Modelling data provenance of road network data produced using Semantic Web Technologies

Michael G. Niestroj[1(✉)], Ivana Ivánová[2], David A. McMeekin[2], and Petra Helmholz[2]

School of Earth and Planetary Sciences, Discipline of Spatial Sciences, Curtin University, Perth, Australia
[1] Michael.Niestroj@postgrad.curtin.edu.au,
[2] {Ivana.Ivanova, D.McMeekin, Petra.Helmholz}@curtin.edu.au

Abstract. The national road transport and traffic authority of Australia, Austroads, has identified a lack of data harmonisation within the jurisdictional agencies, as each agency uses their own data standards and software. Working with road network data is a challenging task as data sets usually do not share the same origin, resulting in a heterogeneous data structure. For instance, the problem exists while working with large road networks across state borders that road authorities change and state connecting road segments appear multiple meters apart due to different possible factors: change of data capturing hardware, data collection method, surveyor and coordinate reference system differences.
In the data harmonisation context, it has been identified that road centrelines provided by different data sources do not share same features although they describe the same asset. Therefore, a developed road network translation algorithm (translating geographic coordinates from road nodes and intersections to a trusted source) will be applied on a selected road network selection, creating new features to enable the identification of how far these road network representations are apart.
Previous work developed a road network representation with ontologies, relations and semantic rules to contribute as a first step towards road network data harmonisation. The use case of this paper is the contribution into road asset data harmonisation by modelling data provenance regarding these ontologies. This is a fundamental development as it helps to understand the origin of data, what the current state is and how the data was processed.

Keywords: Data Provenance · Semantic Technology · Road Assets

1 Introduction

Although the Semantic Web was introduced about two decades ago [2], the potential of using it for road asset data has not been identified by road authorities in Australia. The ongoing problem is that Australia's road authorities use their own data standards and software systems [6]. For instance, road network data is often provided by multiple authorities (such as in Western Australia by Main

Roads Western Australia [MRWA] and the Western Australian Land Information Authority [Landgate]), and, therefore a richer dataset can be provided with Semantic Web Technologies enabling metadata access of all available data sources for a specific road asset.

A first step to an unified road network representation has been identified by Austroads (the Australian main organisation of road transport and traffic agencies) in a business case developed in 2014 that describes the need for standardised and harmonised road asset data in Australia and New Zealand [1].

Data harmonisation is a challenging task as many data aspects (e.g. completeness, life span, origin, quality and use case) need to be considered to provide a unified valid data specification. Semantic Web Technologies have the potential to fill the data harmonisation gap while providing an unified language to connect data with same meanings that do not share metadata identifiers (e.g. object id and name). For example, Cunningham et al. [3] describes nine principles of semantic harmonisation: semantic harmonisation is different from technical harmonisation (all data is available on compatible platforms); distinguish between local data (measurement methods and representation) and global data (used for analysis); separate vocabulary from the structure; use when possible available standard vocabularies; define rules to map between knowledge objects; limit rules to be linked to a single knowledge object; integrate security while limiting data access and user permissions, and trace derived information with provenance; distinguish between how measurements are taken and what the data use case the data have; and find a balance between generic and specific descriptions.

Provenance information provides context and motivation which has lead to the production at hand. More specifically, provenance informs users of data about the origin, updates, handling, use cases, validations and life span. The World Wide Web Consortium (W3C) defined *PROV-O*, an ontology that links the W3C's provenance data model to the Semantic Web using the Web Ontology Language (OWL) 2. *PROV-O* uses three base classes (entities, activities and agents) to define the provenance of the data [5]. Many working groups are exploring the use of data provenance in the geospatial domain [4, 7, 10–12]. For instance, Sadiq et al. [7] developed a data provenance model regarding land management dataflow management systems, Yue et al. [10] provided elementary research into tracking geospatial metadata provenance before the availability of *PROV-O*, and Yue et al. [11] describes how the Open Geospatial Consortium (OGC) Catalogue Service for the Web (metadata to describe geospatial data) can be enriched with geospatial provenance regarding data discovery, service and knowledge level domain.

This paper proposes a data provenance model for an existing road network conflation framework based on ontologies and semantic rules defined for the road network. The previously developed translation algorithm will be improved in this paper to be applicable to a large-scale road network selection. The developed data provenance model uses *PROV-O* to describe the provenance of applied translation methods (selected by the algorithm) and keeps track of data creation, origin and related geographic locations (points, lines and multilines).

The structure of this paper is as follows. The selected data sources will be presented in Section 2. Then, in Section 3 methods of the approach will be explained focusing on the data provenance model, provenance data creation and the translation algorithm. After that, the results of Section 3 will be evaluated in Section 4 while measuring the computation time regarding data creation and ontology reasoning. An evaluation of the translation algorithm and a practical example of the developed provenance model will be shown, followed by our concluding thoughts in Section 5.

