University of Cambridge

Imperial College London

The National Transport Data Framework

Professor Peter Landshoff
Professor John Polak
March 2008

Contents

The National Transport Data Framework

Final Report

Executive Summary

The NTDF is designed to be a resource for data owners to deposit descriptions into a central catalogue, so that people can search for data and find data and understand their characteristics. The need for such a facility was identified in early 2005 following discussions with the DfT, the Treasury, rail operators and experts on road transport. The value of this is to individuals, to commercial organizations, and to public bodies. For example, services that provide better information to travellers will help to make their journey less stressful and persuade them to make more use of public transport. Transport operators need very diverse information to help them plan developments to their services: demographic, geographical, economic etc. And policy makers need a similar range of information to help them decide how to divide their budget and afterwards to evaluate how valuable it has been. The NTDF is a part of the wider issue of the need to address data access, usage, and ownership and the transition from data into knowledge. A development from the NTDF is the National Transport Information Incubator (NaTII) which is piloting a way to identify business cases for data federation.

The NTDF is a repository for information about the data rather than a store of the data themselves; the data remain in the care of the owners of the data, who are responsible for their accuracy and for updating them. Data owners must be able to lay down their own rules for access, giving different privileges to individuals or classes of individuals with the confidence that they will be adhered to. By converting descriptions of the data sets into a semantic web language, the NTDF system then produces metadata in a standard form regardless of the original format, which is searchable by the powerful search language SPARQL.

The NTDF makes use of the experience derived from the government's eScience programme to provide tools to search for what relevant data may be available for a particular application, and to understand their character. Conventional approaches to access to large data sets rely on rigid classifications of the data. They are designed to suit a particular application and it is not easy to adapt them for other uses. The architecture of the NTDF is flexible, based on what is known as the semantic web. The semantic web is a set of technologies governed by the World Wide Web consortium which allows a rich description and consequently better search of diverse data sources. The NTDF provides a tool for data owners to write a description of their data in the powerful formalism Resource Description Framework, or RDF. The diverse data sources which can be described include legacy data formats as well the numerous types of XML, allowing applications to be built which bring together data from disparate sources. RDF is increasingly being used as the formalism for metadata, and data sources themselves are now being created in RDF too. The semantic web relies on links between related concepts. For example, it can easily be taught that a search for cash machines is equivalent to a search through many databases that list banks, hotels, petrol stations or other locations where such facilities might be found and the system can also flag those machines which charge a fee or may do so because of location (small shops, for

example) and those which do not (the majority of machines located at banks). It is not dependent on any particular computer operating system.

Although the NTDF was devised with transport in mind, other sectors of the economy face very similar data issues and opportunities. Some of these are taken up in "*The Power of Information*" report of June 2007 and the Advisory Panel on Public Sector Information added its weight to its recommendations, calling on the government to make the most of archives of data which at present are not being used. No single data collector or user, government departments included, can reliably predict how data may be used when combined with data from other sources. Realising these benefits, therefore, requires permitting greater access to data in order to permit experimentation in developing innovative data applications. Among the obstacles are regulatory and administrative barriers, poor incentives, differing attitudes to intellectual property, and limited awareness and expertise across government and other organisations.

The development of the NTDF relied heavily on a number of use cases that were defined at an early stage in the project and served to generate specific technical requirements that guided and underpinned the development of the technical capabilities of the NTDF. These use cases covered the areas of (i) road network performance, (ii) car park asset inventories and (iii) rail travel demand. The selection of the use cases was guided by the desire to identify as wide a range of different data types and requirements (including institutional settings) as possible, and to do so in such a way as to address significant real world problems, rather than purely artificial examples. Each use case was developed in close consultation with one or more industrial or public sector collaborators. As well as helping to shape our understanding of the technical requirements associated with the NTDF, the use cases were also of value in providing insight into some of the broader institutional considerations that affect the outlook of data owners and data holders. In particular, concern was expressed by several practitioners regarding their liability in the event of erroneous or malicious use of their data.

Indeed, one of the key issues to emerge from the development of the use cases was data quality (see Appendix IV). As the focus of interest shifts increasingly to supporting the re-use of data in multiple applications and contexts, there is growing concern regarding the scope for the propagation and magnification of data errors. Existing approaches to data quality in the transport domain tend to be tied closely to particular applications of the data and do not always provide the information necessary to support intelligent and creative re-use of the data for other purposes. There is therefore a need for improved methods of characterising data quality and publishing such characterisations in an accessible and useful form. To respond to this need, we have proposed a new concept, that of the “data supply chain”. This is the sequence of processes of abstraction, transformation and computation that take place as one moves from a particular real world concept of interest (e.g., the traffic flow on a road link) to its measured counterpart residing in a particular dataset. From this standpoint, the characterisation of data quality of a particular data source essentially becomes the process of fully describing the relevant data supply chain. With such a description in place, the adequacy or otherwise of the data source can be judged in relation to the requirements of new applications, unconstrained by the notions of data used in existing applications of the data. The feasibility of implementing the data supply chain concept was explored in the context of two of the NTDF uses cases – data from SCOOT inductive loop detectors and data from the National Rail Travel Survey. This work indicated that while existing approaches can deal well with some of the requirements associated with characterising data supply chain, several key elements remain problematic. However, without a

move in the direction of properly documenting the data supply chain, it is difficult to see how the new problems of data quality raised by systems such as the NTDF can be properly addressed.

