hence the different trust score for the resulting `lg:property` entity.

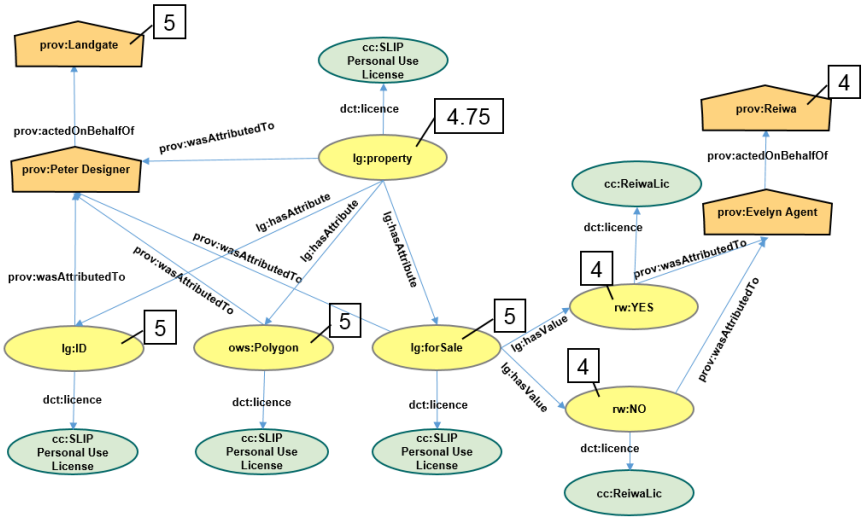

Figure 3. Producing a result for spatial search request - provenance at feature level

Modelling provenance as presented in the example above allows for automatically determining fitness for use of entities, processes and agents in future workflows. Table 1 contains some examples of queries that could be issued (by human or by machine) to provenance enabled spatial resources.

Table 1. Examples of queries on provenance

Query		**PROV elements in the search condition (i.e. the SPARQL 'WHERE' clause)**
1	*Which entities are produced by or derived from Landgate data?*	`prov:Agent` has value '`Landgate`' OR `prov:actedOnBehalfOf` has value '`Landgate`' OR entity `prov:wasDerivedFrom` is an entity which `prov:wasAttributedTo` OR `prov:actedOnBehaldOf` an agent with value '`Landgate`'
2	*Which processes operate on OGC compliant simple feature geometry?*	OGC compliant geometries (and not only those used in are example) are defined in the 'ows: http://www.opengis.net/ont/sf#' namespace. First we verify if this namespace is used in the ontology and if it is, then we can find processes, which used these geometries with search on `prov:used` an entity which '`lg:hasAttribute`' regular expression filter on values containing '`ows:`'
3	*Which entities were generated by OGC compliant processes?*	`prov:wasGeneratedBy` activity having `dct:license` with value '`OGCstandard`'
4	*Which entities were generated in 2016?*	`prov:generatedAtTime` with regular expression filter on values containing '`2016`'
5	*How current is the result?*	`prov:generatedAtTime` on the result (`lg:RESULT`) and on the source entities (`lg:prop4sale`, `lg:suburbs`) and compare with the requirement on currency
6	*Has the result been derived from entities with*	`prov:wasDerivedFrom` an entity with trust score <3, e.g. stored as entities attribute .

	low (< 3) trust score?	

From examples in Table 1, we select the Query 5 *How current is the result?* as the query, which demonstrates the potential of PROV based metadata in search for spatial data resources. In our example, the search result (`lg:RESULT`) was generated on '`xsd:2017-09-28T09:30+8:00`' from datasets `lg:suburb` and `lg:prop4sale`) generated on '`xsd:2016-12-31T09:30+8:00`', i.e. from data sources almost nine months old. In case there was any update on the '`lg:forSale`' status of a property, the result obtained with the process illustrated on Figure 2, will not be valid anymore. However, once the provenance of the search process is registered and documented with PROV and any updates on the source data are also registered with PROV, then rules on the generation time of its elements can be introduced to the search. For instance, if the last update on the '`lg:forSale`' status happened on '`lg:property`' at '`xsd:2017-03-31T09:30+8:00`' than '`prov:generatedAtTime`' on '`lg:prop4sale`' cannot be earlier than '`xsd:2017-03-31T09:30+8:00`'.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we discussed the provenance of spatial resources: its definitions and current use. From the examples in Section 3, there seems to be a consensus in adoption of PROV for modeling lineage for spatial resources. We believe, that the next generation spatial knowledge infrastructure will benefit from this trend and in Section 5 we illustrate the great potential of PROV in a typical example of search for spatial information.