2 Data

The data sets from MRWA and Landgate presented in this paper are quality controlled as provided by public authorities and taken from the Western Australian government data portal (https://data.wa.gov.au). The MRWA data contain intersections and road nodes. The Landgate data contain road nodes, roundabouts and roundabout connectors (nodes within roundabouts) and is available from the Landgate Shared Location Information Platform (SLIP) data with rich metadata information (e.g. MRWA road number, road access right and data capture method), as well as Landgate simplified data (publicly free available with basic metadata). Road network data of both Landgate data sets (SLIP and simplified) share the same coordinate features. The difference between the Landgate SLIP and simplified data sets is that the SLIP data describes a road with many road nodes separated at intersections, whereby the simplified data uses a single data entry for the representation of a road regardless of intersections. OpenStreetMap (OSM) data that have been used in the ontology development will, therefore, be mentioned in minor parts of this paper.

The high-resolution aerial images used to compare the road network data to the road network itself are based on high-quality airborne geo-referenced images provided by the EagleView company.

3 Method

The methods of this paper can be described with two parts. The first part is the development of a provenance model that is applied to MRWA and Landgate road network ontologies (see details of the previous work in [anonymised]). The second part covers the provenance ontology data creation and describes further an improved road network coordinates translation algorithm that translates road network features from MRWA to features of Landgate SLIP data, as latter is used as a trusted road network data source of this paper.

3.1 Provenance model

This section explains the developed provenance models with the help of W3C *PROV-O* ontology [9]. All developed provenance models share the use-case to

trace road asset data entries back to their data source and data creation process. A further application of the provenance models is to provide information about the road assets location described by points, lines and multilines. The provenance graphs are based on five fundamental elements listed below:

1. Activities (blue): processes/algorithms to create data entries (individuals) and indicators for the application of the translation algorithm.
2. Entities (yellow): data individuals (unique data entries), such as: road assets, locations (lines, multilines and points) and data sources.
3. Agents (orange): organisations (Landgate, MRWA and OSM).
4. Types (white): data types (e.g OGC Simple Features, provenance).
5. Properties (white): properties (e.g. *GeoSPARQL* and W3C geo attributes).

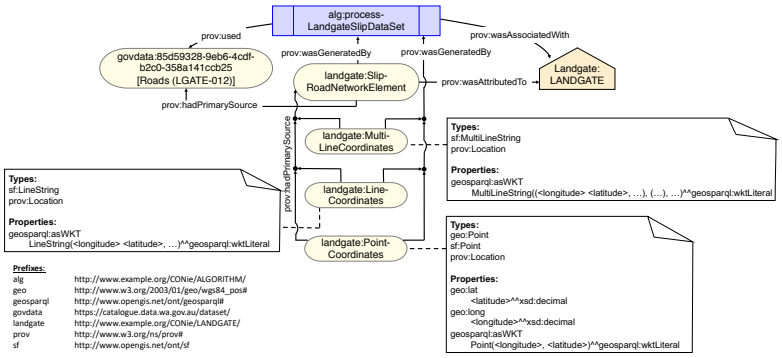

Fig. 1. Data provenance model of the Landgate SLIP data set.

The data provenance model of the Landgate SLIP dataset is indicated in Figure 1. The activity 'alg:processLandgateSlipDataSet' creates Landgate SLIP road network entities and is associated with the agent 'landgate:Landgate'. The activity uses the Landgate SLIP 'LGATE-012' data set as the primary source of the generated road network entities. Each Landgate SLIP road network entity contains a multiline, lines and points that has been extracted from the road network data set while creating the ontology. A multiline is described in the ontology as an OGC Simple Feature 'MultiLineString' and uses the *GeoSPARQL* property 'asWKT→MultiLineString' for the definition in an international standard format. Lines are extracted from multilines and point coordinates accordingly from a line. For instance, if a linestring consists of seven vertices, then each vertex will be an unique data entry (individual). A point is also described by the W3C standard (Basic RDF Geo Vocabulary [8]) for localisation using points defined by their longitude and latitude. All geographic locations (points, lines and multilines) are defined as 'prov:Location'.

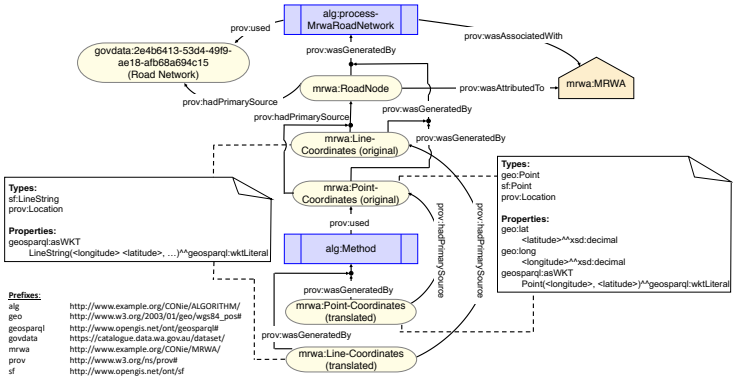

Fig. 2. Data provenance model of the MRWA road network data set.

