

A Metadata-based System for Urban Transportation Data Provision and Analysis

Vivian Salim, Gerardo Trinidad, Nariida Smith, and Leorey Marquez

Transport Futures Project, Building, Construction & Engineering,
Commonwealth Scientific and Industrial Organisation,
Melbourne, Australia
Vivian.Salim@csiro.au

Abstract: Cities are presently faced with increasing pressure from population and activities leading to congestion, pollution and degradation of the living environment. In order to assist decision-makers to assess the changing state of important amenities in a city such as clean air, energy and transport, a metadata-based system for automated data assembly along with a Web-based suite of applications is proposed. This would form the basis for identifying the most effective technology and practice requirements in line with changing needs. Metadata can be defined as structured data, such as access and content details, about data sets. Metadata sets were designed to allow a more efficient use of data from different sources in different places at different times. Metadata based systems can be designed to assemble required data “just in time” for use. This is particularly useful when a broad range of data is required or when the most recent data is needed due to the current climate of rapid change. Both criteria apply to data for city transportation planning today. The process and proposed system for enabling this technology is described in this paper using a study carried out in Melbourne, Australia of the pollution exposure of city residents due to motor vehicle emissions. Using a metadata system, the most recent data at distributed locations (for instance, weather data maintained by meteorologists and traffic data by transport engineers) may be sourced more efficiently. This system must allow interoperability between remote and distributed databases. It should contain an appropriate search filter, for instance checking for the spatial compatibility of data sets. Examples of how spatial compatibility can be analysed is provided using Prolog. A Web-based application is illustrated which features interactive, seasonal models of air-quality.

Keywords: Metadata, data collation, urban transportation, air quality.

1. MOTIVATIONS

The existing urban problems of congestion, pollution and alienation of the natural environment are being exacerbated with increasing pressure from population and activities. With an increasing proportion of the world’s population living in cities (e.g. 80% of the Australian population), it is important for the health of the population and quality of life that these issues are effectively managed. The urban population in turn contributes to environmental stress through the metabolism of cities, which are sources of global warming and encroaching salinity. This paper focuses on the transportation system, which is a crucial contributor to the livability of a city in terms of air quality, noise, safety, provision of services, efficient energy use and mobility of its residents.

In transportation analysis, there are now strong reasons to take into account the triple bottom line of social, economic and environmental costs and benefits. The data needs are thus diversifying and analysts require comprehensive travel and traffic information by hour of day, day of week or season of the year together with supporting data to estimate a wide range of variables such as:

- emissions and pollution from motor vehicles,
- noise from aircraft or road transport,
- social impacts from road safety,
- employment changes with land use,
- impacts and access for disadvantaged groups,
- economic benefits in terms of value of time savings for different socioeconomic groups.

Much of the data required is not traditional transport or traffic data and is not collected by transport agencies. In Australia, this problem is compounded because transportation information is widely distributed amongst federal and state governments and agencies, various corporations, industry associations and other private sector sources. There is no central site for all collections vis a vis in the U.S, the Bureau of Transportation Statistics; the U.K, a website for transport statistics; and Canada, T-Facts, an electronic library of transportation data and information.

Simultaneously, the third great societal revolution is now occurring, from an industrial society to a knowledge-based society, providing more challenges as well as opportunities for cities. Rapid changes are occurring in mobile computing and communications, with the take up of new e-business technologies growing at an exponential rate. A recent study of transport professionals in Australia suggested that the impact of such technologies would be the most important influence on transport in the next 5 to 10 years [Smith et al., 2001]. Tracking the impact of such changes is difficult. Previously, invariance of data between one year and the next might be in question whereas now, invariance within a year is in-doubt. The need for recent data has become paramount.

The combined requirement for multiple types of data and recent updates is particularly onerous for transportation analysis since it is so location specific. Information seekers encounter multiple problems. First they need to determine whether suitable data exists and how it might be accessed. Then they need to obtain the data in a suitable form and interpret it. Furthermore, information is often needed quickly, increasing problems in sourcing.

