Data standards for the New Zealand transport network
September 2017

M Miska, Queensland University of Technology, Australia

NZ Transport Agency research report 627
Contracted research organisation – Queensland University of Technology

ISBN 978-1-98-851256-3 (electronic)
ISSN 1173-3764 (electronic)

NZ Transport Agency
Private Bag 6995, Wellington 6141, New Zealand
Telephone 64 4 894 5400; facsimile 64 4 894 6100
research@nzta.govt.nz
www.nzta.govt.nz

Miska, M (2017) Data standards for the New Zealand transport network. *NZ Transport Agency research report 627.* 43pp.

Queensland University of Technology was contracted by the NZ Transport Agency in 2016 to carry out this research.

Keywords: data model, network model, real-time operation, resilience

An important note for the reader

The NZ Transport Agency is a Crown entity established under the Land Transport Management Act 2003. The objective of the Agency is to undertake its functions in a way that contributes to an efficient, effective and safe land transport system in the public interest. Each year, the NZ Transport Agency funds innovative and relevant research that contributes to this objective.

The views expressed in research reports are the outcomes of the independent research, and should not be regarded as being the opinion or responsibility of the NZ Transport Agency. The material contained in the reports should not be construed in any way as policy adopted by the NZ Transport Agency or indeed any agency of the NZ Government. The reports may, however, be used by NZ Government agencies as a reference in the development of policy.

While research reports are believed to be correct at the time of their preparation, the NZ Transport Agency and agents involved in their preparation and publication do not accept any liability for use of the research. People using the research, whether directly or indirectly, should apply and rely on their own skill and judgement. They should not rely on the contents of the research reports in isolation from other sources of advice and information. If necessary, they should seek appropriate legal or other expert advice.

Acknowledgements

The author would like to acknowledge the input from the steering group members Russell Brown, Russell Pinchen and Ranjan Pant during face-to-face meetings and teleconferencing, as their comments and guidance enhanced the relevance of the research project for practice.

Thanks to the peer reviewers Hayley Davis, Principal Consultant at CH2M, and David Gyles, Principal Analyst at the Department of Transport and Main Roads, Queensland, Australia. Their review ensured that the project was not biased by the author's or steering group's opinions. Also, thanks to Christheo van Rooyen from Fujitsu New Zealand Limited for providing the project team with road measurement data.

Abbreviations and acronyms

API	advanced programming interface
CBD	central business district
CSV	comma separated values
GIS	geo-spatial information system
IoT	Internet of Things
ITS	intelligent transport system
JSON	JavaScript object notation
SQL	structured query language
VMS	variable message signs

Contents

Executive summary

The aim of this research was to find a practical implementation of a standardised reference network model that would support normalising diverse network representations of selected data sources, and support presentation of the information within the context of a standardised reference transport network model.

Based on an extensive review of standards and standardisation efforts around the world, this research project came to the conclusion that existing data standards only focus on either data exchange or application oriented data storage and do not support the key requirements of the New Zealand Transport Agency.

Through 'thinking outside the box' and experimenting with new ideas the project developed a reference network standard founded on the idea of utilising geometry as an identifier for road network elements as well as operational transport data. The model allows users across departments to utilise varying granularity that suits their requirements without jeopardising the consistency.

By linking data sources through geometry, the reference network allows the geo-spatial mashing of data independent of unique identifiers of any subsystem. Therefore, data governance can remain with its owner and small changes will not have an adverse effect on operations. Contrarily, it fosters a culture of building high-quality data sets by cutting down on dependencies that are most commonly the culprit for model degradation and data redundancy.

As a part of the research project work, a web-based demonstrator was developed to showcase the internals of the proposed network model. Further research could include the development of an analytics platform based on the proposed data standard with the aim of driving collaborative innovation in network analysis across authorities and academia.

Abstract

As transport network operations become more and more time critical the requirement for knowledge of the performance and state of the network becomes more significant for decision support. Sensor technology is advancing quickly, particularly with respect to the 'Internet of Things', and there are significant amounts of data and information generated by a number of internal and external sources that relate to transport network and require processing. An inherent embedded component of the data sources is information about the transport networks themselves. Typically, this spatial attribute is developed to meet the specific needs of an application leading to diverse representations of the same transport network.

There is a lack of transport specific data analysis platforms that are resilient to change, are lightweight and cost efficient to maintain. Analytical tools are often tied to specific proprietary systems that compromise adaptability and create latency in terms of keeping up with the rapidly changing developments, and indeed in representing transient changes in transport networks.

This report proposes a method that will support normalising diverse network representations of selected data sources, and support presentation of the information within the context of a standardised reference transport network model.

1 Introduction

1.1 Motivation

Progress in sensor technology provides an increasing amount of data for our transport networks every year. From remote sensing information of our infrastructure (Jin et al 2012) to sensor data from the roads and intelligent transport system (ITS) installations that provide information of movements, speeds and travel times across the network. The challenge is, how to utilise the data in a timely fashion to enhance our ability to anticipate network performance and to act accordingly, rather than being limited to analysing the performance in the aftermath. Despite the large amount of data available, operators of roads and public transport services are still struggling to get the most out of it. Real-time decision support for network operation (Casas et al 2013; 2014; Nantes et al 2016) is deployed in various places, but there is still the need to overcome data issues (Abadi et al 2015).

A key issue is the lack of standardisation in transport data, both for storing operational measurements and the transport infrastructure itself. Further, data sets describing network operation and data sets describing the transport infrastructure develop independently with vastly different speeds. While infrastructure and its attributes change slowly, the enhancements in technology drive a fast change of data in operations. To overcome the drawbacks of missing standardisation and different development speeds, this paper proposes a framework, that detaches data and analytics into two independent components that can be developed independently and communicated through a high-level programming interface. This approach enables network operators to take advantage of new technologies quickly, while maintaining infrastructure descriptions. All transport and infrastructure data is compiled into a network abstraction layer that retains geo-spatial information, focuses on vehicle movements, and is routable to support user set query parameters.

This project aimed to develop a specification for a sustainable data model for historical and real-time analysis of transport operations. It was not an effort to build a transport model for the New Zealand transport network and it included no efforts to rebuild or interfere with existing transport models in the agency. The focus was to establish guidelines/specifications towards a sustainable, high-quality, software agnostic data standard that informs business process in the mid to long term. The data standard defines the format and meaning of data stored for network operations analysis.

1.2 Status quo in network analysis and standards

Today, network analysis is highly data driven, using the vast amount of measurement data from sensors such as inductive loops, Bluetooth, cameras, OptaSense[1] and traffic control systems. As roadside measurement data and control systems by nature provide non-geospatial information, it is mapped into a pre-defined set of nodes and links, or transferred into a geo-spatial information system (GIS) for easy visualisation and further analysis. The process of mapping the data to a network is the key issue with network analysis, as there is no single network nor a single source that is the foundation for the various networks existing across departments. Once mapped, any application that supports network analysis evolves into an organically grown environment that has no links to other applications and endures high maintenance costs, as mapping tables have to be maintained and updated regularly to keep the

[1] http://optasense.com

application functioning. Those costs also lead to a lack of longitudinal datasets for network analysis, as it is only feasible to maintain the latest version of the model due to the management effort required to keep change logs around for comparing today's performance with last year's.

Networks are changing, datasets are changing and requirements for network analysis are changing. Therefore, it is necessary to establish a robust and sustainable process to create network models that adapt over time, are backwards compatible, and can cope with minor changes without disrupting day-today business for network analysts. Network models describe the transport infrastructure and associate operational data in an abstract manner to allow efficient analysis for operational decision support.

1.3 Goal

The goal of this project was to identify and address the core issues for consideration in building and maintaining a reference network model. These included but were not limited to:

- factors to be considered in maintaining the currency of the model, eg what is an appropriate granularity for defining the model
- cost implications for maintaining the model, such as managing changes
- impact on data quality, such as changes to the model over time, and effects of granularity on data stored
- fitness for purpose, eg ability to index diverse data sets without constraining them
- relationship to GIS, while being agnostic to GIS
- support for user queries of data sets, eg automatic cost base routing of the network model.

This provided many challenges that have been categorised in chapter 2 into technical and business requirements and comprise the core requirements of the data standard.

2 Core requirements

2.1 Technical requirements

This section provides an overview of the technical requirements for data management, interoperability and performance that a standard for the New Zealand transport network needs to comply with.

2.1.1 Data management

This project, while broad in its considerations, has a traffic operations focus and therefore, requires any applicable standard to deal efficiently with real-time data and to build and maintain meaningful historical data sets.

Traffic operations data includes a vast amount of information from various sources. Table 2.1 lists the minimum set of data sources to be considered with a priority rating. A high priority dataset is required to perform common traffic operations analysis; medium priority datasets add context; and low priority datasets are not necessary for traffic operations, but would allow a deeper understanding of the transport model as a whole.

Table 2.1 Considered data sources

Area	Data type	Value	Sensors/source	Priority
Operations	Point data	Speed	Inductive loop Double inductive loop Camera Radar	High
Operations	Line data	Travel time	Bluetooth tracking GPS tracking Mobile phone activity Fibre…	High
Operations	Point data	Occupancy	Inductive loop	High
Operations	Point data	Volume	Inductive loop Camera…	High
Road safety	Point data	Crash/incident	TMC report Police report…	High
Operations	Point data	PT schedule at stop	PT operators	High
Operations	Line data	PT lines	PT operators	High
Operations	Polygon data	Signal phasing and timing	STREAMS Scats…	High
Operations	Line data	Emergency vehicle paths	Emergency services GPS tracking…	Medium
Operations	Area data Line data Point data	Events Road closures…	Various	Medium
Asset	Line data	Surface condition	Asset management	Medium
Asset	Point data	Defects	Asset management	Medium

Area	Data type	Value	Sensors/source	Priority
Environment	Polygon data	Weather	Bureau of Meteorology	Medium
Environment	Polygon	Water levels (flood)	Bureau of Meteorology	Medium
Planning	Polygon data	Digital cadastre	Various	Low
Planning	Polygon data	Land use	Land-use planning	Low
Planning	Polygon data	Investment	Investment programme	Low

Note: PT = public transport; TMC = traffic management centre; STREAMS = intelligent transport system platform by Transmax

In addition to the data type and source, it is important to consider the availability of the data as well as the timeliness of the information provided. Raw and adjusted data is valuable for historical analysis, while timely data is required for real-time analysis, even if it requires on the fly filtering. The ideal data standard would be able to store data in different processing stages to allow fast access when needed, and correct and check values for historical, off-line analysis.