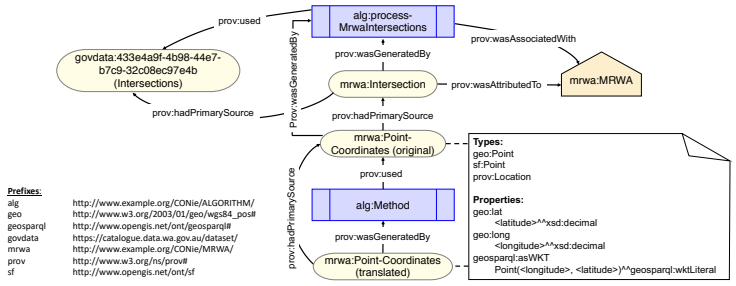

Fig. 3. Data provenance model of the MRWA intersection data set.

Figure 2 shows the data provenance model of the MRWA road network. The activity 'alg:process-MrwaRoadNetwork' creates MRWA road node individuals and is associated with the agent 'mrwa:MRWA'. The activity uses the MRWA 'Road Network' data set as primary source of generated MRWA road network entries. Each MRWA road network individual contains an original line and at least two points that has been extracted from the road network entity while creating the ontology. The difference to the above described Landgate SLIP data provenance model is that MRWA road network locations are indicated as original and translated. The translated point coordinates are generated by the activity 'alg:Method' that uses original point coordinates as the primary source. The translated point coordinates are further processed to a line composed of translated coordinates. Although a line uses translated coordinates, it has provenance to the original line coordinates indicated as 'prov:hadPrimarySource'. Lines are described in the ontology as an OGC Simple Feature 'LineString' and use the *GeoSPARQL* property 'asWKT→LineString' for a standardised representation. Points are described as OGC Simple Feature 'Point' and in addition as W3C points with 'geo:long' (longitude) and 'geo:lat' (latitude).

The data provenance model of MRWA intersections in Figure 3 is compared to the MRWA road network provenance model in Figure 2 simpler designed. The difference is, intersections are only described with points and, therefore, the provenance graph in Figure 3 does not contain lines and multilines. The activity 'alg:processMrwaIn-tersections' creates MRWA intersection individuals and is associated with the agent 'mrwa:MRWA'. This activity uses the MRWA 'Intersections' data set as primary source of generated intersection individuals. Each MRWA intersection contains a point (original coordinate) that has been extracted from the intersection while creating the ontology. The original point is used as the primary source of the activity 'alg:Method' that processes the translation algorithm and creates translated intersection coordinates. Points are described in the ontology as *GeoSPARQL* elements, OGC Simple Features, W3C points and 'prov:Location'.

Data provenance models have also been developed for other data sets, such as MRWA regulatory signs, Landgate roads simplified and OSM data. However, the structure of these models can be derived from the graphs in this section and, therefore, are not further explained.

3.2 Improved road network coordinates translation algorithm

This section describes two relevant parts regarding the coordinates translation approach of this paper. The first part is about RDF/Turtle data creation to create ontology files that include original/translated road network locations and the implementation of data provenance. The second part summarises the application of the improved road network coordinates translation algorithm.

Part 1: The approach of creating a new data sets in the RDF/Turtle file format regarding MRWA intersections is indicated in Algorithm 1. First, the *GeoJSON* data file 'MRWA_Intersections.geojson' is read as an *JSON* object into C. Then, the algorithm addresses each object as A. The unique road node object name is a construct of the string 'MRWA_Intersection', the road name and the road node object identifier.

The original coordinates and the translated coordinates of the intersection are saved into the variables X_O and X_T, respectively. If a coordinate X_O has not been written into the RDF/Turtle file, then it will be written as a data individual into the file with its attributes, such as 'prov:Location', 'prov:Entity' and 'prov:hadPrimarySource'. The attribute 'prov:wasGeneratedBy' is also connected to X_T and covers information about the activity 'alg:Method' that indicates the use of a translation method. A similar approch to write the coordinate pairs of X_O is conducted with the same attributes as before for X_T, except using other attributes for 'prov:hadPrimarySource' and 'prov:wasGeneratedBy'. Finally, the individual is written into the RDF/Turtle file with its metadata and attributes (e.g. 'prov:Entity', 'prov:hadPrimarySource intersection_data', 'prov:wasAttributedTo mrwa:MRWA', 'prov:wasGeneratedBy alg:processMrwa-Intersections' and 'mrwa:hasPointCoordinates'). The previous described ontology features are related to the provenance model in Section 3.1.

Algorithm 1: Create MRWA intersection RDF/Turtle file of a given *geoJSON* data set with original and translated coordinates.