The investment in the NTDF has exposed and started to tackle these issues. Until now transport data owners have not had an illustration of how to share descriptions of diverse data and find other sets. The NTDF has shown a means to do this; data owners were reluctant to share before the creation of the prototype but the creation of the prototype has shown how to share. Another reason for reluctance can be uncertainty by data owners over the quality of their data and over access control. The next step for the NTDF should be to encourage more transport data owners to register their data sources within it. Encouraging users to register data can be backed up by high-level policy guidelines.

Two key meetings of stakeholders in late 2006 produced proposals for three demonstrator applications. Organisations present included the DfT, Highways Agency, DTI, BAA, EEDA, Lockheed, Deloitte, Thales, Transport Direct, Trafficmaster, BBC, Google, BP, BT, O2, Logica CMG, Norwich Union, and RBS. The first application, My Journey, envisages a system to provide customized real-time data to arriving passengers at an airport on onward journeys and the way through the terminal. The second, My Travel Footprint, is intended to be a tool for organisations to visualise their employees' carbon footprint. The third, My Event, envisages a system to get people to and from an event and provide information about it. Each of these applications is being taken forward; all three require the co-operation and federation of data sources from different owners.

As a follow-on from the NTDF, and as a direct result of the investment in it, a consortium led by Cambridge University[1] is performing a pilot study for the creation of a National Transport Information Incubator. This pilot is being funded by the Department for Transport. This will provide the leadership, and a neutral environment, to encourage owners of data that it is in their interests to join together to experiment and so devise new high-value services. It will offer essential tools, and give advice on technical issues and on appropriate business models. The NTDF would be useful as a tool for identifying data sources to populate NaTII propositions. The exploitation of large quantities of data is in its infancy. The development of new services will stimulate the gathering of new types of data and the improvement in the quality of the data already available. To discover what services are possible and what is their value will require data owners to get together to experiment in a safe environment that offers the opportunity to assess proportionate risks. Opening access to data will allow existing businesses and new SMEs to develop novel services that respond to market needs and opportunities. In order to enable this, there needs to be a means for users to log and locate the vast number of datasets. Services based on exploiting large data collections potentially have enormous world-wide markets, which the UK is in a strong position to exploit, because of its investment in eScience. But only if it continues its forward-thinking approach.

This report recommends the creation of a ‘catalogue of catalogues’ for metadata describing data related to transport, enabling an understanding of what data resources there are and a means swiftly to locate them by searching the system. The NTDF could form a key part of such a system. Existing initiatives, notably GovTalk and the ITS Registry, should be brought into the development of the system. The Department for Transport is ideally placed to lead this

[1] The NaTII consortium consists of Deloitte, Lockheed Martin and Thales Research and Technology (UK), working in collaboration with Transport Direct, the Ordnance Survey and other organisations.

initiative. The data sets in the transport environment are perhaps especially diverse in nature and there is a pressing need to create a way to log and search and find these sets using a standard method. DfT expertise and experience in this area is transferable across government and out into the public arena, so providing leadership on data issues as a benefit to the UK public and economy as a whole.

Introduction

Background

A core government aim has been to facilitate better use of, and access to, data for more efficient government and the public benefit.

The National Transport Data Framework (NTDF) and related transport activities were originally an initiative of the Cambridge-MIT Institute (CMI), which organised a two-day meeting on transport in October 2004. This meeting brought together some 50 experts from a variety of backgrounds, including the Chief Scientific Adviser to the Department for Transport (DfT). Among the conclusions of the meeting were:

- It is essential to provide good connections between different transport systems, and good information for passengers;
- There is a need to facilitate the exchange of data, not only to make information available to passengers in real time, but also to allow organisations access to archived data;
- Better ways of mining the archives need to be developed, for example to deduce information about traffic jams from data on bus movements;
- More data need to be collected automatically and archived;
- Data are essential for optimising timetables, devising disruption recovery strategies and innovative payment mechanisms, designing new transport systems, and planning all aspects of transport policy and operations;
- Work is needed to determine what type of data needs collecting that can be most useful. It is possible that much existing data will not be useful as the 'links' between the various existing data sets are not reliable.

After the meeting, Professor Landshoff spent three months visiting various participants on behalf of CMI and then gathered together a group of them in Cambridge in February 2005. These included representatives of Network Rail, three rail operating companies, and experts on roads. A lively discussion reinforced the conclusions of the October 2004 meeting. Following discussions, DfT agreed to fund a manager, with a contribution from CMI, to help create and manage a portfolio of transport activities to be developed by a number of universities. A further meeting in Cambridge in March 2005 brought together academics from Imperial College London College, Leeds, Newcastle and Southampton universities, who agreed that they wished to collaborate.

A proposal was prepared and submitted to the DfT in response to their call for proposals for the Horizons Research Programme. DfT Horizons is an opportunity for innovative researchers to conduct ground-breaking work identifying and investigating areas where change in the external environment (e.g. technological advance, social trends, environmental or economic challenges) could affect the policies or operation of the Department. The NTDF project was approved by the evaluation panel and funded alongside two other projects that also explored the handling and

analysis of large datasets. CMI was the contractor and subcontracted the work to a collaboration between Cambridge University and Imperial College London College.

In 2006 the Stern and Eddington Reviews were published. As captured in "Towards a Sustainable Transport System", DfT's response to the Eddington Review, these set a challenging new agenda for a sustainable transport system that supports economic growth in a low-carbon world. The Eddington Review emphasised the role of agglomerations where national and local initiatives should coincide to deliver economic benefits. The Transport Innovation Fund (TIF) will be used to achieve such outcomes and benefits in specific agglomerations. The boundaries within and between public and private-sector organisations need to be challenged so that data, information and knowledge can flow more easily for the public and/or commercial good. NTDF would help to break down these boundaries to the benefit of all.