Eliot [1999] discussed experience in information locator services emphasizing that data intermediaries have "a primary role as agents of user communities". The paper comments that it makes sense for data content owners to concentrate on managing their data and information resources in the manner that best suits their primary clientele while enabling intermediaries to serve other audiences.

In order to address these issues, this paper outlines the concept of TransPort.au, a demonstration web-portal system for data assembly using metadata.

1.1 Metadata in Data Provision

Metadata sets were designed to allow a more efficient use of data from different sources in different places at different times. These sets are a key resource in the management of the growing

mass of global information in the form of large, distributed, and diversified data archives.

As the range of types of data multiplies, metadata can allow data owners to maintain data control and update. At the same time, it is a method for users of data to dynamically track where data can be found, how it might be accessed, what elements it contains, spatial and temporal time frames of the data sets and what form it is in, and thus whether it is compatible with other datasets.

It is important that standardized metadata is used to describe heterogeneous data sets. Unfortunately as multiple groups address the issue of metadata standardization, multiple standards emerge. For instance, the International Standards Organization (ISO) has recently set up a Metadata Working Group to take responsibility for standards applicable to the specification and management of metadata. The American National Standards Institute (ANSI) also has a data representation standard, and the World Wide Web Consortium (W3C) is developing the Resource Description Framework (RDF) specification.

Urban transportation research will benefit from standardization guidelines that support the geo-spatial aspects of the data since GIS technology will often be used to support different scales of spatial representation. The Australian and New Zealand Land Information Council's (ANZLIC) Metadata Guidelines developed in 1996 are widely accepted in government and the spatial information community. The ANZLIC Metadata Guidelines were influenced by the ISO initiative and will converge to it in future developments [ANZLIC, 2001].

The Australian Spatial Data Directory (ASDD) makes use of this standard in a metadata-oriented initiative dedicated to spatial information [AUSLIG, 2001]. The ASDD aims to improve access to Australian spatial data for industry, government, education and the general community through effective documentation, advertisement and distribution. This paper examines the technical issues involved in going beyond the simple access provision of ASDD or similar directories. Here we propose to use metadata as a platform that allows researchers to obtain an optimised selection of data, using the fullest possible range of available resources.

2. TRANSPORT.AU

The concept of TransPort.au is based on four elements (as shown in Figure 1):

Metadata Repository: Meta-database maintaining information about available data in federal, state or non-government organisations.

Distributed Autonomous Databases: held by different custodians who each allow a description of their data to be maintained in TransPort.au and their data to be accessed either on-line or via other data access facilities.

Data Broker and Task Coordinator: Receives and responds to requests from subscribers, through a WWW based connection, providing raw, processed or analysed data.

Application Suite: from data presentation/viewing and simple statistics to modelling and simulation. An interactive web page featuring results from seasonal air quality analysis of Melbourne, Australia is presented as an example in this paper. Each of the components is discussed in the following sections.

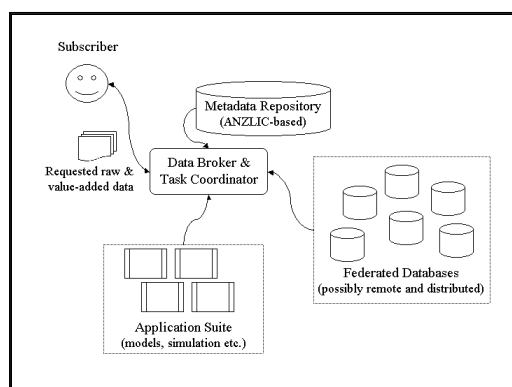


Figure 1. Components of TransPort.au

2.1 Federated and Metadata Databases

The federated databases consist of autonomous databases from different sources and data custodians that could be sited in any part of the world. The metadata repository contains access and content details for those remote or local databases held by organisations who agree to have their data listed.

