COMPUTING
SCIEN
CE

UTMC Compliant Database to Support Technologies of the Future.

M. Bell, V. Suresh, P. Watson and P. Blythe

TECHNICAL REPORT SERIES

No. CS-TR-1085 April, 2008
Abstract

This paper will present the results of an in depth investigation of the standard database format defined by the Common Database project, UTMC10, funded by the Department for Transport on the mid 1990s applied to diverse datasets. It outlines the difficulties encountered in implementing the database platform to embrace different sources, (namely accident, pollutant concentrations, registration plate, journey time etc.) and to include both dynamic (vehicle tracking, tailpipe emissions, bus stop arrival times, car park occupancies etc) and static (public transport time tables, car parking charges etc.) data with different dimensions, sampling frequencies and accuracies. The paper will look at the requirements of Urban Traffic Management and Control compliant database infrastructure in the context of the delivery of wider policy objectives to those relating to traffic. In particular, to create the statistical processing capability that defines the current and changing status of the network, to provide the platform upon which to build a historic picture of the network to allow a statistical forecasting capability and to develop the foundations for model validation and evaluation. In this way the infrastructure needed to deliver policy objectives which not only embrace traffic management but also public transport, air quality, health impacts and climate change will be formulated.
Abstract

This paper will present the results of an in-depth investigation of the standard database format defined by the Common Database project, UTMC10, funded by the Department for Transport on the mid 1990s applied to diverse datasets. It outlines the difficulties encountered in implementing the database platform to embrace different sources, (namely accident, pollutant concentrations, registration plate, journey time etc.) and to include both dynamic (vehicle tracking, tailpipe emissions, bus stop arrival times, car park occupancies etc) and static (public transport time tables, car parking charges etc.) data with different dimensions, sampling frequencies and accuracies. The paper will look at the requirements of Urban Traffic Management and Control compliant database infrastructure in the context of the delivery of wider policy objectives to those relating to traffic. In particular, to create the statistical processing capability that defines the current and changing status of the network, to provide the platform upon which to build a historic picture of the network to allow a statistical forecasting capability and to develop the foundations for model validation and evaluation. In this way the infrastructure needed to deliver policy objectives which not only embrace traffic management but also public transport, air quality, health impacts and climate change will be formulated.

About the author

Prof. Margaret Bell - Science City Professor of Transport and Environment, TORG, School of Civil Engineering and Geosciences Newcastle University

Lakshmi Suresh joined the computing science department as Research Associate in the MESSAGE project in November 2006. She received her Bachelor of Engineering degree in 1996 from the Bharatidasan University, India and the Master of Science (SDIA) with distinction in 2006 from the Newcastle University, United Kingdom. After completing her bachelor’s degree she worked for LogicaCMG and Indigo4 Systems United Kingdom in client server technologies. During her post graduation studies, Lakshmi worked for PrismTech in the Open Splice DDS, a middleware addressing publish-subscribe communications for real-time and embedded systems.

Paul Watson is Professor of Computer Science and Director of the North East Regional e-Science Centre. He graduated in 1983 with a BSc (I) in Computer Engineering from Manchester University, followed by a PhD in 1986. In the 80s, as a Lecturer at Manchester University, he was a designer of the Alvey Flagship and Esprit EDS systems. From 1990-5 he worked for ICL as a system designer of the Goldrush MegaServer parallel database server, which was released as a product in 1994. In August 1995 he moved to Newcastle University, where he has been an investigator on research projects worth over £13M. His research interests are in scalable information management. This includes parallel database servers, data-intensive e-science and grid computing. In total, he has over thirty refereed publications, and three patents. Professor Watson is a Chartered Engineer, a Fellow of the British Computer Society, and a member of the UK Computing Research Committee.

Suggested keywords

UTMC COMPLIANT,
COMMON DATABASE
Keywords: UTMC Compliant, Common Database,

Abstract (REVISIT)

This paper will present the results of an in depth investigation of the standard database format defined by the Common Database project, UTMC10, funded by the Department for Transport on the mid 1990s applied to diverse datasets. It outlines the difficulties encountered in implementing the database platform to embrace different sources, (namely accident, pollutant concentrations, registration plate, journey time etc.) and to include both dynamic (vehicle tracking, tailpipe emissions, bus stop arrival times, car park occupancies etc) and static (public transport time tables, car parking charges etc.) data with different dimensions, sampling frequencies and accuracies. The paper will look at the requirements of Urban Traffic Management and Control compliant database infrastructure in the context of the delivery of wider policy objectives to those relating to traffic. In particular, to create the statistical processing capability that defines the current and changing status of the network, to provide the platform upon which to build a historic picture of the network to allow a statistical forecasting capability and to develop the foundations for model validation and evaluation. In this way the infrastructure needed to deliver policy objectives which not only embrace traffic management but also public transport, air quality, health impacts and climate change will be formulated.