In their review of road data, Transport Direct suggested that there are three information layers: infrastructure; networks and services (which use the infrastructure); and the people that use the networks and services. The Ordnance Survey's Integrated Transport Network (ITN) and the Digital National Framework (DNF) could be seen as the kernel around which additional layers of transport and travel data, information and knowledge could be built. NTDF would facilitate the building of, and access through, these layers.

Consequently, recent agendas, challenges and initiatives have amplified the need for a framework that enables existing information to be found, owners contacted and access gained, so that duplications can be removed, efficiencies gained, better informed decisions made and effective policies and strategies created. There is clearly a need for the NTDF.

Aim

The aim of the NTDF is to provide an effective solution to derive useful information in real time and to analyse archival data from disparate sources. This will help transport users and operators make better decisions and government departments formulate strategy. The data to be handled will be very varied and will range from information from sensors on highways and from CCTV to demographic or pollution data.

The DfT-funded project was to develop the middleware at the core of the framework. An architecture was to be prepared to combine information from a variety of sources, allowing for uncertainties in the provenance of the data. A prototype implementation of the framework software was to be developed to include the following areas:

- An authentication framework identifying the users of the data;
- An authorisation framework, defining access rights to data;
- A core database system for the hub metadata;
- Modules for ingesting new information;
- Metadata handling for processes that create new secondary data, including the tracking of data quality (see Appendix IV) and the propagation of uncertainties;
- A user interface.

Objectives

The specific objectives of the NTDF project were to:

- Develop a prototype implementation of robust and extensible software tools to provide access to large and complex transport datasets. These tools would include metadata, distributed database, statistical processing, user interface and visualisation components.
- Apply this prototype to the specific issue of the consolidation and use of diverse sensor system inputs to the monitoring of network performance.
- Evaluate the performance of the prototype, through consultation with relevant users.

Funding

The NTDF and related transport activities were originally an initiative of the Cambridge-MIT Institute, which contributed £42.5k, later increased to £69.5k. These figures do not include a significant amount of Professor Landshoff's time over the last three years, nor the time of John Patman who, as a consultant employed by CMI, provided links with industry. Peter Landshoff and John Patman were also undertaking other projects for CMI, and their work on these provided valuable links for the transport activities. The Transport Programme Manager was funded to support the NTDF and other transport projects.

The DfT Horizons Programme funding was to create the core middleware. Table 1 summarises the sources of funding for NTDF.

Organisation	Funding
Cambridge-MIT Institute (CMI)	£69,500
DfT – Horizons Programme – NTDF middleware	£275,000
DfT – Transport Programme Manager	£92,676
Total	**£437,176**

Table 1: NTDF Sources of Funding

Parallel Development of NTDF Applications

At the same time, specific applications of NTDF were to be developed. These would be funded either by creating partnerships with industry, or through response to funding calls by EPSRC, the DTI and others:

- The first application associated with the NTDF would be the TIME project, which was just beginning in Cambridge: it would use the city as a test-bed to learn how to gather data and write software to handle the data in real time.
- The next one, initiated by Professor Landshoff on behalf of CMI and led by Professor Polak of Imperial College London College, brought together the five universities that participated in the March 2005 meeting. This was the MESSAGE pollution monitoring project, funded by EPSRC together with the DfT.
- As a result of a discussion between Professor Landshoff and DfT, a consortium of companies was then assembled to respond to a DTI call.

Project Team

The DfT, specifically the Chief Scientific Adviser's Unit, was the NTDF project sponsor. The DfT also funded the transport programme manager, Michael Simmons, with additional funding from CMI, who managed and co-ordinated the project and worked on other transport-related projects. Both contracts were awarded to CMI, with Peter Landshoff as the Principal Investigator. They were later transferred to Cambridge University, following the cessation of most CMI activity.

Professor John Polak, director of the Centre for Transport Studies, was the line manager to the research associate, Dr. Hongqian Liang, at Imperial College London. Mark Hayes, Deputy Director of the Cambridge eScience Centre (CeSC), was the line manager to the research associate, Amine Tafat, based in Cambridge, and to the transport programme manager.

Professor Andy Parker, Director of the Cambridge eScience Centre, Dr. Ken Moody and Dr. Richard Gibbens, Cambridge University Computer Laboratory, were also associated with the project. There was some initial involvement also from Professor John Darlington and Dr. Jeremy Cohen, London Internet Centre, and Professor Marcus Wigan, Napier University and University of Melbourne. Professor Wigan also undertook the evaluation of the NTDF website.

The NTDF Approach

Programme of Work

Work Packages

The work packages set out at the beginning evolved over the lifetime of the project, approved by the Quarterly meetings with the involvement of the key project members and DfT managers.

List of work packages as of September 2007:

•WP1A use case study on road network performance modelling, Imperial College London, May 2006 – December 2007 (see Appendix I),

•WP1B use case study on rail asset conditions (see Appendix I), network performance and passenger demand data, Imperial College London (see Appendix I),

•WP1C use case on car park asset data, Imperial College London (see Appendix I),

•WP2A initial design of architecture and development, mid February – 2nd June 2006.