There are several challenges in automated use of provenance. For example, extraction of the context dependent relevant search parameters from a natural language query is not a straightforward exercise (Ivánová et al., 2013 or Wilson et al, 2011. Moreover, attributing trust to often uncertain agents, entities and activities in a workflow requires formal trust theory and an extension of the current provenance model (de Nies et al, 2013). We did not discussed these two aspects in this paper, however, we will revisit them as part of our future work.

We are currently working on a prototype for leveraging the full potential of PROV for automated use in which we investigate how to reason with provenance on resources' fitness for use, how to automate derived provenance of results for their future reuse. Our longer term goal is further developing models for both, retro- and prospective provenance for spatial data supply chains and for provenance-based reasoning on spatial data resources' fitness for use. These efforts are part of recently approved testbed at Cooperative Research Centre for Spatial Information and, although our work is in its initial stage and there are no outputs to share, we plan to demonstrate some of our results in the very near future.

ACKNOWLEDGMENTS

This work has been supported by the Cooperative Research Centre for Spatial Information, whose activities are funded by the Business Cooperative Research Centres Programme.

REFERENCES

Battle, R. and Kolas, D. (2012) Enabling the geospatial web with parliament and GeoSPARQL, *Semantic Web*, 3(4), http://www.semantic-web-journal.net/sites/default/files/swj176_3.pdf (accessed 25 July 2017)

Buneman P. and Davidson, S. (2010) Data provenance – the foundation of data quality, Workshop: Issues and Opportunities for Improving the Quality and Use of Data within the DoD, Arlington, USA, 26-28 October, 2010, http://www.sei.cmu.edu/measurement/research/upload/Davidson.pdf (accessed 29 September 2017)

Car, N. (2016) Data reuse fitness assessment using provenance, SciDataCon 2016: Advancing the frontiers of data in research, Denver, USA, 11-13 September 2016, http://www.scidatacon.org/2016/sessions/53/paper/47/ (accessed 31 July 2017)

Ciccarese, P., Soiland-Reyes, S., Belhajjame, K., Gray, A.J.G., Goble, C. and Clark, T. (2013) PAV ontology: provenance, authoring and versioning, *Journal of biomedical semantics*, 4(37), doi:10.1186/2041-1480-4-37

Closa, G., Masó, J., Julià, N., Pesquer, L. and Zabala, A. (2017) Web processing service to describe provenance and geospatial modeling, Paper presented at GEOProcessing 2017, The Ninth International Conference on Advanced Geographic Information Systems, Applications and Services, Nice, France, 19-23 March 2017

Cooperative Research Centre for Spatial Information – CRCSI (2017) Towards a Spatial Knowledge Infrastructure, White paper, 26p. http://www.crcsi.com.au/spatial-knowledge-infrastructure-white-paper/ (accessed 31 July 2017)

European Union – EU (2016) GeoDCAT-AP: A geospatial extension for the DCAT application profile for data portals in Europe, Version 1.0.1, 87p, https://joinup.ec.europa.eu/node/154143/ (accessed 25 July 2017)

He, L., Yue, P. and Di, L. (2015) Adding geospatial data provenance into SDI – A Service-oriented approach, *IEEE Journal of selected topics in applied Earth observation and remote sensing*, 8(2), 926-936

Intergovernmental Committee on surveying and mapping – ICSM (2015). Cadastre 2034: Powering Land and Real Property (Cadastral reform and innovation for Australia – A National strategy), 36p, http://www.icsm.gov.au/cadastral/Cadastre2034.pdf (accessed 31 July 2017)

Ivánová, I., Morales, J., de By, R.A., Beshe, T.S. and Gebresilassie, M.A. (2013) Searching for spatial data resources by fitness for use, *Journal of Spatial Science*, 58(1), 15-28

Moreau, L., Clifford, B., Freire, J., Futrelle, J., Gil, Y., Groth, P., Kwasnikowska, N., Miles, S., Missier, P., Myers, J., Plale, B., Simmhan, Y., Stephan, E., and van den Bussche, J. (2011) The Open Provenance Model core specification (v1.1), *Future Generation Computer Systems*, 27 (6), 743-756

de Nies, T., Coppens, S., Mannens, E. and van de Walle, R. (2013), Modeling uncertain provenance and provenance of uncertainty in W3C PROV, in WWW'13 Companion, Proceedings of the 22nd International conference on world wide web, Rio de Janeiro, Brazil 13-17 May 2013, 167-168