Agreement to belong to the federated database system for TransPort.au should not compromise the normal uses of these databases in any way. Only a few requirements are made on the data holders/owners:

1. Willingness to provide data to TransPort.au end users. The databases could be accessible on-line or alternatively, appropriate data import/export facilities could be provided. Access privileges could be restricted to a subset of users, for example government departments, and provision to levy charges for access provided.
2. Description of the database using the ANZLIC guidelines for metadata. Many databases would already have the necessary descriptors. Forms

on the WWW may be used to collect the required metadata. The attributes required for spatial analysis is illustrated in Table 1.

3. Update and maintenance of the metadata would be needed in case of changes to data-structures and semi-automated procedures could be put in place to facilitate this.

2.2 The Data Broker

Using metadata to assemble data for specific research tasks, such as monitoring and forecasting emissions from motor vehicles, consists of the following sub-tasks:

1. Achieving interoperability between the Web-portal and distributed database systems.
2. Identifying the presence and availability of data sets for the tasks at hand,
3. Integrating any heterogeneous data sets,

The first sub-task, global data sharing in a distributed environment may be accomplished via a wide range of solution methods. XML (eXtensible Markup Language) may be used as interchange format between database tables, XML documents on the Web and Prolog clauses. For further details, the reader is referred to Smith et al. [2002].

Interoperability allows querying the metadata repository and exploiting existing web technologies, including intelligent agents. Once the data sets have been found, the second sub-task entails integrating any disparate data sets.

The form of spatial or temporal compatibility of the data sets needs to be analysed before integration. We now show how analysis of spatial compatibility and integration of data sets may be automated via the Resource Description Framework (RDF), a framework for describing metadata, and Prolog, a programming language based on predicate logic. PROLOG was the first logic programming language and is still commonly used today, particularly in inference engines within larger programs. The RDF triple is one of several formal models offered by the core RDF specification [W3C, 2001]. It consists of an ordered 3-tuple of “subject (or resource)”, “predicate (or property)” and “object (or value)”.

For example, say that the data set, “Hourly Population” consists of hourly population distribution estimates for a region. “Hourly Population” is the subject or resource, “spatial extent” is a property and “Melbourne” is a value of that property. The RDF triple is then:

The “spatial extent” of “Hourly Population” is “Melbourne”.

The RDF triple as described previously can be represented in standard PROLOG notation of the form:

```
spatialExtent (“Hourly Population”, “Melbourne”)
```