1 Introduction

Traditionally since the late 1960’s the policy for managing networks has been to reduce delay to traffic. This policy has led to increase in capacity of our networks which in turn has enabled growth, not only in the number of cars, but also in their use. This has led to a detrimental impact on our environment, namely air quality, noise and carbon emissions. The EU set out to develop the overall strategy for air quality and in 1996 the Framework Directive 96/62/EC on ambient air quality assessment and management introduced new air quality standards [8]. A series of daughter Directives [9][10] followed which were concerned with long-term ambient air quality for twelve substances including sulphur dioxide, SO₂, Nitrogen dioxide, NO₂, particulate matter of less than ten microns, PM₁₀, lead, Pb, Ozone, O₃, Benzene, arsenic, nickel, mercury, Polycyclic Aromatic Hydrocarbons, PAHs, Carbon monoxide, CO, cadmium, Cd. In response to this the UK [4] set up legislative Framework for review and assessment of air quality [6][7]. As a result LAs identified Air Quality Management Areas (AQMA), where air quality objectives were being or likely to be exceeded by the Year 2005. By integrating air quality considerations with the Local Transport Plan remedial measures are being delivered along with the provision of air quality information and relevant data to the public [5].

Historically, noise from traffic was caused by free flowing traffic on trunk roads and motorways, DfT [3]. Nowadays, noise problems manifest themselves in urban areas with continuously varying noise throughout the day due to congestion and intrusive noise problems in the early hours of the morning and late at night often caused by commercial vehicle deliveries [1]. Recent European Directives,[11] and associated methodologies WG-AEN (2006) have resulted in a control measure that is a weighted sum of noise levels measured for four hours in each of three periods of the day; namely, day, evening and night. This measure recognises the need for lower levels of traffic noise for evening leisure activity and to avoid sleep disturbance overnight.

The Stern Report [2] provided a wake up call by translating the impact of the traffic pollution on climate change in the context of its implication on our economy. The Stern Report sets out an economic and moral case for taking immediate steps to deliver the 60% reduction of carbon dioxide CO₂ emissions required by 2050. Evidence was presented that suggests that at a cost of ~1% (higher for developed countries) of global GDP invested now would provide a benefit of avoiding 5-20%
hit on GDP if steps are not taken. At the Smart Environment Interest Group Conference in London in November 2007, Graham Pendlebury endorsed the number one priority of the environmental policy of the UK Government was cutting carbon as a global issue acknowledging that air quality, noise, landscape, biodiversity etc seen as quality of life issues usually having a local rather than national impacts. This fact was highlighted in The Eddington Report [12] which acknowledged that transport is vital to the economy, the basic connectivity is good but the main problem is localised but acute crowding or congestion, leading to unreliability. Eddington set out the priorities for congestion management in key urban areas, inter-urban corridors and international gateways.

Congestion problems on motorways are being eased by the implementation of the Automatic Traffic Management, ATM, with measurable environmental benefits, however this improvement is being delivered through more efficient management of capacity which is likely to result in a spreading of the peaks. Against different objectives, such as managing demand, the same technology (ATM) could be used proactively in reducing the volumes of traffic on the road, a measure that is required to address climate change [1]. In urban areas, Transport Infrastructure Fund Partnerships allow authorities to work with the Department for Transport to develop business cases to secure funding, in competition with other Local Authorities, to deliver innovative schemes that aim to reduce congestion.

Local Authorities being responsible for delivering the UK policy regarding congestion have sustainability at the core of their policies. For example Leicester LCC [13] is: “A sustainable city recognised as a model of excellence internationally, which minimises its impact on the local and global environment, and where no-one suffers from serious economic disadvantage”. At the heart of Leicester's aims, actions and measures to tackle congestion is:

- To facilitate more bus, walking and cycling trips, whilst improving network efficiency and the effective allocation of road space

With key goals:

- To ensure that national air quality standards are met, and increase awareness and understanding of air quality issues.
- To promote the prudent use of resources, in particular through patterns of development and transport that make efficient and effective use of existing infrastructure
- To manage the impact of traffic on the global environment and health

In order to achieve these policy objectives Leicester acknowledges the need:

- To realise modular open systems for Urban Traffic Control and Management, UTMC, and testing of tools
- To create a Decision Support System for decision makers and the public
- To use of Information Society Technologies to integrate applications

Fundamental to the delivery of these policy objectives is to enable real-time control of traffic systems and transport operations. Also, in order to affect a substantial modal shift, emphasis will be to operating the network from the user’s perspective, optimising routes for multi-modal travel in a seamless way. This opens up substantial academic challenges to create an UTMC compliant database architecture that will function in real time processing data from diverse monitored and modelled sources. This will require a data platform that can manage huge volumes of information which continuously changes spatially and temporally from both static and dynamic sources.