This leads to a requirement for strong data governance and associated responsibilities. As soon as data is stored in more than one location, the maintenance efforts to keep the information current increase exponentially. There is a trade-off to be made between full control of the data (ie creating a data silo) and accessing data that is out of one's responsibility. The full access has many short-term advantages that allow data optimisation, manipulation and mashing (ie overlaying and combining two or more data sets into a new one), but this comes at the cost of maintenance.

Finally, the lifecycle of data is an important consideration. Commonly, transport network data is stored in its current state, and operational data goes through a cycle of aggregation over time. While this was a necessity when data storage was costly, it prevents analysts from learning from the past. Trend analysis and forecasts based on such data are rendered useless if the data cannot be matched to an operational state of the transport network at a particular point in time. Congestion relief could be part of improved signal control monitoring, but could also be related to changes in public transport or city development. To answer these types of questions one requires a longitudinal dataset that provides context information for the operational measurements. The ideal data standard allows movement back and forth in time considering the changes to the transport network and services.

In summary, with regards to data management, the data standard for New Zealand's transport network should provide:

- richness of included data sources
- raw real-time data and corrected historical information
- a balanced trade-off between dynamic linked datasets and own data storage
- the ability to handle longitudinal datasets.

The following outlines the requirements for the standard to be able to serve more than a purpose build in the long term.

2.1.2 Interoperability

A data standard differs from an interface as it aims to provide broad usage beyond a single purpose implementation. Traffic operations data and the analysis of it is input to various applications such as:

- real-time decision support for operations

- (online) traffic simulation for operations
- real-time dashboard and other reporting
- traveller information
- transport planning
- asset management and maintenance
- data-driven decision support for infrastructure investment.

To open and provide information to all these applications, the datasets need to be linked together through a common reference. As transport networks are commonly captured in a GIS, mapping to a location seems straight forward but has the following issues:

- *granularity:* what is the right level of aggregation to enable detailed analysis without losing the ability to maintain currency of the model?
- *diversity:* it is common for transport agencies to deal with several GIS base models at a time depending on the area or mode of transport, so which one to choose?
- *timeliness:* due to the diversity of GIS base models, changes can take time and cripple or break real-time analysis processes
- *fit for purpose:* a one-fits-all base GIS limits the ability to optimise analysis processes.

With this in mind, the chosen data standard should be able to be referenced from a GIS to build visuals and maps, but not be dependent on a base model provided through the GIS teams. Any dependency would cause issues during updates of the GIS base model, as any data would need to be remapped to align again. In the interim, data consistency is not guaranteed and could cause unexpected results. The ideal candidate for a standard would provide a reference for the data that provides topology and connectivity based on vehicle/traffic movements. While focus is on road usage, it would be desirable to be able to include various modes of transport and to link between them. Further, the reference model (ie a model that is independent from the GIS base, but can refer to it geometrically) should come with a complete set of rules to ensure the reference network is built consistently and can be maintained without deep knowledge of the linked data structures. Possible rule sets could be based on:

- New Zealand road classification standards
- Austroads[2] road classification standards
- harmonic (dynamic) segmentation based on (capacity, geometry, route choice...)
- static segmentation, or others.

Standards are also evaluated by how they can be integrated into existing 3rd party applications, the availability of advanced programming interfaces (APIs) to build new interfaces, and out-of-the-box export features to common data formats such as comma separated values (CSV), Excel, structured query language (SQL) and so on. In summary, with regards to interoperability, the data standard for New Zealand's transport network should:

- provide a reference system to link various data sources in time and space
- provide interfaces and/or export functionality to common data format

[2] www.austroads.com.au

- be GIS compatible, but
- not create a competing GIS base model for the transportation network.

The next section considers performance requirements of the standard to ensure it is sustainable.

2.1.3 Performance

Considering the increase in transport demand and the need to operate our network to its potential, any new data standard for the transportation network of New Zealand requires high levels of flexibility, scalability and security. New technology will keep challenging standards to incorporate the information they are able to absorb, there will be more and more data flowing into transport operations centres, and the data will over time relate more and more to individual travellers.

Every standard will be judged on these upcoming demands in relation to the cost implications of achieving the level of service they provide. The more abstract a standard, the more flexible and sustainable it will be, but the costs of utilisation of such a standard come with high setup and maintenance costs as it requires deep knowledge of the standard, its options and limitations. Straightforward implemented standards will come at a cheaper start price, but bare the risk of becoming obsolete over time. All standards should allow secure storage of information, either by default or option, as data security for transport infrastructure and into operation will become more sensitive.

In summary, with regards to performance, the data standard for New Zealand's transport network should:

- be flexible in its design and not constrained by a fixed data scheme
- be based on a sustainable platform (ie adopt to new data sources and requirements)
- allow end-to-end encryption of data
- apply efficient data storage management
- maintain responsiveness with a forecasted data load for future years.

With the main requirements worked through, the following section looks more closely at the business requirements that inform this project.

2.2 Business requirements

As mentioned earlier the goal of the project was to establish a data standard for the New Zealand transport network primarily for road operations with the focus on real-time and historical analysis. Key requirements for its application, identified through stakeholder consultation, are:

- real-time interrogation of operational data from various sources
- off-line analysis of network operations
- user queries across multiple data sources
- configurable cost-based routing, selection and zoning
- ability to query on the fly during meetings (desired)
- provision of data and analysis results in common data exchange format.

These user base requirements are bound by limitations and restrictions based on business process within the NZ Transport Agency (the Transport Agency), such as:

- no interference with existing GIS models
- agnostic to proprietary software and software components.

While these are the short-term goals of the projects, any standard should have a long-term view. Therefore, the standard design aims to inform and hopefully guide other parts of the Transport Agency to use the standard as guidelines for business change over time to increase the benefit of intelligent and sustainable data management for transportation networks. The heritage of linear asset management and practice of patching new technology to fit the 'old' has led to an environment that does not support or enable change easily. The ideal standard would not be limited to operational data alone, but would be broader in scope. A modular standard without inter-dependencies that can advance with its users would be the ideal.

2.3 Evaluation framework

To ensure a developed or adopted standard meets the requirements established for this project, it needs to be evaluated objectively. Given a project timeframe of 12 months it is not feasible to consider every challenge in detail, but to take into account core requirements, such as:

- network model has granularity to represent the road network sufficiently to perform directional flow analysis
- network model is resilient to network changes and can be adapted at low cost
- network model uses linked data sources and business rules for data accessibility rather than additional data storage infrastructure
- network model is able to handle diverse data sets without constraining them
- network model has a relation to GIS but remains application agnostic to any specific GIS
- network model can support user queries of data sets, such as automatic cost-based routing
- network model supports time series data rather than only a single point in time data

The requirements above are considered the core requirements for any proposed way forward.

3 Solution design

This chapter focuses on the design of a solution that conforms with the requirements outlined in the previous chapter. It provides a review of existing solutions and approaches with their advantages and disadvantages, focusing on GIS and mathematical models, as these are the most suitable for building the foundation for a network analysis focused data standard. A network analysis platform is proposed that consist of two major elements: a network model that describes the infrastructure and a tool box that can host several analysis algorithms.

3.1 Review of existing solutions and approaches

Barcelo et al (2010) discuss the need for standardisation for traffic data and its collection, and Tamminga et al (2012) followed up on standardised exchange of data to enable frameworks that can adapt to change and are resilient to missing information. As standardisation, especially across departments, cities or even countries take extended periods of time and are often redundant by the time they are implemented, most efforts today fall back onto solutions from GIS (Abdul et al 2016), structured text formatted data such as GeoJSON (Butler et al 2016) and mark-up languages, such as CityGML (Gröger et al 2014) or OpenDRIVE (Richter 2015). Only a few projects use mathematical representations, as they usually do not link to the infrastructure geometry and make data visualisation and data exchange more challenging.

3.1.1 Geo-spatial information system data

A GIS is a computer system for capturing, storing, checking and displaying data related to positions on the earth's surface. A GIS can show many kinds of data on one map, which enables people to see, analyse, and understand patterns and relationships.

With GIS technology, people can compare the locations of different things to discover how they relate to each other. For example, using GIS, the same map could include road surface conditions, and crash sites. Such a map would help people determine black spots in the road network.

A GIS can use any information that includes location. The location can be expressed in many ways, such as latitude and longitude, address, or zip code. Many different types of information can be compared using a GIS. The system can include data about people, such as population, income, or education level. It can include information about land use, and it can include information about sites such as storm drains, roads and electric power lines.

3.1.1.1 Data and GIS

Data in many different forms can be entered into a GIS. Data that is already in map form can be included in a GIS. This includes information such as the location of rivers and roads, hills and valleys. Digital, or computerised, data can also be entered into a GIS. An example of this kind of information is data collected by satellites that show land use - the location of farms, towns, or forests. A GIS can also include data in table form, such as population information. GIS technology allows all these different types of information, no matter their source or original format, to be overlaid on top of one another on a single map.

Putting information into a GIS is called data capture. Data that is already in digital form, such as images taken by satellites and most tables, can simply be uploaded into a GIS. Maps must be scanned, or converted into digital information.

A GIS must make the information from all the various maps and sources align, so they fit together. One reason this is necessary is because maps have different scales. A scale is the relationship between the distance on a map and the actual distance on earth. A GIS combines the information from different sources in such a way that it all has the same scale.

Often, a GIS must manipulate the data because different maps have different projections. A projection is the method of transferring information from the earth's curved surface to a flat piece of paper or computer screen. No projection can copy the reality of the earth's curved surface perfectly. Different types of projections accomplish this task in different ways, but all result in some distortion. To transfer a curved, three-dimensional shape onto a flat surface inevitably requires stretching some parts and squeezing other parts. A world map can show either the correct sizes of countries or their correct shapes, but it cannot do both. A GIS takes data from maps that were made using different projections and combines them so all the information can be displayed using one common projection.

3.1.1.2 GIS maps

Once all the desired data has been entered into a GIS, it can be combined to produce a wide variety of maps, depending on which data layers are included. For instance, using GIS technology, many kinds of information can be shown about a single city. Maps can be produced that relate information such as average income, car ownership and travel patterns to a spatial representation. Any GIS data layer can be added to or subtracted from the same map.

GIS maps can be used to show information about number and density. For example, a GIS can be used to show how much congestion there is in different areas compared with the travel demand. They can also show what is near what, such as which homes and businesses are in areas prone to flooding.