•WP3 feedback for implementation informed by use cases and WP5 development, customers and WP4 evaluation

•WP4 evaluation, Imperial College London

•WP5 management, development and reporting, responsibilities of programme manager, January 2006 – December 2007 (23 months)

Deliverables

The agreed deliverables evolved slightly over the lifetime of the project, approved by the Quarterly meetings. Deliverables were produced to schedule until the 'Feedback for implementation', due in July 2007, which was produced in November 2007. Two further deliverables, this report and an evaluation, have been produced. See also the section on Findings and Deliverables below. Funds for the project were released in stages on production of deliverables spaced out over the lifetime of the project.

The project deliverables were:

- month 3: Survey of potential users of NTDF underway and subjects of Case Studies determined
- month 6: Report on WP1 –- Initial specification of a general architecture for the NTDF
- month 9: Agreed workplan for software development available
- month 12: Interim Report from WP4 – Evaluation of initial architectural design
- 20th April (month 15): Interim Reports from WP1A and WP1B case studies
- 20th July (month 18): Feedback for implementation key stages completed (this deliverable was cancelled)
- 20th October (month 21): Final evaluation (WP4) started (conflated with last deliverable)
- 20th December (month 24): Final Report covering the software developed (WP2), the case studies (WP1) and project evaluation (WP4)

Communications, reporting and wiki (on CD-ROM)

A domain name was purchased and a website was created at www.ntdf.org.uk. A wiki, now available on CD-ROM, was created to facilitate communications between the project members.

The project was steered by quarterly meetings ('Q-meetings'), the first of which was held in January 2006. These included at least the DfT project manager, the principal investigator, the co-investigator, the programme manager, the two research associates, and the Deputy Director of the Cambridge eScience Centre. Others also attended, such as colleagues from the London Internet Centre and Cambridge Computer Laboratory. These quarterly meeting with DfT representation meant that the DfT were able to provide feedback and be kept informed of developments throughout the project.

Two progress reviews took place, the first after six months, conducted by the head of the Cambridge MIT Institute, and then in January 2007 by the Professor Ian Leslie, Cambridge Pro-Vice Chancellor for Research.

Further meetings by telephone or in person took place as required. At some stages of the project, for example when the early prototype was under development between January and June 2007, these took place very regularly between the programme manager, the Deputy Director of the eScience Centre and the two research associates at least.

Description of developments

The work of the project can be divided into several phases.

December 2005 to April 2006

There was a delay in recruiting the two middleware developers and they did not take up their posts until early February 2006.

In March 2006 Michael Simmons, Ken Moody, Hongqian Liang and Amine Tafat visited Joanna White and Philip Proctor of the Highways Agency, in Bristol, to discuss access and issues around the MIDAS highway loop data. This resulted in the NTDF being granted access to a derived MIDAS source, MIDAS Gold.

In March 2006 the project team visited the National Traffic Control Centre in Birmingham, and made contact with experts such as Ivan Wells of the HA, Steve George (driving force behind www.help2travel.co.uk) and Timothy Hilgenberg (of SERCO).

These two visits gave the team valuable links to experts in the transport field. One significant discovery was the existence of the ITS Registry (www.itsregistry.org.uk), an extensive repository of transport data using the UML metadata mark-up language (so distinct from the NTDF approach to use semantic web technology, specifically RDF and OWL for metadata).

Through these visits the NTDF project made contact with Mary Gosden of the Highways Agency, leading a survey of data sources within the Agency.

April 2006 to December 2006

Work was done during this phase on the development of the middleware and on examination of MIDAS (HA highways loops) data and MIDAS Gold (a snapshot of the raw MIDAS data), on Cambridge SCOOT data thanks to colleagues in the TIME-EACM project in Cambridge, and on mark up languages for describing physical sensors and processes.

A technology survey document was produced as a deliverable as a result of the work done during the period up to June 2006.

The three use-case descriptions in June 2006 (see Appendix I) identified important issues and areas for further investigation which could be worked on in future projects: the whole area of data quality and how to begin to address it in the context of a metadata system (see Appendix IV); the proposal that as part of this the whole data source needs to be described including the physical sensor or source, and the operational aspects of any devices; and the proposal that survey data can be enhanced through an understanding of the actual process of conducting the survey, captured in metadata. The network flow use case informed the development of the SCOOT/weather application in the prototype.

In July 2006 there was a meeting between the National Rail Travel Survey managers and John Polak and Michael Simmons. It was agreed that work could be done on characterising aspects of the survey process with a view to later work on NRTS survey data (then estimated to be due in mid-2007). Following this meeting Hongqian Liang began work on the characterising the NRTS Functional Specification. In the event, the NRTS survey data was not released at the time of this report.

During the period June 2006 to December 2006 the two researchers worked on the development of the middleware and data quality issues.

Two seminal meetings, on 12th September and 8th December 2006, brought together a number of interested organisations in what was termed the 'Data Value Group'. A participant exercise at the 8th December event produced the three projects described in section 8. The two meetings were put together by Peter Landshoff, John Patman and Michael Simmons and took advantage of their contacts in industry and government that had been made by their other work for CMI.

January 2007 to July 2007

Professor Ian Leslie, as reviewer of the project at the 12-month stage on 14th December 2006, recommended the development of a demonstrator prototype. At the review meeting, he raised a concern that there was a requirement to deliver over a tight timescale but there was a lack of clarity of what the prototype was supposed to do. It should provide a user interface for discovery of data, a description of the data and information on data quality and provenance, and a system of access controls.

A demonstrator prototype specification was drawn up and an initial demonstrator system developed from January 2007 to June 2007.