Function `main()`:
- C ⟵ load file 'MRWA_Intersections.geojson' as JSON
- **foreach** $A \in C$ **do**
  - objectName ⟵ MRWA_Intersection + A[road name] + A[id]
  - X_O ⟵ original coordinates of A
  - X_T ⟵ translated coordinates of A
  - **if** X_T *not written in RDF/Turtle file* **then**
    - write ⟵ X_T into RDF/Turtle file
    - write ⟵ X_T is 'sf:Point'
    - write ⟵ X_T is 'prov:Location'
    - write ⟵ X_T is 'prov:Entity'
    - write ⟵ X_T is 'geo:Point'
    - write ⟵ X_T is 'geo:lat'
    - write ⟵ X_T is 'geo:long'
    - write ⟵ X_T 'geosparql:asWKT' 'Point(<long>, <lat>)'
    - write ⟵ X_T 'prov:hadPrimarySource' X_O
    - write ⟵ X_T 'prov:wasGeneratedBy' 'alg:Method [1..8]'
  - **end**
  - **if** X_O *not written in RDF/Turtle file* **then**
    - write ⟵ X_O into RDF/Turtle file
    - write ⟵ provenance respective to X_T with 'sf:Point', 'prov:Location', 'prov:Entity', 'geo:Point', 'geo:lat', 'geo:long' and 'geosparql:asWKT'
    - write ⟵ X_T 'prov:wasGeneratedBy' 'alg:processMrwaIntersections'
    - write ⟵ X_T 'prov:hadPrimarySource' 'mrwa:Intersection'
  - **end**
  - **if** *A not written in RDF/Turtle file* **then**
    - write ⟵ A with its metadata into RDF/Turtle file
    - write ⟵ A is 'prov:Entity'
    - write ⟵ A 'prov:hadPrimarySource' 'govdata:...(Intersections)'
    - write ⟵ A 'prov:wasGeneratedBy' 'alg:processMrwaIntersections'
    - write ⟵ A 'prov:wasAttributedTo' 'mrwa:MRWA'
  - **end**
- **end**

End Function

The creation of RDF/Turtle files from the data sets MRWA road network, MRWA regulatory signs, Landgate simplified data set and Landgate SLIP data follows up the logic of creating the above described intersections RDF/Turtle file. The difference is, that the data sets regarding road nodes require additional layers for lines and multilines.

Part 2: The process of applying the improved coordinates translation algorithm and to write MRWA road nodes and intersections with translated coordinates into *JSON* files is indicated in Algorithm 2. The main function summarises the algorithm. At the beginning, data individuals will be loaded from the file 'dataset_JsonLD_individuals.json' as *JSON* into D (grouped by data source). Landgate data (connectors and road nodes) and MRWA data (road nodes and intersections) are copied into F_D (grouped by road names).

It is indicated that results are saved with new features into F_D, such as translation methods, original and translated coordinates, and distances in meters that inform how far original/translated coordinates are apart. The translated road nodes and intersections are saved into 'Road_Network_MRWA.geojson' and 'Intersections_MRWA.geojson', respectively.

Algorithm 2: Apply coordinates translation algorithm and write results of translated road nodes and intersections into *GeoJSON* files.

Function `main()`:
- D ⟵ load file 'dataset_JsonLD_individuals.json' as *JSON*.
- F_D ⟵ get Landgate SLIP data (connectors road nodes) and MRWA data (road nodes and intersections) from D.
- F_D ⟵ run translation methods 1–8 and add results to F_D.
- write ⟵ MRWA road nodes with added metadata about translated coordinates in original MRWA *JSON* data format into 'Road_Network_MRWA.geojson'.
- write ⟵ MRWA intersections with added metadata about translated coordinates in original MRWA *JSON* data format into 'Intersections_MRWA.geojson'.

End Function

For the application of the translation algorithm offsets are used in the translation methods 1–8, so that only adjacent coordinates within a certain range will be considered as valid neighbours (see Table 1). These offsets are based on experience values and not associated with measurement uncertainties regarding the location accuracy of road assets. Therefore, to translate the coordinates of MRWA intersections to Landgate SLIP connectors and road nodes (methods 1 and 2), Landgate SLIP data coordinates must be within a distance of 6.00 m to the related MRWA intersection. The same offset is applied while translating MRWA road nodes to already translated MRWA intersection (method 3). If an MRWA road node has not been translated before method 4, then an MRWA road node will be translated to the nearest Landgate SLIP road node coordinate within an offset of 25.00 m. The translation methods 5–7 have an offset of 16.00 m. Translation method 5 translates not before translated intersections to surrounding Landgate SLIP road nodes. Method 6 translates MRWA intersections to the nearest translated MRWA road node or in between of an MRWA left/right carriageway, so that the position of an intersection will be in between of these two MRWA lanes. Translation method 7 is applied to translate MRWA intersec-

tions in between two Landgate SLIP road nodes, and is applied when MRWA provides for a road node one lane data, whereby Landgate SLIP data represents the same road node with two lanes. If an intersection has been not translated, or processed by the methods 6 or 7, then a related MRWA road node will be translated to this intersection within an offset of 6.20 m to enable a smooth road network representation (road nodes and intersection will be connected). These translation methods are applied to the data sets in the order of 1 to 8 and the offsets have been adjusted for the selected road network data set.