**Emerging Technologies**

The project MESSAGE (Mobile Environmental Sensing System Across Grid Environments) jointly funded by the EPSRC and DfT plans to deliver pervasive sensors which will co-exist with legacy systems to considerably enhance the monitoring capability across a network. The sensors monitor, vehicle occupancy, position, movement, temperature, humidity, noise, CO and NO2, and have a data logger and zigbee low cost, low-power, wireless mesh networking standard to collect data through a gateway to a remote central server. These inexpensive monitoring sensor systems operate for one year with the same battery and the measured signals are stored in their raw state. The data is calibrated, using the manufacturer's specifications, quality assured and cleaned. In this way the MESSAGE creates a data platform to inform, understand, assess and evaluate impacts of policies such as:

- Delivering accessibility,
- Tackling congestion,
- Safer roads
- Better air quality
- Reduced noise and carbon emissions
- Implementation of ITS

The MESSAGE database aims to give integration and coherence of data and analysis of the pervasive and legacy monitoring systems to provide network state estimation and be able to communicate with air quality, noise and carbon emissions models, supply data for validation, supporting the formulation of scenarios to resolve network problems and to evaluate the impacts of changes. As the data builds up over time a data warehouse stores the processed data which allows data mining, and knowledge discovery to enrich the fundamental understanding of the movement within the
network providing smarter management and modelling to deliver government policy.

**Standardisation of data**

The data from the UTMC and pervasive monitoring systems has first of all to be ratified and incorrect data needs to be flagged. Statistical methods of extrapolation and reference to historic records allow values to be assigned to missing and to substitute bad data. The policy is to store the data in its raw format for research purposes. Statistical methods have been developed to automatically carry out the ratification. In order to associate diverse data sources with each other and to provide consistency of interpretation of the data across all sources, there is a need to ensure that the units and the dimensions data are the same along with a measure of the reliability. For example traffic flow from remote acoustic sensing, axle counters, SCOOT, Automatic Number Plate Recognition, ANPR etc, need to be made comparable by post processing using empirical relationships derived based on validation exercises. Using direct observation surveys algorithms will be developed to ensure congestion measures based on occupancy from remote acoustic sensing are comparable with a congestion measure based on loop detector systems such as from SCOOT system. These surveys for the MESSAGE acoustic will be complete by the end of April and will be presented at the conference.

The value of integration of data into one database platform is to be able to establish the cause and effect of changes made to the network or events that occur. Therefore, another issue is synchronisation of data records in time of occurrence and for a fixed duration. As the data sources are from independent detector systems often the start time and end times of reporting periods and sampling frequencies are different. For example, when the SCOOT system is rebooted, the data collection will start at a random point in time. This means that by using weighted averages, data is synchronized over sampling period time, chosen to be on 5 minute intervals from a specific point in the master clock defined to be 00:00hrs GMT. The individual vehicle data (ANPR) has to be aggregated to flows in into the same fixed 5 minute periods, 1 minute data (air quality). When data is available at lower frequencies (meteorological conditions over 10 minutes) then data can be duplicated or extrapolated.

In order to compare traffic data in different parts of the network the data associated with road sections need to be standardised to a geographic layout of the network. In the context of UTMC compliance this can be difficult due to the inconsistency with which networks are defined for monitored and modelled data. For example traffic flow associated with a SCOOT detector is assigned to a generic link defined as stop line to stop line. Specific traffic monitoring counters and ANPR monitor measure flow at a point. Often these flows are used to validate modelled data. However we move to the modelling framework, the flows associated with traffic models have a different node link specification to those of air quality models which usually are completely different from the SCOOT node link diagram. Increasingly, probe vehicle data from sensors is being used for management decisions. In this case, second by second data may be available for vehicles moving along roads (bus tracking) or alternatively a snap shot of a vehicle position is available (ISIS). If the ultimate aim is to find a way to seamlessly associate modelled data with measured data and in the future move from a system to an individual entity optimisation then it is important to have a standardised network layout specification within the UTMC compliant database.