July 2007 to December 2007

In July 2007, following the Q-meeting, Professor Landshoff decided that the prototype was not of an acceptable standard. Mark Hayes was then given the brief of re-developing it. By mid-October 2007 he had rebuilt the system and incorporated two demonstrator applications: Cambridge SCOOT and weather correlations. The application showing Cambridgeshire road casualties with GPS position on a map through time was developed by Luke Hur of CeSC. Transport Direct developed an application showing CO2 footprints for individuals from travel-to-work data. This application linked in to the My Travel Footprint case which may be put through the NaTII pilot project, which runs between September 2007 and April 2008.

The rebuilt system also was populated with metadata for transport and related data sources from the Cambridge Travel for Work partnership and other travel-to-work data via O2 on the Leeds Arlington site; Cambridge SCOOT data; Cambridge weather data; and diverse data from sources recommended by Transport Direct. When asked for feedback, selected users were also asked to add their own metadata description for datasets relevant to them. Imperial College London College felt unable to provide any sources of data.

A meeting in July 2007 in Leeds between the HA RIF group and the NTDF project, with Geoff Scott of BT Research, was useful but it did not lead to an application.

The NTDF team met Landmark, the collectors of some UK car park data, via Transport Direct to discuss possible incorporation of their XML into the NTDF system, to tie in with the car park asset use case but no conclusion was reached. The Update on Use Cases (Appendix II) deliverable in May 2007 gives details of work on data characterisation including how the car park case could be characterised.

The NRTS data release was delayed until after the end of 2007, so could not be incorporated into the system.

Professors Polak and Landshoff met in August 2007 to discuss the way forward and both agreed on the need to bring the project to a satisfactory result. With the agreement of the DfT, Q-meetings after July 2007 were stopped in favour of direct communication with the DfT Chief Scientific Adviser's Unit. The two researchers were asked to produce papers for dissemination.

Adoption of Semantic Web technologies

The Semantic Web is an extension of the usual ubiquitous web technologies which allows the addition of *meaning* to web content. For example, an individual's WWW home page will not only contain textual information on that person intended to be read by humans, but also machine readable information from a set of pre-defined concepts common to all persons. Such a set of concepts is known as an ontology and also serves to *define* a person as a higher level concept.

For example, the Friend-of-a-Friend (FOAF) ontology [foaf] contains the full name, geographical location and contact details for an individual. As the name implies, it also allows for links to other people known to an individual. It is often used for defining social networks on the web.

[foaf] http://www.foaf-project.org/

Links may also be made between ontologies to identify concepts that have the same meaning. In this way, the Semantic Web offers a means of integrating previously independent "silos" of data. As more data becomes widely available on the web, the links forged between different data sets will benefit from an exponential network effect. It will become easier to locate data sources of interest, to manipulate and visualise them and to compare them with related data sets. The NTDF aims to kick-start this network effect for transport-related data sets.

Ontology

Ontology is a term borrowed from philosophy that refers to the science of describing the kinds of entities in the world and how they are related. New languages will enable the development of a new generation of technologies and toolkits. Semantic Web languages are based on a so called Description Logic (DL), a family of knowledge representation languages, which can be used to represent the terminological knowledge of an application domain in a structured and formally well-understood way.

Ontologies: Ordnance Survey examples

Ordnance Survey is building several topographical ontologies, including one for buildings and places. Table 2 illustrates some transport-related entities, highlighting possible synonyms and the way in which the entity is defined.

Concept	Synonym	Definition	Pseudo-Language
Airport		An airfield containing a terminal building and used for transporting passengers and cargo [10]	An Airport is a kind of Place. An Airport has part a Terminal Building. An Airport has purpose Transport by Aircraft.
Bus Station	Bus Terminal	A place where public transport buses begin, break or end their journey and at which passengers may embark or disembark. [1]	A Bus Station is a kind of Place. A Bus Station has purpose Public Transport by Bus. A Bus Station has part a Building that has purpose Public Transport by Bus.

Concept	Synonym	Definition	Pseudo-Language
Car Park		An open area of ground or a building where vehicles, [1] mainly cars, [10] may be parked. [1]	A Car Park is a kind of Place. A Car Park has purpose Parking of Cars.
Railway Station	Station	A building and platforms by a light railway network or railway network where a light rail vehicle or train may stop to pick up [1] or transport [10] goods (Freight) [10] or passengers. [1]	A Railway Station is a kind of Place. A Railway Station has purpose Transport by Railway. A Railway Station has part a Building that has purpose Transport by Railway.
Service Station		An establishment beside a road selling petrol and [2] providing [10] refreshments to motorists; [2] Has a petrol station. [10]	A Service Station is a kind of Place. A Service Station has purpose Provision of Refreshments for Motorists. A Service Station has part a Building that has purpose Provision of Refreshments for Motorists. A Service Station has part a Petrol Station.
Underground Station	Tube station, Metro station	A railway station on an underground railway.	An Underground Station is a Railway Station. An Underground Station has purpose Transport by Underground Railway.

Table 2: Ordnance Survey Places Ontology - Transport-Related Entities

Ontologies: NTDF examples for weather and SCOOT

The NTDF uses a subset of the Dublin Core ontology (see http://dublincore.org/) for basic 'library card' information on a data set, such as the publisher, the physical format of the data, its geographical/temporal coverage and whether the data is public domain or access restrictions apply. The complete set of Dublin Core identifiers used in the NTDF is detailed below.