With GIS technology, one can also look at change over time. Satellite data can be used to study topics such as how much of the polar regions is covered in ice. A police department can study changes in crime data to help determine where to assign officers.

A GIS often contains a large variety of data that does not appear in an onscreen or printed map. GIS technology sometimes allows users to access this information. A person can point to a spot on a computerised map to find other information stored in the GIS about that location. For example, a user might click on an intersection to find lane configuration, traffic controller in use, or phasing of the intersection.

GIS technology makes updating maps easy. Updated data can simply be added to the existing GIS program. A new map can then be printed or displayed on screen. This skips the traditional process of drawing a map, which can be time consuming and expensive.

People working in many different fields use GIS technology. GIS maps can show a region's future needs for public services like parking, roads and electricity. There is no limit to the kind of information that can be analysed using GIS technology, but one should note a GIS does not create the dataset that is being analysed.

3.1.2 Structured text formatted data

To avoid being locked in to software-specific data formats a growing number of data sets are stored in a structured text format. These formats can be parsed by various applications and are human readable, which increases significantly the ability to reuse the information over a long period of time. The following introduces the most common structured text formats used for data exchange.

3.1.2.1 Comma separated values

In computing, a CSV file stores tabular data (numbers and text) in plain text. Each line of the file is a data record. Each record consists of one or more fields, separated by commas. The use of the comma as a field separator is the source of the name for this file format.

The CSV file format is not standardised. The basic idea of separating fields with a comma is clear, but that idea gets complicated when the field data may also contain commas or even embedded line-breaks. CSV implementations may not handle such field data, or they may use quotation marks to surround the field. In addition, the term 'CSV' also denotes some closely related delimiter-separated formats that use different field delimiters. These include tab-separated and space-separated values. A delimiter that is not present in the field data (such as a tab) keeps the format parsing simple. These alternate delimiter-separated files are often even given a .csv extension, despite the use of a non-comma field separator. This loose terminology can cause problems in data exchange. Many applications that accept CSV files have options to select the delimiter character and the quotation character.

The common format and MIME type for CSV files (RFC 4180) proposes a specification for the CSV format, and this is the definition commonly used. However, in practice the term 'CSV' might refer to any file that:

- is plain text using a character set such as ASCII, various Unicode character sets (eg UTF-8), EBCDIC, or Shift JIS
- consists of records (typically one record per line), with the records divided into fields separated by delimiters (typically a single reserved character such as comma, semicolon or tab; sometimes the delimiter may include optional spaces)
- has the same sequence of fields in every record.

Within these general constraints, many variations are in use. Therefore, without additional information (such as whether RFC 4180 is honoured), a file claimed simply to be in CSV format is not fully specified. As a result, to utilise the data one should write a script or program to produce a conforming format based on the meta data of the CSV file.

3.1.2.2 (Geo) JavaScript object notation

JSON (JavaScript object notation) is a lightweight data-interchange format. Like XML it is easy for humans to read and write. It is easy for machines to parse and generate. JSON is a text format that is completely language independent but uses conventions that are familiar to programmers of the C-family of languages, including C, C++, C#, Java, JavaScript, Perl, Python and many others. These properties make JSON an ideal data-interchange language.

GeoJSON is an open standard format designed for representing simple geographical features, along with their non-spatial attributes, based on JavaScript object notation.

The features include points (therefore addresses and locations), line strings (therefore streets, highways and boundaries), polygons (countries, provinces, tracts of land), and multi-part collections of these types. GeoJSON features need not represent entities of the physical world only; mobile routing and navigation apps, for example, might describe their service coverage using GeoJSON. The GeoJSON format differs from other GIS standards in that it was written and is maintained not by a formal standards organisation, but by an internet working group of developers.

3.1.3 Mathematical representation

Another way to store transportation data is in the form of a mathematical model. Mathematical models are optimised to store data for calculations, such as simulation. When looking at large-scale transport models, the cell transmission model, proposed by Carlos Daganzo, is popular for analysing traffic phenomena. For more analytical tools the data is commonly stored in vectors, matrices, or data cubes for fast interrogation though mathematical processes. Mathematical models are generally utilised within applications where they are set up once and produce ongoing results such as in micro/meso modelling of traffic at operational level forecasting models. They are generally not made for data exchange or to accommodate data outside the immediate scope.

Generally, all geo-spatial enabled data formats can represent the road network accurately, with some including connectivity on link and/or lane level for routing. What most standards fail to capture is the change to the network overtime, which enables before and after performance studies at a more granular network level. This is because most standards are built with a specific application in mind, and are optimised to perform best for this application. While this is a reasonable strategy, the major drawback is loss of flexibility to adapt to changes in those applications over time.

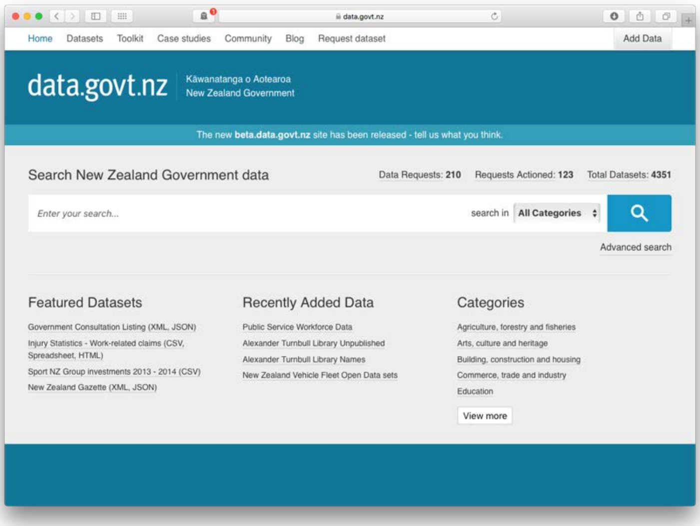
3.1.4 Sample implementations

This section provides a snapshot of publicly accessible transport data to provide more transparency into the decision-making process of transport agencies. The list is compiled to provide an overview of different directions the agencies have chosen around the world. When looking at examples of data standard implementation it becomes clear that the majority are data warehouse applications based on structured text or GIS information for common use or software package based data for specific purposes. No standard specific for transport or network data is reflected in the implementations.

3.1.4.1 Department of Internal Affairs, New Zealand

The New Zealand Government's open data website data.govt.nz is a collection of datasets from various areas, including transport. Datasets are provided mostly as CSV files through HTML or through a web based API.

Figure 3.1 Screenshot of the open data portal from the New Zealand Department of Internal Affairs



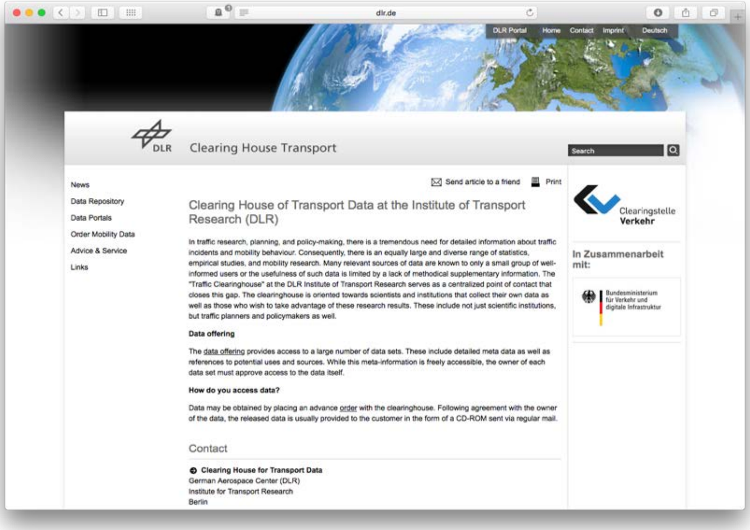
The data does not follow a consistent standard, and is provided in a tabular format that is then exported into CSV. While the data is easily human readable, accessing the data programmatically requires individual processing that connects the data sheets with meta data that makes is accessible for further processing.

3.1.4.2 German Aerospace Centre, Germany

The Clearing House of Transport Data at the Institute of Transport Research within the German Aerospace Centre is not an open data portal and focuses rather on data discovery. The idea is to make collected data available for research and practice and to reduce the cost for collecting redundant data.

Metadata can be browsed freely to allow users to become familiar with the dataset. If the dataset suits the user's needs, they can place an order at the site that will be re-directed to the owner of the dataset.

Figure 3.2 Screenshot of the transport data clearing house from the German Aerospace Centre

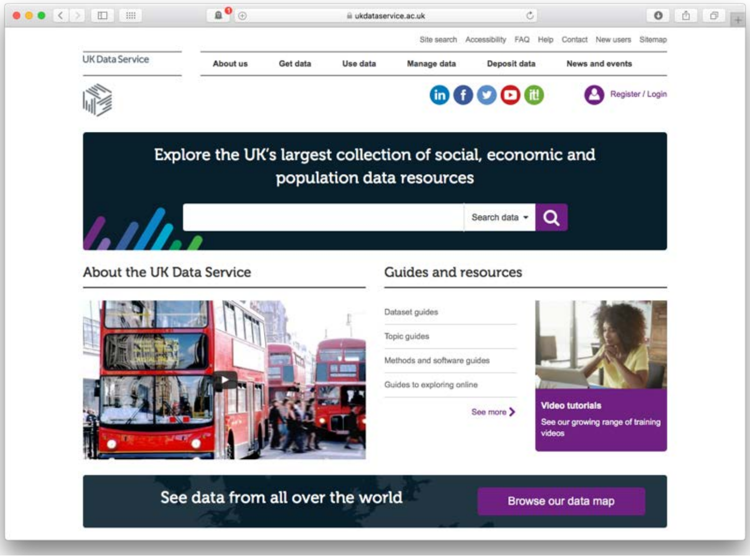


To provide a standardised set of metadata, the website adopts the Dublin Core Schema. The Dublin Core Schema is a small set of vocabulary terms that can be used to describe web resources (video, images, web pages etc), as well as physical resources such as books or CDs, and objects like artworks.

3.1.4.3 Economic and Social Research Council, United Kingdom

The UK Data Service website, like other government websites aims to provide as much data as possible to research and the public to benefit from external analysis. The UK has been a pioneer in aggregating various data portals to standardise access and governance of the data sources. This has led to an increase in data usage and registration of new data sets in a short period of time. The data provided is commonly structured formatted text that requires parsing for usage in analysis. While text files remain readable, the lack of a strict format and meta information provides a risk for consistency.