Basic Attributes	
Element name	Definition and Usage Notes
Dataset Title	A unique, informative title for the dataset.
Creation Date	Date when the dataset was brought into existence.
Last Revision	Date when the dataset content was most recently revised.
Abstract	A brief narrative summary of the content of the dataset.
Parameters List	Parameters (variables) represented in the dataset.
Keywords	Keywords assigned by the author or custodian of the dataset.
Spatial Attributes	
Element name	Definition and Usage Notes
Spatial Representation	Method used to spatially represent geographic information. Use "text/table" for spreadsheets, databases or text files as in XML, "vector" for point/line/polygon/volume information stored by coordinates in a GIS system.
Spatial Extent	Coordinates defining the spatial extent of the dataset expressed in longitude and latitude, in decimal degrees (negative for south and west directions). Provides a coarse, "first pass" set of coordinates for rapid spatial searching and/or display.
Inclusion Polygon	Boundary enclosing the dataset, expressed as the closed set of (x, y) coordinates of the polygon (last point replicates first point). Use to provide a machine-readable definition of a finer, irregular shaped region within which the actual dataset occurs.
Exclusion Polygon	Boundary enclosing exclusion area, expressed as the closed set of (x, y) coordinates of the polygon (last point replicates first point). Use to designate significant "holes" in the polygon described above in the inclusion polygon.
Reference System Name	Name of the spatial reference system used in the dataset. Eg GDA94 (Geocentric Datum of Australia); AGD66/84 (older Australian Geocentric Datum); WGS84 (World Geodetic System); GRS80 (Geodetic Reference System 1980).
Data Quality Attributes	
Element name	Definition and Usage Notes
Dataset Status	The status in the completion of the dataset. Eg completed; ongoing; under development; planned; historical; archive; obsolete.
Maintenance, Update Frequency	Frequency of changes/additions made to the dataset after completion. Eg continual; irregular; not planned; unknown;
Completeness	A brief assessment of the completeness of coverage, completeness of classification and completeness of verification of items in the dataset. Typically used to flag any data gaps or poor quality data within the dataset. Statements should be explained in as quantitative a manner as possible.
Logical Consistency	A brief assessment of the degree of adherence of logical rules of data structure, attribution and relationships. Data structure can be conceptual, logical or physical. Typically used to report the results of any tests run on the dataset for logical consistency - e.g. tests whether values are within valid ranges; tests whether internal logic is obeyed (e.g. no start dates earlier than end dates); tests whether referenced objects all exist; tests for double entries or missing data elements; plus specific logical consistency tests for GIS vector data (are all polygons complete, etc.).
Temporal Accuracy	Accuracy of the temporal attributes and temporal relationships of features (as quantitative as possible).
Parameter Accuracy	A brief assessment of the reliability assigned to features in the dataset in relation to their real world values (as quantitative as possible).
Lineage Statement	A brief history of the source(s) and processing steps used to produce the dataset.
Usage and Access Attributes	
Element name	Definition and Usage Notes
Classification	Handling restrictions imposed on the dataset for security concerns. Eg unclassified; restricted; confidential.
Access Constraints	Any restrictions of legal prerequisites that may apply to the access and use of the dataset including licensing, liability and copyright. Eg intellectual property rights; copyright; patent; patent pending; trademark; license restricted; public domain.
Dataset Usage	Provides basic information about specific application(s) for which the dataset has been or is being used by different users.
Usage Limitations	Applications, determined by the custodian for which the dataset is not suitable (if applicable).
Available Medium	The format in which the dataset is available. Eg online; off-line; electronic (CD, disk, tape etc.); hard copy.
On-line Access	URL link to www-accessible interface or file which permits the data to be remotely queried or accessed.
Contact Info	Basic information such as contact person(s), position(s) and organization(s) whom queries about the dataset can be directed.
Mailing Address	Provide complete mailing or postal address.
Electronic Mail	Provide complete email address.
Telephone	Include country and area codes.
Facsimile	Include country and area codes.

Table 1. Metadata Attributes for Spatial Data

If its spatial unit is "Census Collection District" or "CCD", then in PROLOG notation, we have:

spatialUnit ("Hourly Population", "CCD")

Similarly, the following statements in Prolog say that the data set "Resident Population" is a census on the resident population of Australia with the spatial unit of the data being "CCD".

spatialExtent ("Resident Population", "Australia")
spatialUnit ("Resident Population", "CCD")

Further, say we have the tables "Ozone Level" and "Particulates" which are described as:

spatialExtent ("Ozone Level", "Melbourne")
spatialUnit ("Ozone Level", "Grid")
spatialExtent ("Particulates", "Sydney")
spatialUnit ("Particulates", "Grid")

Using Prolog, simple inference rules may be formulated which determine the appropriate method for integrating two data sets. For instance, the following set of rules performs "Spatial Join" which retrieves and joins all pairs of data sets with

overlapping extents and having the same spatial unit.

method (Data1, Data2, "Spatial Join") ←
 spatialExtent (Data1, Extent1) &
 spatialExtent (Data2, Extent2) &
 overlap (Extent1, Extent2) &
 spatialUnit (Data1, Unit1) &
 spatialUnit (Data2, Unit2) &
 Unit1 = Unit2.

The set of rules below performs "Spatial Interpolation" and retrieves and interpolates pairs of data sets with overlapping extents but with different spatial units.

method (Data1, Data2, "Spatial Interpolation") ←
 spatialExtent (Data1, Extent1) &
 spatialExtent (Data2, Extent2) &
 overlap (Extent1, Extent2) &
 spatialUnit (Data1, Unit1) &
 spatialUnit (Data2, Unit2) &
 not (Unit1 = Unit2).