**MESSAGE Database architecture**

The architecture of the MESSAGE is illustrated at a simple level with the real-time UTMC Compliant database at its centre in the Figure 1.

![MESSAGE Database Architecture](image)

**Figure 1:** Simplified Systems architecture for the MESSAGE project monitored data and modelled data. QA= Quality Assurance, APPS=Applications
The database needs to handle massive volumes of dynamic, static and geographically distributed data. In MESSAGE, with reference to figure 1, the approach has been to separate the database into a real time component which is UTMC compliant and the Data Warehouse. The raw data from various traffic data sources like SCOOT, AURN, loop detectors, accidents, car parks, traffic signals and pervasive sensors, has to be stored in a standard way. The UTMC standards designed for the DfT, has been adopted in MESSAGE as the platform to integrate the disparate data sources.

**REAL TIME DATABASE:** This is UTMC compliant so that it can interface with available legacy systems and emerging technologies. The data from the sensors are deposited in the database with minimum time delay.

As a first step in the database design, metadata provides a generic description of the device and configuration information is followed by the dynamically monitored data. Using air quality as an example, the generic data will include type of equipment, location for example with respect to kerb side and measurement accuracy. Configuration data relates to the thresholds as defined by the Department of Environment, Food and Rural Affairs (DEFRA). Dynamic data is the concentrations of pollutant for example CO in ppm. In addition there are support data for the air quality measurements identifying faults, nature of the fault, device history etc. Sensors in MESSAGE are dynamic as well as static, the GPS data (Northing and Easting) will be recorded along side the dynamic air quality measurements.

As a first step in the data capture is to clean the data from various sources as described in previous section. The raw data is stored in the same format as it is received and stored in the warehouse for the research purposes. Second step is to average and synchronise the data and process in to UTMC compliant format. This data can be stored in the real time database for typically over the recent 12 months.

**DATA WAREHOUSE:** Data warehouse is the storage for the quality assured historic and current raw data. It fulfils the core statistical processing and analysis of data for evaluation. In the MESSAGE project, the main demonstration will use the LEICESTER historic data held in the instrumented city facility.

In order to efficiently process the large volumes of data available for Leicester a more fundamental aspect of data organisation will be developed to reply to various questions that are posted on the data. The importance of the data warehousing architecture is its ability to scale to accommodate: the growth of data as it is continuously loaded; the increase in reporting requirements and the need to address various business or scientific needs demanded of the database. It is important to note that a warehouse may in fact be several linked databases attached to different monitoring or modelling systems. Therefore, it is not necessary for these to be UTMC compliant. However, the outputs of the statistical processing that occurs within the warehouse have to be UTMC compliant in order to be presented to the real time database for comparison with the current data.

The warehouse architecture for MESSAGE will mainly focus on organising the data that is both measured and modelled and includes meteorological conditions, pollution and traffic. The data has to be dimensioned according to time, device, location and fleet. The data warehouse is built based on the summary levels required, the users needs from the specific datasets, indexing and partitioning to finally achieve an efficient data structure. Various workload analysis help to decide where an index needs to be created, where the tables should be combined and where summaries should be created.

**Statistical Processing:** This has many facets which depend on the use to which the data is put. For example, standard statistical processing allows typical daily profiles to be derived for different data sets such as traffic flow, delay, congestion, CO, NO\textsubscript{2} etc. Data mining is a statistical process of analysing the data to find the intelligence in the data and uses various statistical and machine learning techniques to discover the trends and patterns of traffic or pollution building up over time and location. This knowledge improves the quality of the decision making. Application of statistical tests will enable the changes occurring in the network over time to assess the impacts of traffic management measures.

**Evaluation:** Traffic and air quality models are used to explore solutions to be implemented to solve problems occurring in the network. Examples include relocation of queues to open space to enable the natural ventilation of the built environment to disperse toxic tail pipe emission, implementation of park and ride to affect a model shift to address heavy congestion along the radial route in to the city.

The data warehouse plays an important role in supplying data to validate the base case and following the implementation of the traffic management measures to evaluate the actual changes taking place on street in the network. In addition, the measured data can be compared to the modelled data in order to establish the success with which the model has provided the correct solution to the problem.
**DATA MARTS:** Data marts are subject-specific or application-specific subsets of the main data warehouse. The major difference between the data mart and the data warehouse is the scope of the information they contain. The data mart represents a smaller component of the data warehouse therefore it implementation time is shorter.