Table 1. Defined offsets of the road network translation methods.

Translation Method / Offset	**Max.**
Method 1/2: Translate MRWA intersection to Landgate SLIP connector / road node	6.00 m
Method 3: Translate MRWA road node to MRWA intersection	6.00 m
Method 4: Translate a not previous translated MRWA road node to a Landgate SLIP road node	25.00 m
Method 5: Translate a not before translated MRWA intersection to a Landgate SLIP road node	16.00 m
Method 6: Translate MRWA intersection to translated MRWA road node or in between MRWA left/right carriageway	16.00 m
Method 7: Translate MRWA intersection to centre of two Landgate SLIP road nodes	16.00 m
Method 8: Translate MRWA road node to MRWA intersection (if intersection is not translated or processed by method 6 or method 7)	6.20 m

4 Results

The results of this paper are described in three section. The first section covers the computation time comparing two road networks, generating ontology files, applying the road network translation algorithm, and measures the processing time of the *Pellet* reasoner. The second section evaluates the road network translation algorithm regarding MRWA and Landgate SLIP data. Finally, the third section covers the provenance model in Protégé. The evaluation has been done with a MacBook Pro (2017) with an 3.1 GHz Intel Core i5 central processing unit (CPU) and 16 GB 21330 Mhz LPDDR3 random-access memory (RAM).

4.1 Computation time

This project includes the use of three developed *Python* scripts for data processing. The first script 'GeoJSON to TTL' reads GeoJSON data from different data sources (Landgate, MRWA and OSM) and creates RDF/Turtle data

entries of each individual, so that the results can be inserted into the developed ontologies. Before the processing of the second script 'create individuals', the onologies (MRWA, OSM and Landgate) must be merged and saved in the *JSON-LD* (*JSON* for Linked Data) format for simpler data handling. The second script then extracts all *JSON-LD* individuals from the merged ontology and saves it as *JSON* data file. The result is used for the third script 'translate road network' that translates MRWA road network coordinates to Landgate SLIP data. The scripts running time in Table 2–4 is each taken from an average of 10 measurements.

Table 2. Running time of the script 'GeoJSON to TTL'.

Script / Data	**Landgate**	**MRWA**	**OSM**
GeoJSON to TTL (road network of 0.2 km^2)	0.09 sec.	0.05 sec.	0.02 sec.
GeoJSON to TTL (road network of 1.3 km^2)	0.96 sec.	0.68 sec.	0.49 sec.

The running time of the first script to create data in an RDF/Turtle format is indicated in Table 2. As one can see, the script processed in any case is less than one second. The MRWA and OSM data sets are faster produced than the Landgate data sets. The reason for this is related to the amount of metadata from each data set/source. For instance, OSM provides 9 different metadata fields (e.g. osm id, name, highway and man made) whereby MRWA has a count of 24 different metadata fields (e.g. road, road name, common usuage name and carriageway). Landgate provides the SLIP data set with 34 metadata fields and also represents the road network as a simplified data set. Therefore, the Landgate data is regarding to the MRWA and OSM data larger which results in a linear relation to the processing time.

The processing time of the second script to create individuals in an JSON format, and the third script to apply the improved coordinates translation algorithm are indicated in Table 3. To create individuals of the merged ontology is not time intensive. The small road network has an area of about 0.2 km^2 and contains 1173 individuals that are created in 0.12 seconds. The larger road network has an area of 1.3 km^2 and contains 7825 individuals that are created in 1.26 seconds. The improved translation algorithm creates new MRWA road node and intersection features of the smaller road network in 0.15 seconds to align with the Landgate SLIP data road network representation. The execution time of the improved translation algorithm applied on the larger road network takes 5.51 seconds. This shows that the effort increases dramatically with the count of individuals. The reason for this is that the algorithm compares each road object with surrounding objects. When possible a grouping into road names occurred to minimise the impact of not related road assets. For the work with road networks larger than 1.3 km^2 this algorithm requires to be updated as currently with 6 times more individuals the processing time increased by a factor of 36.

Table 3. Running time of the scripts 'create individuals' and 'translate road network'.

Script / Data	**Road network of 0.2 km²**	**Road network of 1.3 km²**
Create individuals	0.12 sec.	1.26 sec.
Translate road network	0.15 sec.	5.51 sec.

Table 4. Running time of reasoning road network data with Pellet.

Dataset / Reasoner	**Pellet**
Road network of 0.2 km² (original)	29.78 sec.
Road network of 0.2 km² (translated)	29.18 sec.
Road network of 1.3 km² (original)	1,224.89 sec. (20.4 min.)
Road network of 1.3 km² (translated)	2,490.82 sec. (41.5 min.)