Figure 3.3 Screenshot of the open data portal from the Economic and Social Research Council



3.1.4.4 Delft University of Technology, The Netherlands

A different approach was taken by the team behind the Regiolab Delft website, which provides transport data to researchers and government for analysis. In addition to the datasets, which mostly consist of minute by minute traffic flow data from motorways in the region formatted as CSV files, the website also offers analysis tools and a library of scripts that can be downloaded for off-line analysis.

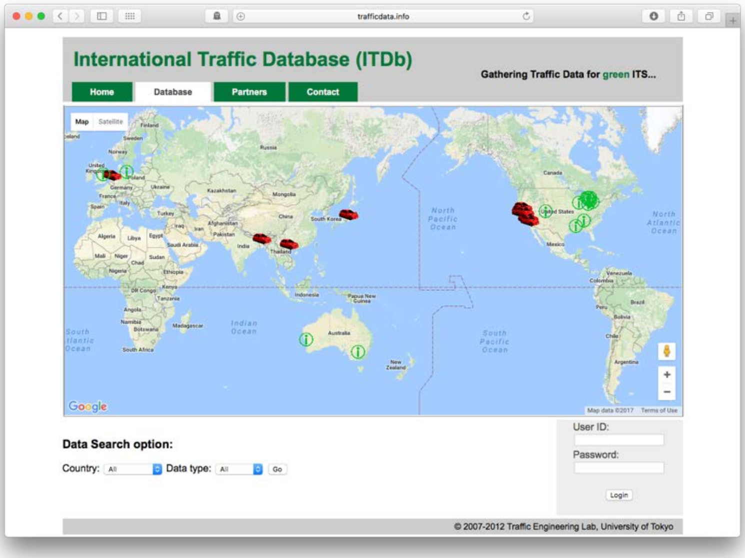
Figure 3.4 Screenshot of the transport data portal of Regiolab Delft

Regiolab Delft has its roots in the need for preserving collected data for research and trend analysis and is funded by local and federal government since 1999.

3.1.4.5 University of Tokyo, Japan

The International Traffic Database, hosted by the University of Tokyo in Japan, is the result of a collaboration between Japan, Europe and the United States of America. The goal of the project was to provide calibration datasets for emission models in the three regions to provide trust among researchers and governments that the local emission models are accurate.

Figure 3.5 Screenshot of the International Traffic Database

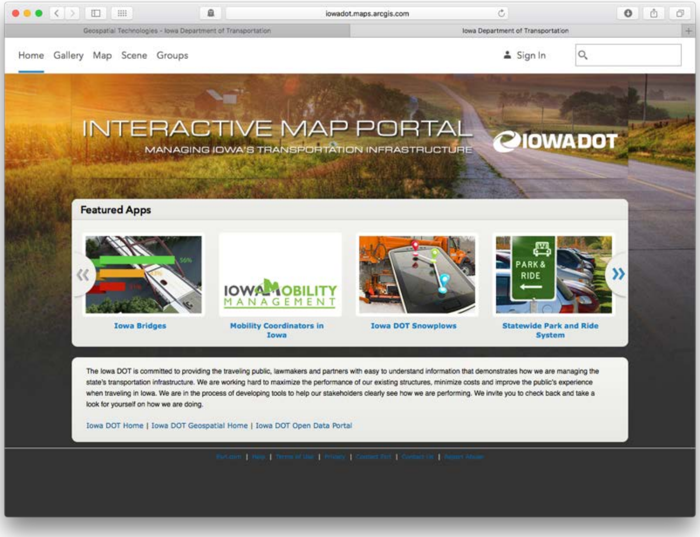


The project highlighted that traffic data (eg detector and vehicle type data) is not sufficient to understand vehicular flows. Network data is required to be able to replicate findings and results. As the project team did not standardise the network data, the utilisation of the information provided was limited. Efforts to extend the dataset with network information are ongoing. The data available is in CSV format and is standardised across the three regions. The standard is not compatible with any other standard and hence only valid for sources from this site.

3.1.4.6 Interactive Map Portal Iowa, USA

The Interactive Map Portal from the Iowa Department of Transportation is a GIS-based web portal, built using the ARC GIS platform from Esri[3]. Various traffic measurement data and vehicle information is presented in a geo-spatial web interface that users can browse through.

Figure 3.6 Screenshot of the Interactive Map Portal from Iowa



Not all data is available for download, but open data can be exported into shape files and other structured plain text formats such as CSV and GeoJSON.

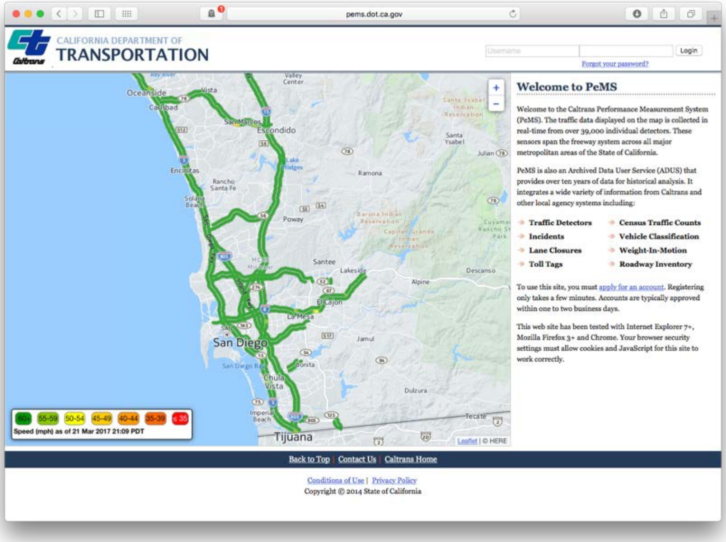
3.1.4.7 Caltrans Performance Measurement System, USA

The Caltrans Performance Measurement System (PEMS) was one of the first web portals to provide real-time information of the traffic network to the public and allow data download for off-line analysis. The project was driven by the notion that data driven decision support is only possible if you collect data and allow data to be accessible and to be transparent in policies.

Despite push-back from many transport authorities, the PEMS site has shown that opening data to show the status quo of the transportation network can lead to better investment in data collection infrastructure with support from the public and agencies.

[3] Esri is an international supplier of GIS software, web GIS and geodatabase management applications, based in Redlands, California.

Figure 3.7 Screenshot of the PEMS website in California



3.1.5 Common analysis tools

While data warehouses provide rich data sets in various sensor or application-oriented formats, they are generally not designed to enhance analytical capabilities, but for data exchange only. To provide analytics from raw data one requires additional analytics software. Generally, there are two types of analytics software package:

- specialised software packages (ie visualisation, modelling)
- mathematical toolbox based systems.

While specialised software packages typically provide an easy to use interface, they have limitations regarding data manipulation. When utilising these software packages one must understand they are not visualisation tools, such as GIS, but mathematical software packages that allow processing of data in various ways. These systems require advanced knowledge in mathematical modelling and techniques despite their user friendliness to provide reliable and useful outcomes.

The following provides a brief overview of two systems commonly used in research and practice. The specialised software package Tableau, and the general mathematical toolbox MATLAB.

3.1.5.1 Tableau

Tableau is a data analytics software that allows visual data interrogation through drag and drop. It is built for speed and ease of use to spot visual patterns in data. Data can be sourced from spreadsheets, SQL databases, Hadoop or cloud services and combined without writing a single line of code. The software is used by various transport agencies (eg Singapore Land Transport Authority, TransUrban) to investigate traffic patterns in the network. A key benefit of Tableau is its ability to blend disparate data sources with a couple of clicks, which makes it easy to provide visual representations in graphs, maps and charts.

3.1.5.2 MATLAB

The MATLAB platform is designed to solve engineering and scientific problems with a matrix-based language and a series of mathematical tool boxes that allow for data manipulation and analysis. It provides a powerful tool for computation and visualisation, but in comparison to Tableau leans more towards a programming or development environment rather than a data interrogation platform. MATLAB scripts can be exported as Java code for integration in software applications and hence provide an easy entry to deploying software based on prototype implementations of new algorithms. While analytical models such as MATLAB allow for simulation, specialised packages for traffic simulation such as Aimsun provide significant advantages in respect to ease of preparing the input data and collection of results.

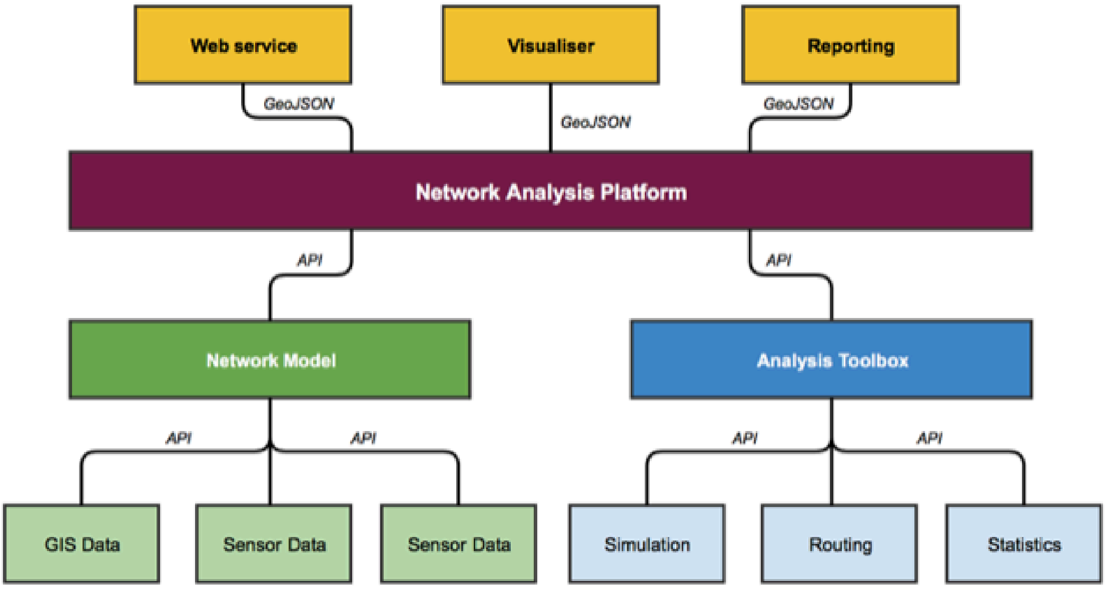
In general, tools for analysing data in transportation acknowledge the fact that they come from various sources and allow data connections to databases, local files and cloud storage. Recent developments, such as IBM Watson (High 2012), Tableau (Murray 2013), Power BI (Hart 2017) and Spotfire (Kaushal et al 2004) provide interfaces to multiple sources via simple drag and drop actions. The risk of these systems is the lack of control over data quality. Like traffic simulation models that, without careful calibration and validation, provide solutions that are not useful, these big analytic platforms can provide misleading results if not used with care.