In the prototype MESSAGE architecture applied to Leicester data marts will be implemented namely pollution and traffic. With reference to figure 2, the pollutants like CO, NO2, NOX, NO, particulates, ozone and CO2 are recorded. The data are summarised over various levels depending on dimensions such as time location, meteorological conditions and devices. For example if we consider time dimension, various grouping conditions are formed based on the day, week, month, season, year, bank holidays and so on. Thus various subjects for analysis are created and pollution profile for different conditions are analysed. This kind data organisation is called “STAR SCHEMA” with the FACTS listed vertically holds the main information and the “DIMENSIONS” displayed horizontally explain each value of the dimension and can be joined to the fact tables as needed.

**REAL TIME EVENT PROCESSOR:** The real time traffic data from sensors and other sources like SCOOT or detectors have a very high data flow rate and the volume of data received every second is quite high. For few seconds, a raw data will cross hundred thousand records for a particular region. To analyse the data in real time, various aggregations, grouping and comparisons has to be made. The most recent data which is accumulated within the last few minutes and historic profiles has to be compared with the real time streaming events. The process of storing the information and in few seconds or minutes, querying the database to retrieve the information will have considerable impact in the performance of the database. Some novel methods of processing the real time events have to be adapted.

Let us consider a simple use case of how congestion builds in a network with respect to time and space. This requires the measure of congestion from various data sources like SCOOT, pervasive sensors and remote detection monitored in real time each five minutes to be compared with the average congestion profile from the statistical analysis of data stored in the database to determine when and where the congestion levels exceed a threshold specified by the traffic signal control operator. To achieve this for a city, millions of records of data need to be scanned in seconds.

In MESSAGE, we are intending to implement a prototype to demonstrate this ability of real time processing of diverse data sets.

**SCENARIO MODELS:**

These models are deterministic in nature similar to the mathematical models in which parameters and variables provide a replication of the current state of the network commonly referred to as the base case or a modelled scenario. In our architecture, the data described in the scenario is specified in the data marts and fed into the traffic model. This is then used to predict the traffic flow, delays, stops and congestion expected when the scenario is implemented.

**ONLINE MODELS:** Online models can take two forms. The first include empirical models which result from data mining and statistical processing within the warehouse. An example would be an estimate of PM2.5 from measured CO. Secondly, with advancements in the computational processing power micro simulation models can produce results in real time. Online models have an important role in enhancing the information presented to the operator to enrich his decision making capabilities.

**Challenges of UTMC compliance**

Several difficulties have been encountered during the implementation of UTMC standards to data captured in the wider context of the MESSAGE project to create the integration needed to deliver the UK government policy. Some of these will be described briefly in this section. The specification of nodes and links across the diverse networks of SCOOT, traffic, air pollution, noise etc. is unclear. It is essential that the conceptual network map directly on to the standard geographical co-ordinates. This is particularly important when synchronising spatial data from dynamic with static sources and associating
point measurements to entire links. In the case of noise and air pollution modelling in canyons, widths of roads and heights of buildings introduce further complication.

In the official UTMC compliant handbook [16] the term emissions (which relates to that measured in the tailpipe) is confused with concentrations (which is the level of pollution actually measured at a location resulting from the tail pipe emissions)

Given the diverse data sources and the various types of systems available in the market to make the same and similar measurement, it is suggested that a standard protocol is agreed to label the data within the database. A protocol for differentiating modelled from measured data needs to be addressed also.

Finally the concept of UTMC compliance in the context of a database schema itself has been found to be difficult to fully understand. The availability of test data set from various data sources to illustrate the various elements of the schema would prove to be very useful to many.

CONCLUSION

This paper has made a start in creating a schema which will allow real time processing of diverse data sources to extract useful knowledge regarding the state of the network, to inform models to generate solutions to manage detected or forecasted events and to evaluate the impact of actions taken by an operator. The schema has two main components, the real time database and the data warehouse. It is essential that the former is UTMC compliant in order to interface with the hardware and software currently used by the local authorities. To maximise the benefits of the data sources provided by the legacy systems which have their own databases (often not UTMC compliant), MESSAGE recognises the need for a protocol for a data communication channel to seamlessly pass datasets between different components of the systems.

Acknowledgements

EPSRC and DfT for funding the message project. Leicester and Gateshead city council for supporting data access and through out the demonstrations.

References


[13] [see Annex A – Local Development Framework for Leicester – Core Strategy – DTM 12.07.06]


[15] Oracle 10g Data Warehousing (Paperback) by Lilian Hobbs (Author), Susan Hillson (Author), Shilpa Lawande (Author), Pete Smith (Author)