Substituting in the appropriate values for Data1, Data2, Unit1, Unit2, and Extent1 and Extent2 in the above, it is now possible to determine that using these two methods:

- "Hourly Population" can be integrated with "Resident Population" using a simple spatial join method.
- "Hourly Population" can be integrated with "Ozone Level" using a spatial interpolation technique.
- It is not possible to integrate "Hourly Population" and "Particulates".

The process above is done without human intervention.

2.3 Application Suite: Air Pollution

Motor vehicles contribute heavily to the urban air pollution load in developed countries. Estimates for Melbourne Australia, a city of 3.5 million people suggest that, 83% of CO, 63% of SO₂, 41% of VOC and 16% of PM₁₀ emissions in 1996 were due to motor vehicles. An Australian survey in 1997 showed urban air pollution as the environmental issue of greatest concern to urban citizens [Smith, 1997]. Pollution is related to motor vehicle emissions via the mechanisms of the urban air shed. Concentrations vary across the city depending upon the weather. Where interest in human health and comfort is uppermost, the spatial distribution of the urban population must be considered in conjunction with the pollutant distribution.

Marquez et al. [2002] conducted a case study for Melbourne. This study showed the advantages of jointly modelling emission rates and distribution across the city, then distribution of air pollution, via an urban airshed model, and finally population exposures based on activities. It was found that the major pollutant hazards vary markedly between summer and winter. Hence, seasonally sensitive

models can be of value in selecting alternative measures to limit pollutant impacts.

The results of this case study may be viewed online at <http://www.dbce.csiro.au/biex/cityvital/> (current as at June 2002). Figure 2 illustrates the Web application that presents the findings of that case study. A drop-down combo box is used to display the spatial distribution of NO₂ or O₃ in winter or summer, or to show the seasonal population distribution in Melbourne.

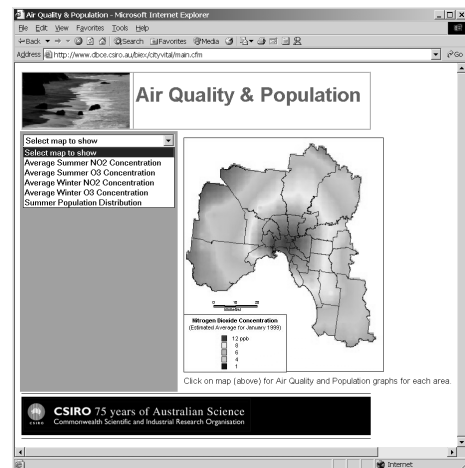


Figure 2. Seasonal Air Quality for Residents of Melbourne

Clicking a region opens a window containing the graphs of NO₂ and O₃ levels for that area in winter and summer by time of day and a graph depicting the hourly population variation in winter and summer (see Figure 3).

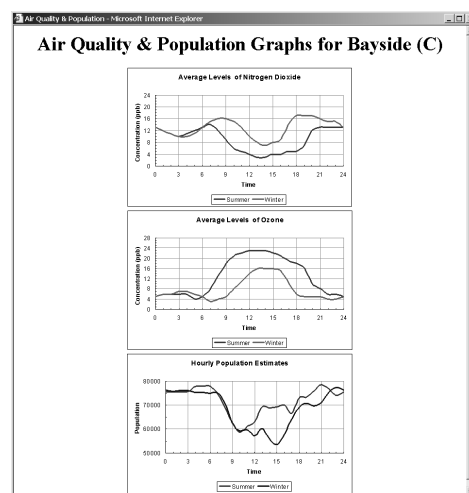


Figure 3. Graphs of Hourly Levels of Pollutants and Population Distribution in the Bayside Region

3. CONCLUSIONS

This paper described a conceptual metadata-based Web portal for "just-in-time" or "when needed"

data provision. This portal can also assist analysis needs through an application suite, as illustrated here for the examination of emission patterns due to motor vehicles in Melbourne, Australia.