3.2 Network analysis platform

Based on the reviewed material it became clear that the goals of this project did not align with an existing standard. The application driven standardisation efforts could only provide partially what was required and it was necessary to explore new ways forward. The need to represent data visually to provide easy to understand information of complex relationships is the strength of GIS, but these systems are not designed to create new datasets based on traffic flow theory, and are generally limited to filtering and geo-spatial manipulation. Data exchange formats allow for an easy exchange between analysis tools, but system-wide standardisation is required to benefit from it. Mathematical models are key to cleanse raw data, to filter out information from raw data and to provide algorithms that turn this information into knowledge. This feature is not part of any exchange format or GIS, but would benefit from the capabilities of both.

A sustainable standard for the New Zealand transport network should not be limited to an application but rather describe a transport network analysis platform (see figure below) that can exchange information with existing data warehouse solutions through common data exchange format, manipulate and analyse those inputs, and finally, be able to export results into format that can be visualised in a GIS environment. Expanding the data standard for the New Zealand transport network to a network analysis platform provides a sustainable pathway into the future. Introducing a middleware layer that enables standardisation on a higher, more abstract level allows for the integration of new data sources over time, as well as integrating new algorithm and interrogation methods as research progresses. Data and its sources have been and will keep changing rapidly, with more and more information becoming available through new sensor technology and reduction in production costs of the devices. This will enable new ways to analyse the transport network over the years to come, and any standard to be introduced nationally should be able to adapt accordingly.

Figure 3.8 Proposed network analysis platform



It is recommended to build standardised interfaces based on physical and economical processes as well as on aspects of asset management. These interfaces should be used to feed data into the platform, to manipulate the data, and to create results based on user defined queries. As they do not rely on specifics of data or algorithms, these interfaces can be maintained over time by adding bridges for new data sources as they become available.

All computational results as well as data within the network analysis platform should be accessible through GeoJSON exports as it is a machine and human readable standard that is designed to preserve geo-spatial context of data. The two key elements of the framework, the network model and analysis toolbox are defined in more detail below.

3.3 Network model

The approach for the network model is to take a traffic flow-oriented spatial representation, rather than the physical road network. The model describes flows between terminals (origins/destinations) using routes. Routes consist of one or more paths that are linked to an underlying representation of the physical road network including topology, connectivity and direction. Paths are connected through switches or hubs, depending on the knowledge about split fractions, to represent merging and diverging flows. Therefore, the network reference model is geo-spatially based on the physical road network, but independent from it to make it resilient to change and/or errors. All network elements have a unique identifier, and have a timestamp for their activation and retirement to capture longitudinal changes in the network. The following sections describe the elements within the network reference model.

3.3.1 Hubs

Hubs connect flow paths where traffic is merging or diverging. Hubs, in contrast to switches, have no knowledge about the distribution of flows between the connected paths.

Figure 3.9 Illustration of a hub connecting flow paths

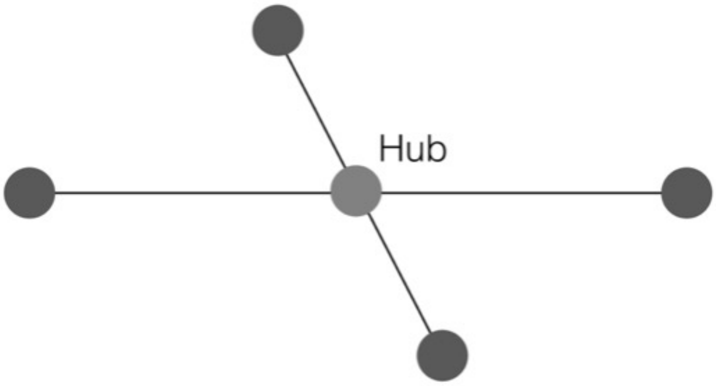


Table 3.1 Hub definitions

Attribute	Description
hub_id	UUID that uniquely identifies the area Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	Timestamp in GMT that indicates the start of utilisation/or initialisation Format: YYYY-MM-DD hh:mm:ss
active_until	Timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
name	Free text up to 255 characters (ie intersection identifier)
comment	Free text up to 255 characters for comments
geometry	Point of hub as latitude and longitude

3.3.2 Path

A path is a connection between hubs, hubs and terminal, terminals and switches, and switches and hubs.

Figure 3.10 Illustration of a flow path

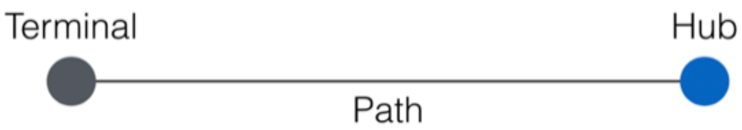


Table 3.2 Path definitions

Attribute	Description
path_id	UUID that uniquely identifies the connector Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	Timestamp in GMT that indicates the start of utilisation/or initialisation Format: YYYY-MM-DD hh:mm:ss
active_until	Timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
from_id	UUID of the terminal, hub, or switch where the path originates
to_id	UUID of the terminal, hub, or switch where the path ends
name	Free text up to 255 characters (ie street or area identifier)
comment	Free text up to 255 characters for comments
geometry	Linestring that represents the path

3.3.3 Route

Routes are a collection of one or more paths between terminals.

Figure 3.11 Illustration of a route between terminals

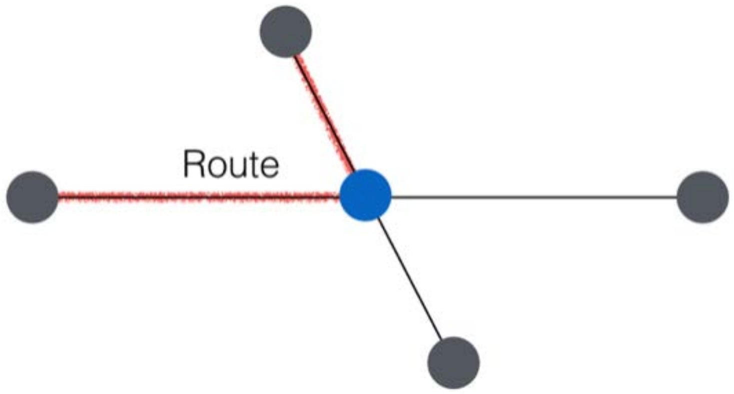


Table 3.3 Route definitions

Attribute	Description
route_id	UUID that uniquely identifies the connector Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	Timestamp in GMT that indicates the start of utilisation/or initialisation Format: YYYY-MM-DD hh:mm:ss

Attribute	Description
active_until	Timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
path[]	Array of path UUIDs that uniquely identify the route.
name	Free text up to 255 characters
comment	Free text up to 255 characters for comments

3.3.4 Switch

Switches, like hubs, connect paths and contain information about the distribution of flows between the connected paths.

Figure 3.12 Illustration of a switch connecting flow paths

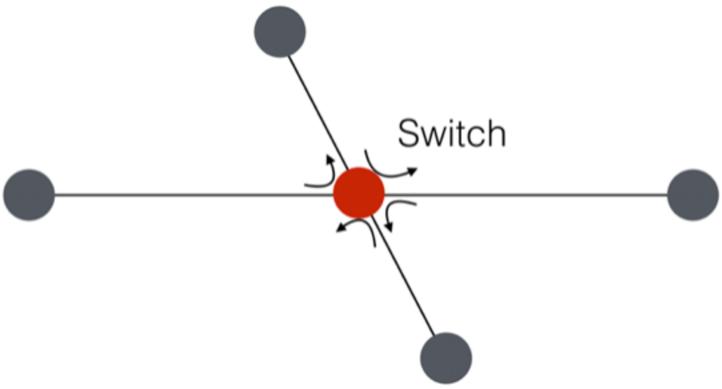


Table 3.4 Switch definitions

Attribute	Description
switch_id	UUID that uniquely identifies the link Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	Timestamp in GMT that indicates the start of utilisation / or initialisation Format: YYYY-MM-DD hh:mm:ss
active_until	Timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
paths[]	Array of path UUIDs that connect to the switch
fractions[]	Split fractions, corresponding to path array
name	Free text up to 128 characters for comments
comment	Free text up to 255 characters for comments
geometry	Location of switch

3.3.5 Terminal

Terminals are origins and destinations within the real physical network that are of interest for analysis. Terminals do not represent just a single point, but the general area with a given radius (ie CBD).

Figure 3.13 Illustration of terminals on a map

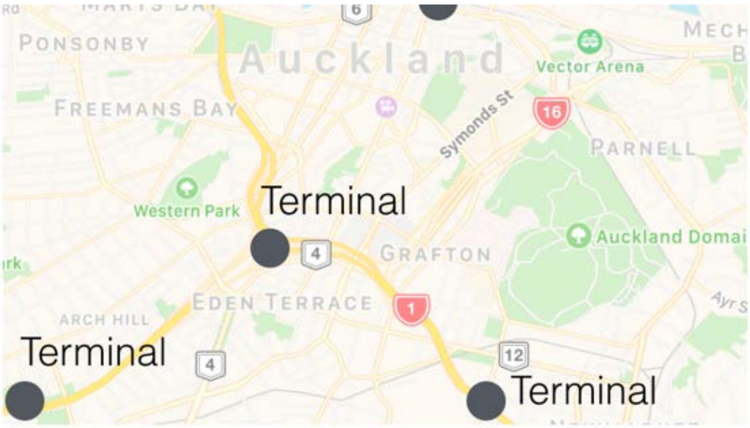


Table 3.5 Terminal definitions

Attribute	Description
terminal_id	UUID that uniquely identifies the link Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	timestamp in GMT that indicates the start of utilisation/or initialisation Format: YYYY-MM-DD hh:mm:ss
active_until	timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
radius	maximum radius for node aggregation
name	free text up to 128 characters for comments
comment	free text up to 255 characters for comments
geometry	location of terminal

3.3.6 Sensor

Sensors are data availability markers that represent access to a data source about the path itself (ie detectors, BT scanners, counts, costs). A path can have one or more sensors with any combination of underlying data. A radius attribute for each sensor is added to allow for aggregation of data from various sources or detection systems (ie multiple loops, cameras). Note that sensors also include settings of control elements, such as traffic lights, variable message signs (VMS) and others as they act as an input to the system.