Term	Definition
identifier	an unambiguous reference to the data resource
Subject	a set of keywords or keyphrases to describe the resource
Title	a free text name given to the resource.
publisher	the entity responsible for making the resource available
description	A short free text description of the resource
contributor	the entity responsible for contributing metadata on the resource
access rights	a public/private tag indicating access rights (see also section on Access Control
coverage	the spatial or temporal coverage of the resource
Format	the file format or physical format of the resource
Creator	the entity responsible for creating the resource
Type	the nature or genre of the resource, e.g. numerical/text/graphical
Source	a primary source from which the resource is derived
Relation	a related resource
Date	the date the resource was described
language	the language of the resource and its documentation

There is a also limited ontology built into the system which makes use of the keywords field to enhance search results. For example, "highway", "motorway" and "street" are all subclasses of the concept "road". "NO2" and "noise" are both instances of "pollution".

We have also developed ontologies describing the messages generated by the SCOOT induction loop system and the archived data for the Cambridge weather service at http://www.cl.cam.ac.uk/research/dtg/weather/

These are available in RDFS format at

http://www.ntdf.org.uk/scoot.rdfs and
http://www.ntdf.org.uk/weather.rdfs

Generating ontologies for specific data sets is a time consuming, manual process. However, there are tools (such as Protege: http://protege.stanford.edu/) which allow the user to focus on the concepts they are describing rather than the detailed XML syntax.

RDF, OWL and SPARQL

The basic components of the Semantic Web (Resource Description Framework, or RDF & Web Ontology Language, or OWL) have been adopted by the World Wide Web Consortium or W3C and the wider web development community as relatively mature standards. SPARQL Protocol and RDF Query Language, or SPARQL, a query language for the Semantic Web, is currently going through the W3C standardization process.

The Web has revolutionized access to information for individuals, the private and public sectors alike. As an extension of the protocols behind the Web, the Semantic Web promises to do the same for data. The Semantic Web would therefore seem to be promising approach to building a system such as the NTDF, whose purpose is to enable easier access to data and integration between distributed data sources. By converting descriptions of the data sets in the semantic web language RDF, the NTDF system then produces metadata in a standard form regardless of the original format, which is searchable by the powerful search language SPARQL.

XML and UML

XML is often used as a format for publishing structured data on the web. Examples in the transport sector include NaPTAN and MIDAS Gold. There are many development tools and programming libraries available for processing XML. XML itself is extremely flexible: one may invent XML tags for almost any conceivable application. RDF and OWL are usually expressed in XML, although using a highly constrained set of standard tags for describing data and the relationships between data sets. XML alone is almost too flexible: no two parties would necessarily define the same set of tags (or schema) for a particular application.

UML (the Unified Modelling Language) was originally intended as a way of documenting object-oriented software systems but may also be used to describe physical systems (e.g. an aircraft), business processes and organizational structures. UML is often expressed in a graphical form for human consumption but may also be serialized as XML. The ITS Registry (http://www.itsregistry.org.uk/) is a source of UML models for the transport sector. In its current form it does not easily lend itself to inclusion in the NTDF repository. However, if XML versions of the UML models were made available at well-defined URIs, either on the NTDF or the ITS Registry web sites, it would be possible to provide links to these from metadata records

within the NTDF.

Findings and Deliverables

NTDF Prototype

NTDF Architecture

Figure 1 illustrates the NTDF architecture.