Table 4 shows the reasoning time of *Pellet* in Protégé regarding the previous developed road network ontologies and Semantic rules (see [anonymised]). *Pellet* needs about 29 seconds to reason the smaller road network ontology. The reasoning of the larger road network has been done with a Google Cloud virtual machine with a very large RAM of 624 GB. The outsourcing to the Google machine was compulsory as with the amount of 88 semantic rules and a set of 7825 individuals the reasoning was not executable on a MacBook Pro with 16 GB RAM. After about 19 minutes the reasoning with the MacBook Pro failed with the error message 'GC overhead limit exceeded' indicating that not enough RAM is available. The reasoning with the Google hardware was done in approximately 1224 seconds (20.4 minutes) for the not translated larger road network, and 2490 seconds (41.5 minutes) for the translated larger area. While reasoning with *Pellet* with the virtual machine, Protégé used up to 520 GB RAM indicated by the Windows Task Manager. It has to be said that for the larger road network reasoning an average of seven measurements has been taken instead of ten. The reason for this was due to the limit of the free trial from the Google Gloud virtual machine service. Thereby, two of seven measurements had a running time of 68.6 minutes and 72.0 minutes, and the other five measurements were in a range of 27.6–28.6 minutes.

4.2 Road network

The results of the improved road network coordinates translation algorithm for the selected road network selection of 1.3 km^2 are indicated in Figure 4. It shows the original MRWA road network with road nodes (green polylines) and intersections (pink circles [mostly overlapped]) underneath the translated MRWA road network (black polylines) and colourful translated intersections (green, orange and red). The intersection circles have been coloured using a data filter in QGIS3

(free and Open Source Geographic Information System) that highlights intersections regarding translated distances of less than 2 m in green, between 2 m and 5 m in orange, and larger than 5 m in red. The translated MRWA data set also contains pink polylines for not translated road nodes, as well as white circles for not translated intersections.

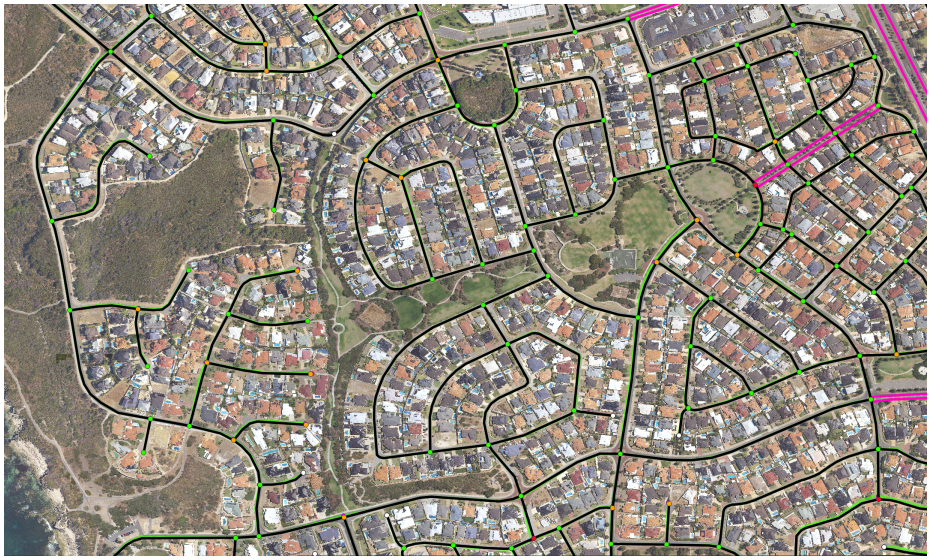

Fig. 4. Visualisation of the improved translation algorithm in an selected road network.

The road network translation of intersections are evaluated by comparing the colours of intersections. For the evaluation of the translated road nodes, the green polylines (original data) and black polylines (translated data) need to be analysed. This means, if a green polyline is visibly strong, then the related road node has been translated with a larger distance compared to a less visible green polyline. The colourful representation of translated polylines is currently not possible in QGIS3, as a polyline contains two or more vertices and each vertex point is translated individually using the best translation method regarding its relation to the trusted Landgate SLIP data set. In the future, it will be possible to represent each vertex point in an individual colour as the required metadata is available and an extended QGIS3 filter can be developed. For example, the data of a vertex point is available as: '[{ "LONGITUDE": 115.702661, "LATITUDE": -31.679257, "OLD_LONGITUDE": 115.702661, "OLD_LATITUDE": -31.679257, "METHOD": "Method 3.2: Translate Mrwa Road Node to translated MRWA intersection (translated by Method 2).", "DISTANCE_METERS": 0.043742 }, {<next original/translated point attributes>}, ...]. In the future, this data can be evaluated with QGIS3 filter queries.

Table 5. Evaluation of the improved translation algorithm.