Figure 3.14 Illustration of sensors collecting data on a flow paths

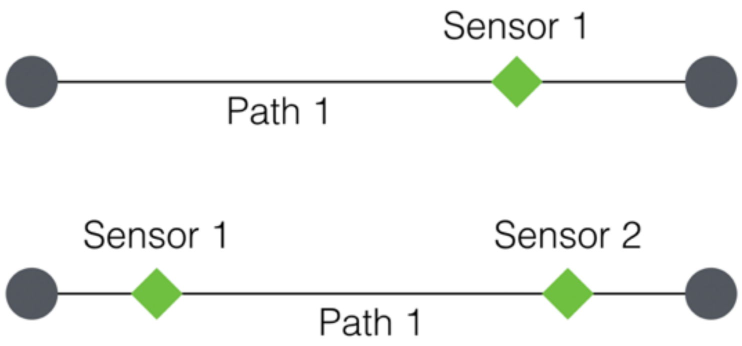


Table 3.6 Sensor definition

Attribute	Description
sensor_id	UUID that uniquely identifies the marker Format: 36 characters (32 alphanumeric characters and four hyphens) as per Open Software Foundation standard
active_from	Timestamp in GMT that indicates the start of utilisation/or initialisation Format: YYYY-MM-DD hh:mm:ss
active_until	Timestamp in GMT that indicates the end of utilisation Format: YYYY-MM-DD hh:mm:ss
sensor_type	Data type
data_format	Format of data
dynamic_link	End-point of data access
geometry	Location of sensor

Sensors also include the settings of control elements, such as traffic lights, VMS and others as they act as an input to the system.

3.4 Analysis toolbox

The initial minimum analysis toolbox allows the user to create the network model, and a cost-based routing algorithm that enables users to route the network based on impedance provided by any attribute available. There are two ways to create a network reference model: creation from terminals, or creation from sensor data. While the first builds mainly on GIS information, the second considers the overall knowledge of the flows in the network, which can potentially lead to a set of unconnected network reference models.

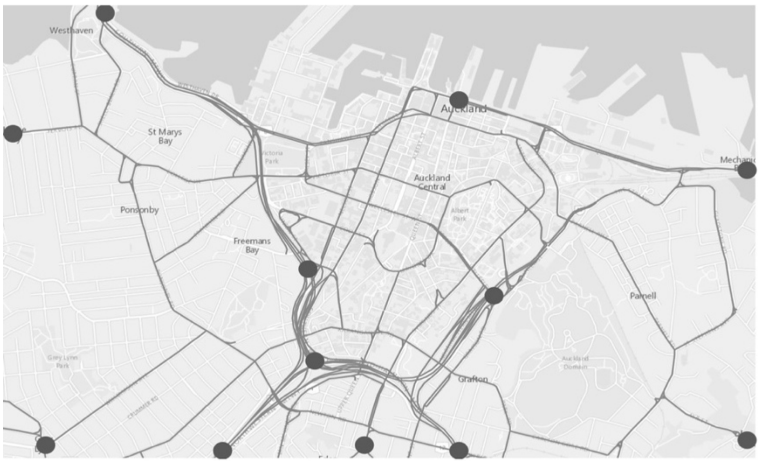
3.4.1 Network creation from terminals

Creating the network reference model with terminals is the standard approach, as it follows a more traditional GIS-based creation of a network model. The starting point is a map of the area in any desired resolution.

Figure 3.15 Illustration of terminals placed on a map interface

On this map, the user is adding terminals, which are points of interest that are used as origins and destinations of flow to be analysed. Terminals are not fixed locations in a GIS sense, but merely indicators of an area of interest.

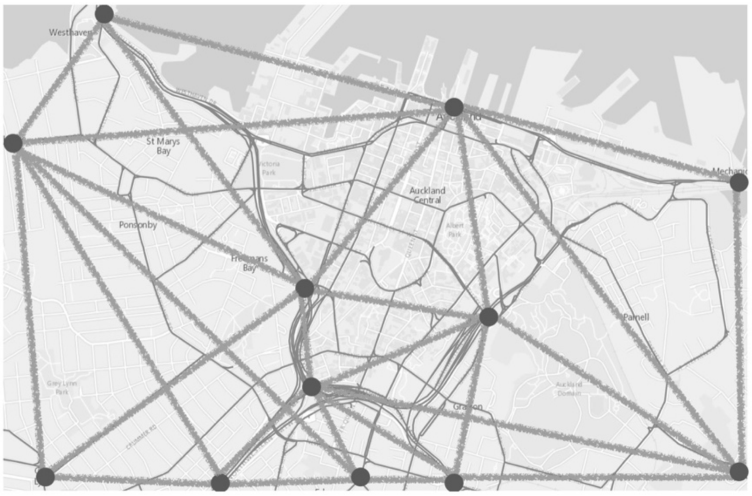
Figure 3.16 Illustration of connecting path between terminals



However, the density of terminals in any given area will affect the resolution of the reference model: For a transport planning application, the CBD of Auckland could be one terminal and the analysis would show aggregated measures for travel from and to the CBD. This would probably not be enough detail for traffic operations, which is more concerned with the utilisation of the network in the CBD and surroundings. Such analysis would require a higher density of terminals for the CBD area.

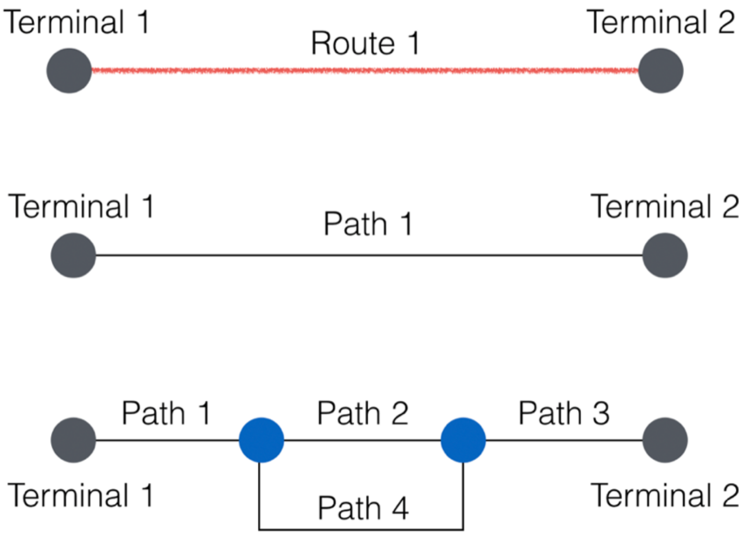
Once the terminals are placed, the latitude and longitude of their position are extracted and passed to a mapping tool that overlays the terminals onto the latest GIS network model available. Ideally, the GIS network model has sufficient information to extract road configuration and lane by lane connectivity throughout the network, but as a minimum requirement it should be routable. Closest neighbour search within the GIS model provides an array of GIS nodes that are included in the terminal area of interest. Those nodes are then aggregated into a GIS area that is used for building routes.

Figure 3.17 Illustration of a simplified flow path



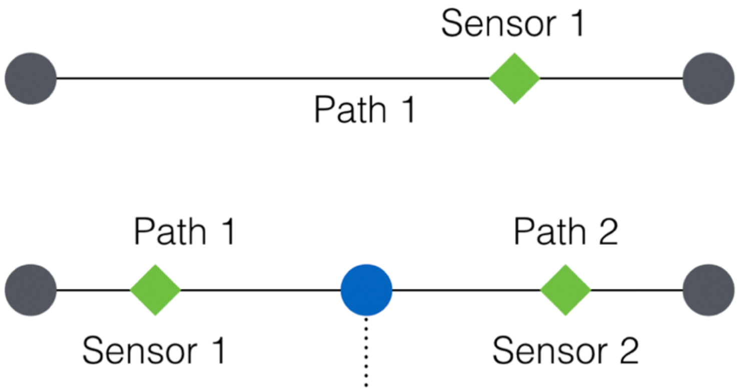
Routes are determined by running an algorithm for finding shortest paths in a weighted graph. Depending on the level of detail provided by the GIS network, the weights are based on free-flow travel time, road hierarchy and distance. The abstract construct of a route stores the path based on the GIS base model geometry, but remains independent from it as it does not take over any identifiers, hence is resilient to any changes being made within the GIS model going forward. For example, if route 1 between terminal 1 and terminal 2 is represented today by one single path (path 1), but in future, due to the construction of an alternative route, through three paths (path 1, path 2 and path 3), the created network model would not be affected, as geometry has not been altered. Results and map-based visualisations remain valid.

Figure 3.18 Illustration of routes containing several flow paths



However, the flows of route 1 might significantly change and cause issues with the analysis as the system is unaware of the change in the physical road network. To prevent this, a regular update of routes is recommended. Hence, a route between two terminals is based on paths that share the geometry of the latest GIS reference.

Figure 3.19 Illustration of flow paths splitting

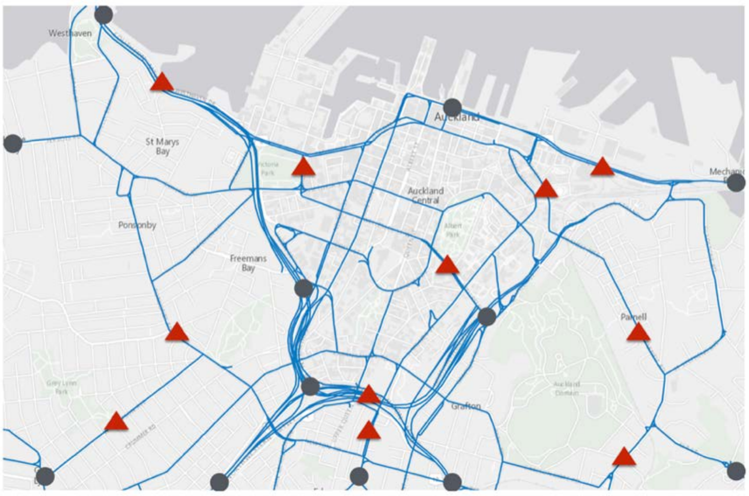


If sensor data is available for routes, and is updated independently from the GIS base layer, it is possible to detect network changes due to anomalies in the measurements. Considering a path between two terminals with a single sensor on it: If a new road/intersection is added to this path, and is relevant to the overall network performance, the sensor on path 1 would report abnormal numbers that would trigger an alert. If a second sensor was added to the new road to measure influx from the added road, the system would detect a violation of flow conversation on the link, and report the issue with an approximate location of the abnormality (ie between the two sensors).