Open Geospatial Consortium – OGC (2009) CSW-ebRIM Registry Service - Part 1: ebRIM profile of CSW, OGC® Implementation Standard, 45p, http://portal.opengeospatial.org/files/?artifact_id=31137 (accessed 31 July 2017)

OGC (2012) – GeoSPARQL – A Geographic query language for RDF data, OGC® Implementation Standard, 57p. http://www.opengis.net/doc/IS/geosparql/1.0 (accessed 31 July 2017)

OGC(2013). OGC® OWS-9 Cross Community Interoperability (CCI) Conflation with Provenance Engineering Report, OGC® Engineering Report, 2013, 37p, www.opengeospatial.net/def/doc-type/per/cci-conflation-provenance (accessed 31 July 2017)

OGC (2014). OGC® Testbed 10 Provenance Engineering Report, OGC® Engineering Report, 2014, 79p, http://www.opengis.net/doc/ER/testbed10/provenance (accessed 31 July 2017)

OGC (2015a). OGC® Testbed 11 Data broker specifications, OGC® Engineering Report, 2015, 36p, http://docs.opengeospatial.org/per/16-045r2.html (accessed 31 July 2017)

OGC (2015b). OGC® WPS 2.0 Interface standard, 2015, 133p, http://docs.opengeospatial.org/is/14-065/14-065.html (accessed 31 July 2017)

OGC (2016). OGC® Testbed 12 Semantic portrayal, registry and mediation, OGC® Engineering Report, 2016, http://docs.opengeospatial.org/per/16-059.html (accessed 31 July 2017)

Reed, T. W., Gulland, E-K., West, G., McMeekin, D. and Moncrieff, S. (2016). Geographic Metadata Searching with Semantic and Spatial Filtering Methods, In *Proceedings of the GEOProcessing 2016 : The Eighth International Conference on Advanced Geographic Information Systems, Applications, and Services*, 2016, pp.85-92, IARA: Venice, ISBN 978-1-61208-469-5

Scheider S. and Ballatore, A. (2017) Semantic typing of linked geospatial workflows, *International Journal of Digital Earth*, http://dx.doi.org/10.1080/17538947.2017.1305457 (accessed 31 July 2017)

Standards Australia – SA (2004) AS/NZS ISO 5127:2004 Information and documentation – Vocabulary, Standards Australia, 2004, 152p, ISBN 0-7337-6137-2

SA (2006) AS/NZS ISO 19119:2006 Geographic information – Services, Standards Australia, 2006, 72p, ISBN 0-7337-7258-7

SA (2011). AS/NZS ISO 19115.2:2011 Geographic information – Metadata Part 2: Extension for imagery and gridded data, Standards Australia, 2011, 44p, ISBN 978-1-74342-001-0

SA (2015). AS/NZS ISO 19115.1:2015 Geographic information – Metadata Part1: Fundamentals, Standards Australia, 2015, 167p, ISBN 978-1-74342-970-9

Wilson, G., Devillers, R., & Hoeber, O. (2011) Fuzzy logic ranking for personalized geographic information retrieval. In: Kudělka, M., Pokorný, J., Snášel, V., & Abraham, A., eds. *Proceedings of the Third International Conference on Intelligent Human Computer Interaction (IHCI 2011)*, Prague, Czech Republic, August, 2011, Springer-Verlag, 2013, 111–125.

W3C (2013a). PROV-Overview, An overview of the PROV family documents, W3C Working Group Note, https://www.w3.org/TR/prov-overview/ (accessed 31 July 2017)

W3C (2013b). PROV-DM: The PROV Data Model, W3C Recommendation, https://www.w3.org/TR/2013/REC-prov-dm-20130430/ (accessed 29 September 2017)

W3C (2014). Data Catalog Vocabulary (DCAT), W3C Recommendation, https://www.w3.org/TR/vocab-dcat/ (accessed 31 July 2017)

Yue, P, Wei, Y., Di, L., He, L., Gong, J. and Zhang, L. (2011) Sharing geospatial provenance in a service-oriented environment, *Computers, Environment and Urban systems*, 35(4), 333-343