Type	**Description**	**Colour**	**Count**
Intersection	not translated	white	7
Intersection	translation less 2.0 m	green	89
Intersection	translation between 2.0 m and 5.0 m	orange	19
Intersection	translation larger 5.0 m	red	3
Road node	translated	black	180
Road node	not translated	pink	14

After the visual inspection of the translated road network, the improved translation algorithm can be evaluated as shown in Table 5. One can see, that identical MRWA and Landgate SLIP data exists as indicated by 7 not translated intersections. An intersection translation of less than 2 meters occurred in several cases, followed by an intersection translation between 2 m and 5 m. Three times an intersection was translated with more than 5 m, that is about 2.6 % from the total of 115 intersections. The road network contained in total 194 road nodes, whereby 14 of them have not been changed. These not translated road nodes are part of roads that are represented by multiple lanes and at present are not supported. The evaluation proved that the used MRWA and Landgate SLIP data sets share, in some cases, the same features in this selected road network. However, in most cases the road network representation of both data sets do not share the same features although describing the same road asset.

4.3 Provenance

The evaluation of the developed provenance models is done with Protégé after reasoning the ontology with *Pellet* (see Figure 5). Therefore, an arbritary MRWA road node individual ('mrwa:MRWA_RoadNode_PeelWy_10626730') has been selected to present the result of the provenance model. Protégé provides a variety of information regarding the reasoned ontology while viewing the description, annotations and property assertions frames, such as ontology classes, connected road nodes, original coordinates, translation info and provenance. The reasoned information from *Pellet* is highlighted with a yellow background, and entries with a white background are part of the ontology file.

As one can see while looking into the description view, the classes 'owl:Thing' and 'prov:Entity' are part of the ontology. The reasoner added the information ':Intersection' (connected parts of an intersection), ':Road' (part of a road), 'mrwa:Geometry' (contains geometry elements, e.g. 'mrwa:hasLineCoordinates') and 'mrwa:RoadNode' (is a road node). These descriptions are overall valid for MRWA road nodes. It is recommend to implement this information during the ontology creation to reduce the reasoning time, as it is an intensive process.

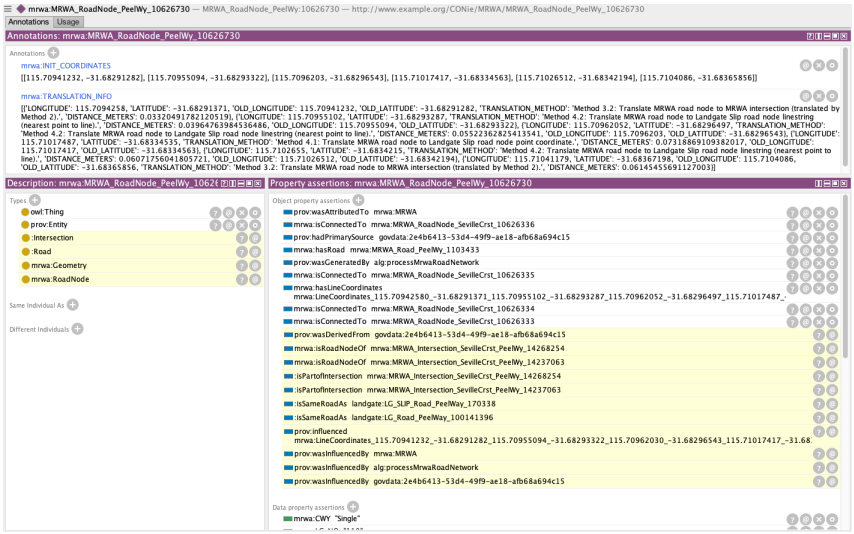

Fig. 5. Evaluation of the developed provenance model while viewing an arbitrary MRWA road node after the ontology was reasoned with *Pellet* in Protégé.

The object property assertion view can be evaluated with the same principle as above. The provenance object property assertions 'prov:wasAttributedTo', 'prov:hadPrimarySource' and 'prov:wasGeneratedBy' are part of the ontology, as well as the property assertions regarding MRWA and Landgate data, such as 'mrwa:isConnectedTo' (connected road nodes and intersections), 'mrwa:hasRoad' (part of road) and 'mrwa:hasLineCoordinates' (translated polyline of the road node). The reasoner added 'prov:wasDerivedFrom' (data source), 'prov:influenced' (original polyline of the road node), 'prov:influencedBy' (agent, data creation algorithm and data source), 'mrwa:isRoadNodeOf' (connected intersections), ':isPartOfIntersection' (connected intersections) and ':isSameRoadAs' (same road within multiple data sources). The ontology has the potential to reduce the reasoning processing time by implementing currently reasoned property assertions into the ontology creation process, such as 'prov:wasDerviedFrom', 'prov:influenced' and 'prov:wasInfluencedBy'.

The outcome of the evaluation shows that the model is well implemented and *Pellet* is able to reason the provenance model. However, the evaluation of the reasoning processing time in Section 4.1 demonstrated that ontology reasoning is an intensive process. Therefore, to prevent the reasoning time a further investigation into writing more ontology classes and provenance assertions while creating the ontologies as explained above can be conducted. The evaluation of this outsourcing can show if a better reasoning time can be achieved.