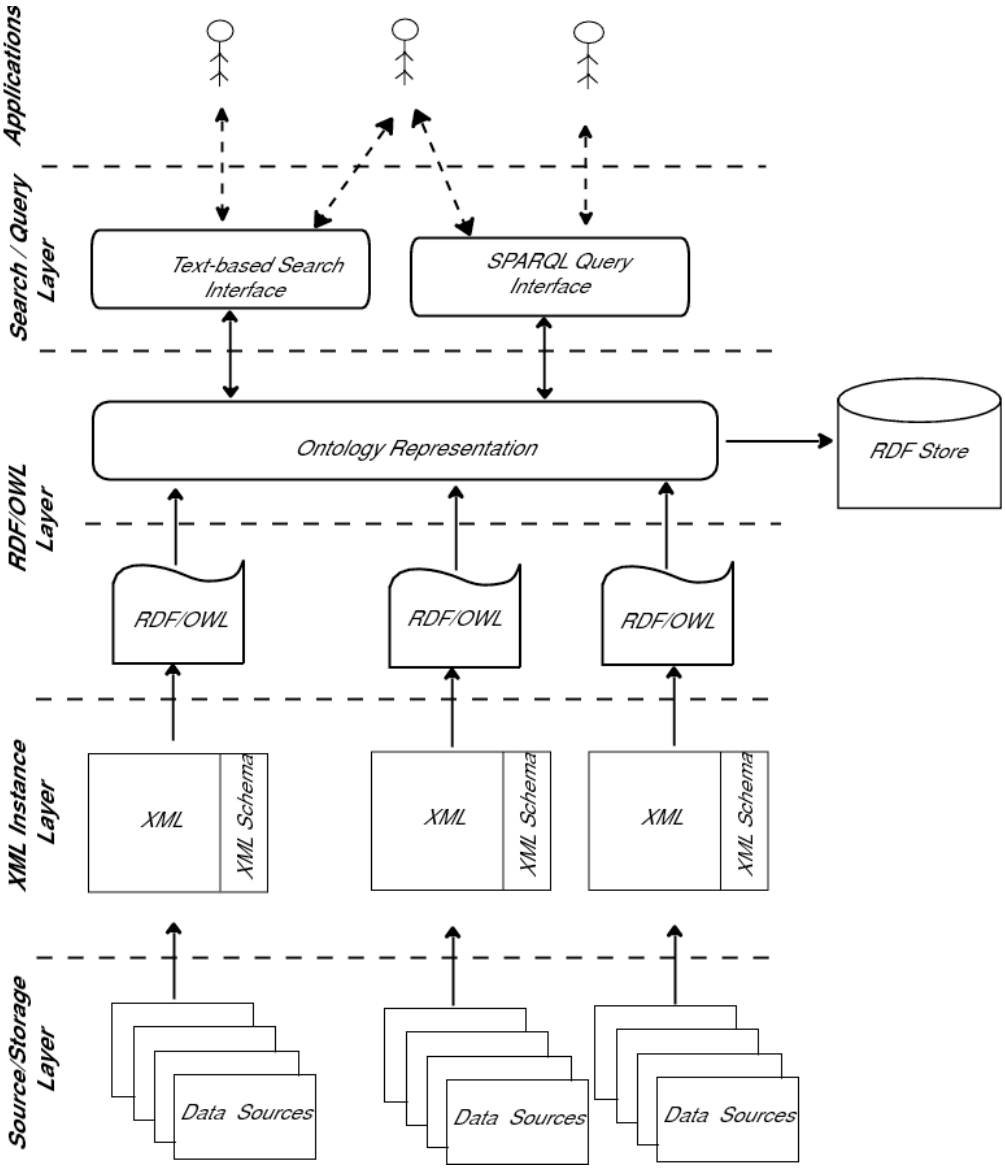


Figure 1: NTDF Prototype – Architecture

The NTDF Architecture comprises:

- User interface / application;
- Search / query layer;
- NTDF ontology;
- RDF / OWL layer;
- XML instance layer;
- Data source / storage layer.

} ***NTDF Middleware***

Data Source / Storage Layer

End users will interact with the NTDF middleware through the query and search layer. Simple keyword searches and an advanced SPARQL interface provide access to the repository, allowing the identification of relevant data sources for a particular application.

NTDF Middleware

The NTDF "middleware" enables transport data owners and transport service providers to discover new sources of data and use them to build their own applications. This middleware functionality is organised around three levels:

- ***Repository***. The repository is intended to store transport metadata and the semantic representation of different transport data sets. The metadata and semantic representations, which may be from descriptions provided by owners of their datasets, are converted and stored as RDF triples.
- ***Semantic and metadata description***. The NTDF middleware supports both user-defined metadata and a Dublin Core description of each data set. Dublin Core includes tags identifying the publisher of the data, it's spatial and temporal scope, physical format and any associated access rights. See http://dublincore.org/. Data holders describe their data by filling in a series of web forms. This information is automatically converted to RDF and included in the repository.
- ***Search and Query***. Retrieving information from the NTDF repository can be done in two different ways. A simple keyword search tool is provided to access the repository content and discover information about transport data sets. In addition it is possible for expert users to query the NTDF repository via a SPARQL interface. SPARQL is an SQL-like formal query language for RDF. As of November 2007, it is a Proposed Recommendation in the W3C standards process (i.e. one step away from being adopted as a fully-recommended standard.)

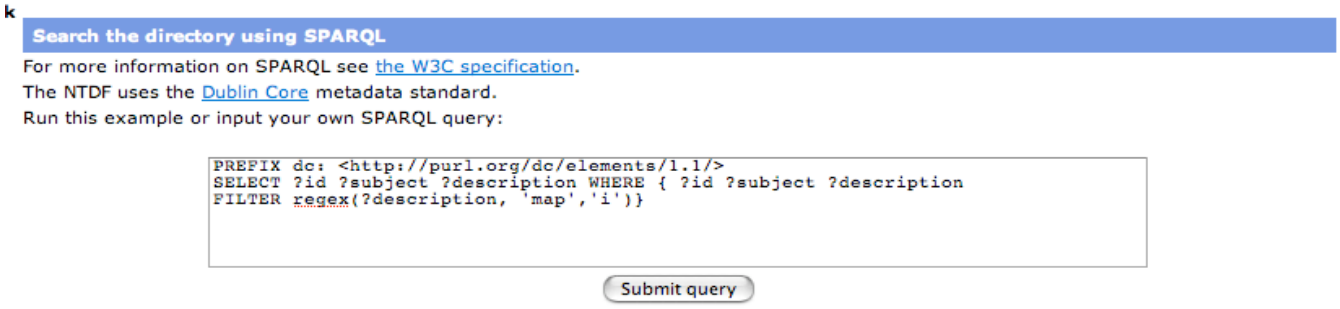


Fig 2: Screenshot of SPARQL query from NTDF prototype

[subject] -> [predicate] -> object
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/date] -> 9/8/2007
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/title] -> AMEE carbon footprint database
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/format] -> XML
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/rights] -> public
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/subject] -> carbon footprint CO2
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/coverage] -> UK
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/language] -> English
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/publisher] -> http://www.dgen.net/
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/contributor] -> mah1002@cam.ac.uk
[http://blog.co2.dgen.net/] -> [http://purl.org/dc/elements/1.1/description] -> An open database of carbon footprint informatior
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/date] -> 9/8/2007
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/title] -> NAPTAN - National Public Transport Node database
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/format] -> XML
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/rights] -> public
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/subject] -> public transport bus trains stations stops NAPTAN
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/coverage] -> UK
[http://www.naptan.org.uk/] -> [http://purl.org/dc/elements/1.1/language] -> English

Fig 3: Screenshot showing RDF triples for NTDF metadata

NTDF Applications

An example application was developed in order to illustrate how the NTDF can be used to demonstrate an aspect of network flow (see the first use case in Appendix I). In this scenario, we provide a common view across an archive of Cambridge weather conditions and traffic data from the SCOOT loop system in Cambridge. The system automatically generates a CSV file containing daily cumulative rainfall and SCOOT vehicle counts, between any two dates. We provide an example Microsoft Excel spreadsheet file with macros for generating line graphs and scatter graphs, showing any correlation between traffic levels and rainfall levels.