3.4.2 Network creation from sensor data

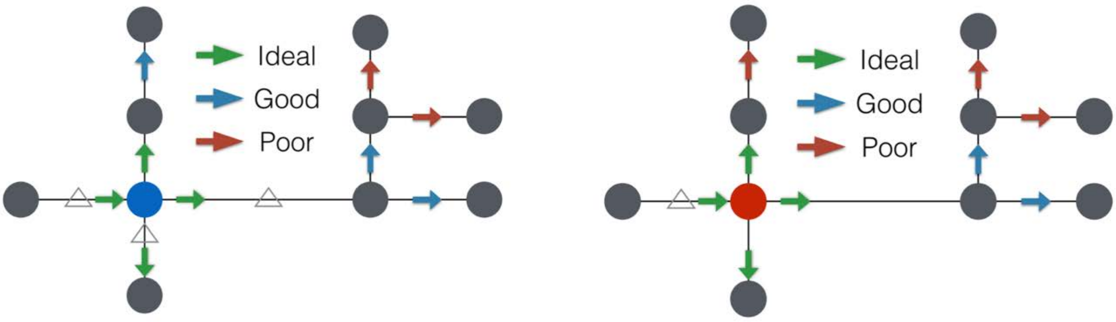
An experimental, but more holistic approach is to build the network model from sensor data. This enables users to get a better understanding of the information available throughout the transportation network. The starting point is again a map of the area in any desired resolution. The next step, however, matches sensors with the latest GIS base model to generate paths with a given level of information based on the presence of a sensor.

Figure 3.20 Illustration of hubs and switches across the network



By looping through all links with high information, and running a routing engine to each terminal, every path is assigned an information value. Along the path, information can increase or decrease based on the network element that is passed along the way.

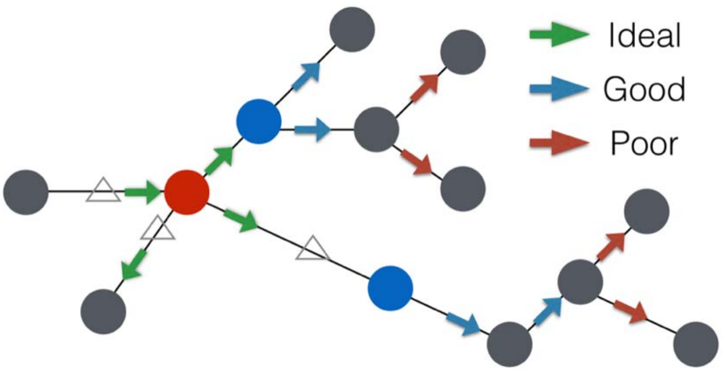
Figure 3.21 Illustration of information flow through hubs and switches



For example, if traffic flow is measured across an arterial road that leads into a roundabout with three exits that are not monitored (ie hub), the assumed knowledge of traffic flow after the roundabout decreased significantly. If the exits are equipped with sensors (ie switch), the level of confidence in the traffic flow information remains high.

The result will be one or more connected networks, representing the level of knowledge one has. This step can be done for various attributes and inform decisions on investment in detection required for a better understanding of the transport network. While this approach has many advantages over the standard GIS approach, it requires a high level of detail on detection equipment when it comes to placement and coverage. Unfortunately, this level of detail is not standard nor best practice in agencies around the world and would require a significant investment in time and effort.

Figure 3.22 Illustration of confidence in data through network



3.4.3 Cost-based routing

User queries that use cost-based routing require the selection of at least one data set that contains the attribute(s) to be used as link impedance. A modified Floyd Warshall algorithm (Hougardy 2010) calculates the shortest path between all network nodes and the result can be interrogated visually by the user.

3.5 Alignment with core requirements

Reflecting on the core requirements set out by the project steering committee, this section shows if and how the requirements are met.

3.5.1 Accurate representation of the road network

The proposed network model carries the geometry attributes from the GIS base, and maintains this attribute for its lifetime. Hence, when any results of analytics on the network model are exported to a mapping system or GIS, the alignment with the road network is maintained and only limited by network changes due to construction work or other infrastructure changes.

3.5.2 Network model represents traffic flow

During the process of generating the network model from GIS base data, a directed graph that aligns with possible vehicle movement through the network is built. So, independent from GIS construction rules (ie one link per direction, bidirectional links), the network model can represent traffic movements and is able to export these to a map environment.

3.5.3 Network model allows user defined routing

As routing analysis with user-defined cost functions was a priority, the network model graph allows for user-defined values as an impedance for edges. Values are required to be positive to ensure that standard algorithms like Dijkstra can be used without checks for negative loops.

3.5.4 Resilience to network changes

By carrying the geometry of a path, rather than a reference to a GIS link, the proposed network model is resilient to not only minor, but any change to the network and representing GIS. To align the network model with a new or changed GIS model, the two are joined by location. For major changes, further manual editing might be required, but minor changes, such as alignment or link splitting, are handled without user interaction.

3.5.5 Ability to deal with a variety of data sources

The proposed network model includes a sensor element that acts as an abstract container for transport data in JSON format. Adding a new data source to the platform requires a one-time generation of a descriptor for the source that can be used from that point onward. Sensors are not limited to measurement data, but can also include information actively provided to road users.

3.5.6 Agnostic to (proprietary) software packages

The solution is based on common and open standards that do not require a licence or proprietary software component to work. Export to proprietary formats, such as ESRI shapefiles, is possible if desired, but it is recommended to leave the core independent and text based for sustainability. Most third party software that deals with visualisation and/or analytics of data has import filters for GeoJSON and would allow direct integration out of the box.

3.5.7 Avoiding data replication

Sensor elements of the network model can retrieve data from existing databases, web services and manual upload. Hence, it is possible to harvest existing data stores without duplication. Dependencies to data sources out of immediate control will require arrangements that outline the availability of the data over time to ensure access.

3.5.8 Provision of longitudinal data-sets

With geometry being the underlying reference between the network model and the real world, a frequent re-occurring generation is recommended to ensure minimal adjustments over time. Large construction projects and network changes can lead to loss in GIS data that has been preserved in the network models to allow before and after analysis.

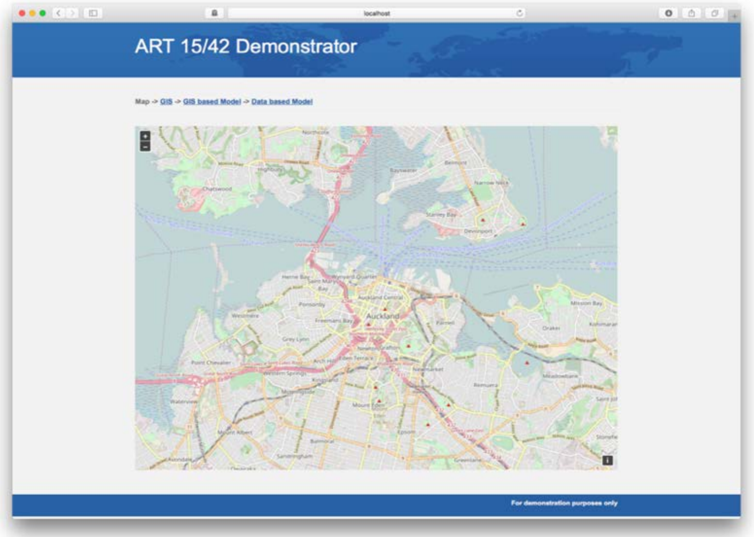
4 Platform prototype

A platform as proposed in chapter 3 requires a minimum development effort of six month to provide the basic infrastructure and functionality. To accelerate this process, a platform prototype that replaces the interface with more limited connectivity was developed as a web application consistent of a visualiser and backend to showcase the network model and its performance.

4.1 Visualiser

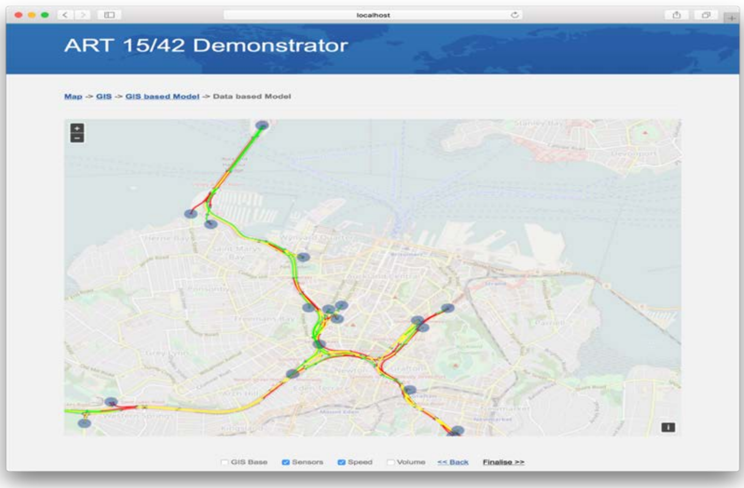
The visualiser is built as a web application and uses the OpenLayers framework (Perez 2012) for visualisation, as it allows easy access to basic maps as well as GeoJSON based layers. While the network model itself is independent from the map layer, it is a helpful tool to provide information to users.

Figure 4.1 Screenshot of the map- based visualisation



Users are able to add data as new layers, create new data-sets through cost-based routing, and to edit and export the results in a GeoJSON format.

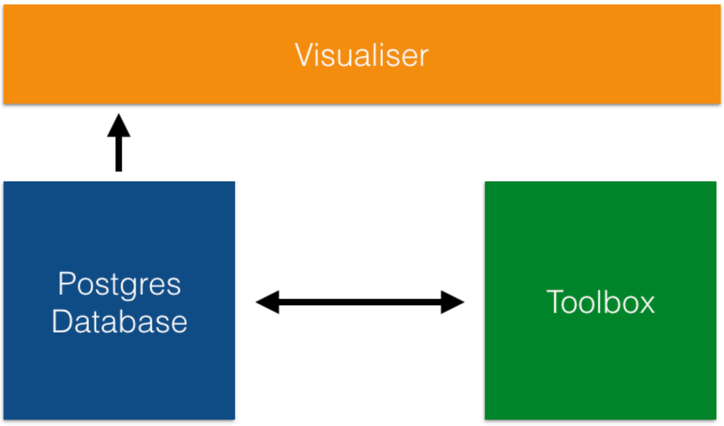
Figure 4.2 Screenshot of data visualisation in the demonstrator



4.2 Backend

The backend of the prototype consists of a Postgres database with an interface that communicates through the visualiser with the user.