5 Conclusions

In this paper, we proposed a data provenance model for a road network translation process. The advantage of data provenance is that data characteristics, such as handling, ownership and changes can be traced back to its source. Data sets have been used from MRWA (intersections and road nodes) and Landgate (SLIP and simplified). The developed provenance models support the trace of data sources, organisations, created entities, locations and the use of a road network coordinates translation algorithm. The road network translation algorithm provenance indicated in the graph representation that a translation method has been applied but not in detail which of eight available methods. In the future, the translation algorithm provenance relations can be extended to the exact translation method with provenance about why a certain method has been chosen.

This paper showed an improved road network coordinates translation algorithm to translate MRWA data to Landgate SLIP data, as latter often showed better road centreline features in the selected road network selection and was, therefore, used as a trusted source. The road network translation algorithm was developed to indicate the differences and similarities between road authority data sets while describing the same road asset. The studying area covered a road network of about 1.3 km^2. The evaluation of the translation algorithm showed that in most cases the road centreline representation and the resulting position of intersections differs in a range of 0–2 m. The interactive inspected differences in the road centreline representation are assumed due to measurement uncertainties and errors in on-screen digitising.

This paper described the use of three different scripts: create RDF/Turtle data, extract data individuals and translate road network. The running time of each script has been evaluated comparing a road network selections 0.2 km^2 and 1.3 km^2. The results showed, except for the translation algorithm, that the script running time has a linear relation to the size of a data set. The translation algorithm processing time increased by the factor 36, although the bigger road network was just six times larger as the smaller road network. To minimise the influence of potential surrounding objects regarding a translation, the algorithm grouped when applicable road assets into road names. However, it is obvious that other techniques need to be developed as the application on a larger scale would result most likely in a exponential higher processing time.

In addition, this paper used road network ontologies and semantic rules and, therefore, the evaluation of the reasoning time of the OWL 2 reasoner *Pellet* was part of the evaluation. It was required to outsource the reasoning over the 1.3 km^2 large road network to a Google Cloud virtual machine with a very large RAM of 624 GB due to the intensive processing required. To reduce the reasoning time in the future, it is recommended to include current reasoning results (e.g. classification of road nodes, connectors, roundabouts and extended provenance [the current model writes provenance only in one direction and relies therefore on ontology reasoning]) where possible into the ontology creation approach. Each into the ontology added relation can be removed as a Semantic rule, as currently 88 Semantic rules are used to identify relations, so that an

improved reasoning time can be achieved. How much time can be saved, is part of further investigation.

The outcome of this paper can be used as base for further developments, such as the above described provenance model extension or the development of an ontology based route planing approach.

References

1. Austroads: Austroads Annual Report 2015-16. Tech. Rep. AP-C20-16, Austroads (Oct 2016)
2. Berners-Lee, T., Hendler, J., Lassila, O.: The Semantic Web: A New Form of Web Content That is Meaningful to Computers Will Unleash a Revolution of New Possibilities. Scientific American (May 2001)
3. Cunningham, J.A., Van Speybroeck, M., Kalra, D., Verbeeck, R.: Nine principles of semantic harmonization. In: AMIA Annual Symposium Proceedings. vol. 2016, p. 451. American Medical Informatics Association (2016)
4. Harth, A., Gil, Y.: Geospatial data integration with linked data and provenance tracking. In: W3C/OGC Linking Geospatial Data Workshop. pp. 1–5 (2014)
5. Lebo, T., Sahoo, S., McGuinness, D.: Prov-o: The prov ontology (04 2013), http://www.w3.org/TR/2013/REC-prov-o-20130430/
6. Niestroj, M.G., McMeekin, D.A., Helmholz, P.: Overview of standards towards road asset information exchange. ISPRS Technical Commission IV Symposium (Volume XL-4) (2018)
7. Sadiq, M.A., West, G., McMeekin, D.A., Arnold, L., Moncrieff, S.: Provenance ontology model for land administration spatial data supply chains. In: 2015 11th International Conference on Innovations in Information Technology (IIT). pp. 184–189. IEEE (2015)
8. W3C: Basic geo (wgs84 lat/long) vocabulary. (02 2006), https://www.w3.org/2003/01/geo/ (last accessed 2 April 2019)
9. W3C: Diagrams. prov graph layout conventions. (12 2012), https://www.w3.org/2011/prov/wiki/Diagrams (last accessed 2 April 2019)
10. Yue, P., Gong, J., Di, L.: Augmenting geospatial data provenance through metadata tracking in geospatial service chaining. Computers & Geosciences **36**(3), 270–281 (2010)
11. Yue, P., Gong, J., Di, L., He, L., Wei, Y.: Semantic provenance registration and discovery using geospatial catalogue service. In: Proceedings 2nd International Workshop on the Role of Semantic Web in Provenance Management, Shanghai, China. pp. 23–28 (2010)
12. Yue, P., Wei, Y., Di, L., He, L., Gong, J., Zhang, L.: Sharing geospatial provenance in a service-oriented environment. Computers, Environment and Urban Systems **35**(4), 333–343 (2011)