3.1 Implementation Issues

The system is currently in its conceptual phase, with some technical issues examined in order to ascertain scientific feasibility. Despite the availability of technology, the portal would only be possible with the cooperation of data custodians. Data suppliers range from national and state government bodies, to non-government agencies to various industry corporations, research organizations, and individual private sector sources. The desire and/or ability of these data suppliers to provide and maintain metadata depend upon the primary function of the organization and their access to resources or funds for that purpose. Large government departments, such as census bureaus, have significant motivation for supporting a suitable metadata framework. However, other agencies may not have the resources or incentive to do so.

In view of this, the next steps in the development of this project are:

- Contact stakeholders in government and industry across Australia to identify key databases for the initial system together with information on current access conditions and privacy issues.
- Engage potential users in a web-based survey which includes trial access to the prototype. The aims are twofold, to estimate market interest in various system features to allow user responsive design and to gauge the willingness to pay.

3.2 System Benefits

A system such as TransPort.au is particularly timely due to a number of reasons. As shown in this paper, the technology for it is now available. In fact, the TEMSIS project uses a metadata information system and integrates distributed information platforms to facilitate remote cooperation of local environmental authorities in Germany and France [Denzer et al., 2000].

There is a need for this sort of system in any country where resource constraints affect data supply. This is particularly true in the study of the urban environment, which is interdisciplinary in nature and requires a broad range of data. At the same time, new technology generates large amounts of data. Frequent updates of data may be needed, which would be costly for the end-user to purchase "just in case". The metadata system allows the most recent data at distributed locations to be sourced efficiently.

The system would therefore greatly ease and semi-automate the data seeking task for end-users, ensuring that the best available data for their purpose is obtained. The portal would not replace or compete with current data sources. Data holders maintain control over their own data and in some cases, could gain some revenue from use of their data. Through analysing metadata, developers can gain a greater insight into the needs of end-users in establishing the relevance of a particular set of data.

Increasing the range of data that can be accessed and the range of people who can access it should significantly improve analysis for the triple bottom line of social, economic and environmental costs and benefits.

4. REFERENCES

- ANZLIC, The ANZLIC Standard, Feb 2001, <http://www.anzlic.org.au/asdi/metgidv2.pdf> Accessed 26 October 2001.
- AUSLIG, 2001, Australian Spatial Data Directory <http://www.auslig.gov.au/asdd/>, Accessed 29 August 2001.
- Denzer, R., Guttler, R. Houy, P., and TEMSIS Consortium, "TEMSIS - a transnational system for public information and environmental decision support", *Environmental Modelling and Software*, 15, 235-243, 2000.
- Eliot, C., "Experiences with Information Locator Services", *Journal of Government Information*, 26, 3, 271-285, 1999.
- Marquez, L., Smith, N., Trinidad, G., and Guo, J., "Activity Patterns and Pollution Exposure: A Case Study of Melbourne", *European Journal of Transport and Infrastructure Research*, 1, 4, 371-390, 2001.
- Smith, N., "Economics of Benefit and cost, Urban Air Pollution in Australia, An Inquiry by the Aust. Acad. of Tech. Sciences & Eng., 1997, http://www.ea.gov.au/atmosphere/airquality/urban-air/urban_air_docs.html, Accessed 5 November, 2001.
- Smith, N., Trinidad, G., and Salim, V., "Relevant Data for Transportation Analysis: A Search Framework using Metadata", presented at The 81st Annual Meeting of the Transportation Research Board, Washington D.C., USA, Jan. 13-17, 2002.
- Smith, N., Ferreira, L., and Mead, E., "Stake Holder Insights: Working Paper 4, NTS Study of E-business and Transport", February 2001, <http://www.nts.gov.au/media.htm>, Accessed February 2002.
- W3C. Semantic Web Activity: Resource Description Framework (RDF). 8 June 2001. <http://www.w3.org/RDF>. Accessed 19 July 2001.