Figure 4.3 Illustration of the backend components



While this design is sufficient for a prototype, a full scaled system would require a more scalable and robust setup.

5 Conclusions, recommendations and future research

5.1 Conclusions

The purpose of the project was to develop a New Zealand data standard that provides a cost-effective, robust and sustainable solution for transport network analysis. The core requirements were that the standard:

- uses existing data as a foundation without the need for redundancy, to provide cost-effective data management
- supports existing business processes, to avoid changes to day-to-day business that could result in resistance to adoption
- provides flexible data schemas to adapt to change, to build a sustainable solution that is robust to changes in data formats
- is agnostic to proprietary software and software components, to avoid lock-in into software environments.

After reviewing existing standards and analysing standard approaches around the world, the project concluded that data standardisation alone would not provide the required benefits and a more elaborate platform was needed to enable network analysts to provide up-to-date decision support in a cost-efficient and sustainable way.

The researcher accordingly sought to find a method that would support normalising diverse network representations of selected data sources, and support presentation of the information within the context of a standardised reference transport network model. Such a method comes with various challenges.

When considering the factors in maintaining the currency of the model (eg what is an appropriate granularity for defining the model), it is most important not to break the physical connection between the physical transport network and the data that is collected in various formats from multiple sources. The proposed solution leverages the power of GIS to work with geo-spatial relationships without depending on any specific GIS in particular. Instead of using unique identifiers for network elements and data points, the proposed solution stores information against a geometry. This powerful connection allows varying levels of granularity among users without breaking the fundamental relationship between the network and its data.

Small network changes (ie link splitting, realignment of geometry) will have no effect on operational analysis, and adding sensor or other data sources does not depend any longer on links, path, or section sizes. Therefore, the reference model itself can adapt over time to increase overall data quality without having an impact on operations.

Similar benefits are shown when considering the ability to add diverse datasets without constraining them. When linking a data set to the reference model, the user defines a geometry (ie point, path, area) and direction to enable mapping of the information on a geo-spatial level. If implemented across the department, data collections can be mashed on a geo-spatial level without the loss of governance of the data itself. However, it should be noted that business processes are required to ensure continuity of data access.

There is a key difference between the proposed reference model and a GIS. While a GIS is similar to any other database driven by object identifiers (IDs) that can be mapped to other data, the proposed model links data purely by location and direction, using the physical transport network as a common base. Data is mapped to a location and needs to be mapped to an ever-changing GIS model repeatedly over time to maintain an ID level mapping. However, this is only necessary to maintain a reference model for historic comparison, as it provides snapshots of the infrastructure over time.

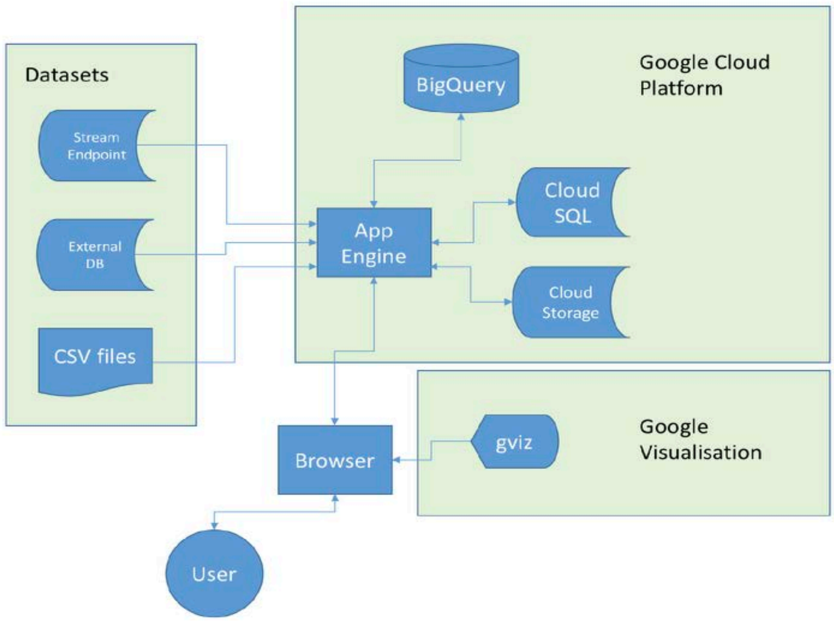
Further, the proposed reference network supports user queries of data sets like cost-based routing of the network model. With geometries being the glue between data and the physical road network, a myriad of routing possibilities opens up to the users (eg select all VMS for traffic approaching incident A).

Sections 5.2 and 5.3 outline recommendations and future research steps to enhance this work.

5.2 Recommendations

Initial results show it is possible to build a network model based on existing GIS and sensor data that is resilient to minor changes, cost efficient and software agnostic. As a next step, it is recommended to implement a full version of the prototype, using one of the big cloud platforms (eg Google, Amazon, MS Azure) for fast deployment. Platform deployment should be to enterprise standards to ensure outcomes are ready to use in a real-world setting.

Figure 5.1 Design of full- scale implementation of analysis platform



A full-scale implementation as depicted above, would run on a fully managed cloud platform that does not require dedicated hardware and can be scale dependent on the demand. Users connect through a browser to the platform. The browser uses a visualisation engine to provide feedback to the user based on results and functionality provided by the application engine. The application engine connects various data sources, ranging from databases to on-the-fly data uploads, with big data query engines that are readily available in the cloud platform, such as BigQuery, CloudSQL and CloudStorage for non-SQL data transactions. The main application can be updated over time to increase capabilities and analysis performance.

Further, to promote the work and to receive input from other agencies, it is recommended to open the platform to synchronise efforts, resources and ideas. Joined development would provide a more cost-effective and timely establishment of best practice and standardisation.

The open platform should further provide a sandboxed access to trial new developments and explore future directions in a safe environment that does not affect the production version.

5.3 Future research

If the platform is established, future research should focus on enhancements in analysis tools and automated reporting. By providing real-world data access to research groups, and encouraging them to use the provided standardised interface the advances in network analytics tools could be boosted significantly. New solutions could be used and tested before moving them into the production version. Enabling this future research would require:

- establishment of the platform
- developing public facing APIs
- developing guidelines for tool developments
- developing rules of engagement to avoid intellectual property issues.

With these in place, the platform could quickly become a thriving hub for network analysis.

6 References

Abadi, A, T Rajabioun and PA Ioannou (2015) Traffic flow prediction for road transportation networks with limited traffic data. *IEEE Transactions on Intelligent Transportation Systems 16*, no.2: 653–662.

Abdul, J, M Alkathiri and MB Potdar (2016) Geospatial Hadoop (GS-Hadoop) an efficient mapreduce based engine for distributed processing of shapefiles. *International Conference on Advances in Computing, Communication, & Automation (ICACCA) (Fall),* IEEE, 2016.

Barceló, J, M Kuwahara and M Miska (2010) Traffic data collection and its standardization. Pp1–10 in *Traffic data collection and its standardization*. J Barceló and M Kuwahara (Eds). New York: Springer.

Butler, H, M Daly, A Doyle, S Gillies, S Hagen and T Schaub (2016) *The GeoJSON format*. No. RFC 7946.

Casas, J, A Torday, J Perarnau, M Breen and A Ruiz de Villa (2013) Present and future methodology for the implementation of decision support systems for traffic management. *36th Australasian Transport Research Forum (ATRF),* Brisbane, Queensland, Australia.

Casas, J, A Torday, J Perarnau, M Breen and A Ruiz de Villa (2014) Decision support systems (DSS) for traffic management assessment: notes on current methodology and future requirements for the implementation of a DSS. *Transport Research Arena 5th Conference: Transport Solutions from Research to Deployment.*, Paris, France, 14-17 April 2014.

Gröger, G, TH Kolbe, C Nagel and KH Hafele (2014) OpenGIS city geography markup language (CityGML) encoding standard (OGC 12-019). Version 2.0.0. OGC 12-019. *Open Geospatial Consortium 2014*.

Hart, M (2017) *Get started with Power BI service*. Accessed 27 July 2017. https://powerbi.microsoft.com/en-us/documentation/powerbi-service-get-started/

High, R (2012) *The era of cognitive systems: an inside look at IBM Watson and how it works*. IBM Corporation, Redbooks.

Hougardy, S (2010) The Floyd Warshall algorithm on graphs with negative cycles. *Information Processing Letters 110*: 279-281.

Jin, H, M Miska, E Chung, M Li and Y Feng (2012) Road feature extraction from high resolution aerial images upon rural regions based on multi-resolution image analysis and gabor filters. *Remote Sensing-Advanced Techniques and Platforms*: 387-414.

Kaushal, D and C Naeve (2004) Analysing and visualising expression data with Spotfire. *Current protocols in bioinformatics*: 7.9.1-7.9.25. New York: John Wiley and Sons.

Lv, Y, Y Duan, W Kang, Z Li and F-Y Wang (2015) Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems 16*, no.2: 865-873.

Murray, DG (2013) *Tableau your data: fast and easy visual analysis with tableau software*. John Wiley & Sons.

Nantes, A, D Ngoduy, A Bhaskar and E Chung (2016) Real-time traffic state estimation in urban corridors from heterogeneous data. *Transportation Research Part C: Emerging Technologies 66*: 99-118.

Perez, AS (2012) *OpenLayers cookbook*. Packt Publishing.

Richter, A (2015) Master the challenge of generating OpenDRIVE road networks based on real world data. *8th OpenDRIVE User Meeting*, Stuttgart, 15 October 2015.

Tamminga, G, M Miska, E Santos and S Hoogendoorn (2012) Design of open source framework for traffic and travel simulation. *Transportation Research Record: Journal of the Transportation Research Board 2291*: 44-52.

Xin, W and ES Prassas (2014) Development and implementation of a real-time big-data management architecture for effective adaptive traffic signal control. *21st World Congress on Intelligent Transport Systems: Reinventing Transportation in Our Connected World, ITSWC,* Detroit, US, 2014.