A second application uses data on road casualties in Cambridge to plot accidents through time and across the city in Google Earth. This visualisation helps to discern patterns of road accidents, their location and frequency.

A third application, showing CO2 footprint for employees travelling to work (using data in the system) is planned.

NTDF User Interface

We provide a web interface to the NTDF middleware so that all functionalities are accessible via a web browser. NTDF users can use the web portal to add descriptions of their data to the repository, list the contents of the repository, search the repository by keyword or SPARQL query.

We describe in the next paragraphs each of these tools and interfaces. Figure 4 shows the main NTDF web interface from which users can invoke all NTDF functionalities.

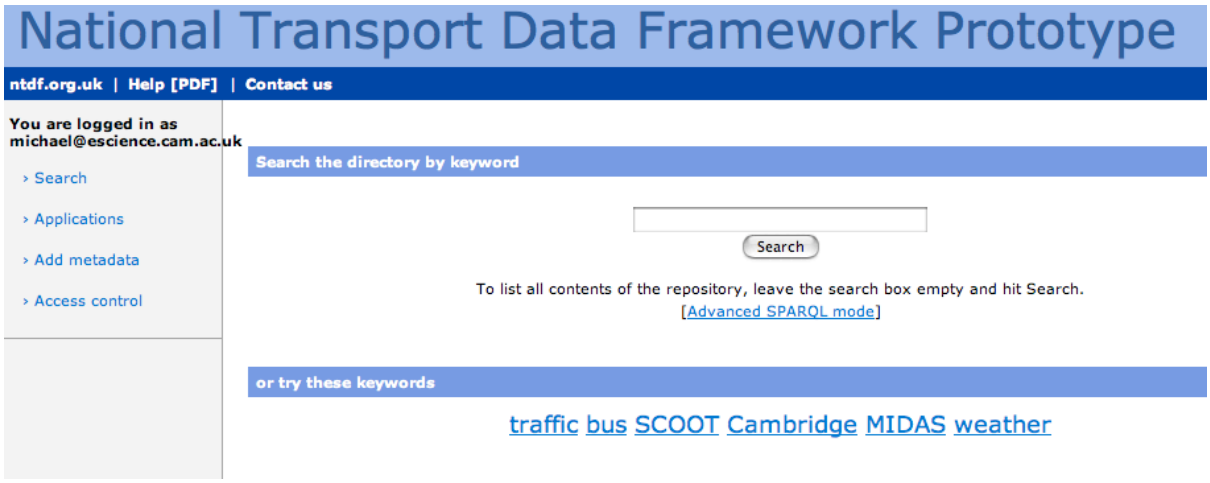


Fig 4: screenshot of main NTDF user interface

NTDF Prototype Implementation

The metadata repository was implemented as an RDF store using the open source Redland libraries (see http://librdf.org/). The underlying persistent storage is provided by the Sleepycat/Berkeley database.

Our reason for choosing Redland is motivated by the fact that it is a relatively well established product with comprehensive documentation and APIs available for many different programming languages. As the NTDF software is accessible as a web service, we adopted the Perl API as this is one of the most popular languages for web scripting.

NTDF Access Control (see also Appendix III)

The original project proposal highlighted the importance of an authentication and authorization framework for the NTDF prototype. The UK eScience Security Taskforce defined "authentication" as the "establishment and propagation of a user's identity". "Authorization" is defined as "controlling access to services based on policy". In other words, you are authenticated when I know who you are and you are authorized when I know who you are and therefore what you have access to.

NTDF users are identified by a simple user-id (which happens to be the same as their email address) and a password. Access to the prototype requires registration and a simple web based login using the basic HTTP authentication protocol[http-basic-auth]. The Apache web server at http://www.ntdf.org.uk/ (currently hosted by the Cambridge eScience Centre) provides this authentication service. In a production version of the NTDF, SSL would be layered over this to provide end-to-end encryption of the login details.

The prototype allows the definition of authorization policies in XACML format.[XACML] Policy descriptions are stored on the NTDF server and given a unique URL. The XACML processing model allows remote data providers to reference these policy files as part of their own access control procedure. Use of XACML therefore lends itself to the NTDF principle that data owners should retain control of their own assets. The following links give more detail (but see also Appendix III).

[http-basic-auth]: http://www.ietf.org/rfc/rfc2617.txt
[XACML] http://www.oasis-open.org/committees/xacml/

Use Cases

Road Network Performance Use Case

This Use Case informed the core development of the NTDF prototype. The Use Case highlighted the need to provide a means for users to get information on how the road network was operating, which would be in the form of an online system drawing on several real time and non-realtime sources. Such systems exist already, but the aim of the NTDF was to encourage data owners to allow their data to be brought into combination with others' in novel ways.

In the first months of the project the project team made contact with the Highways Agency and gained access to MIDAS data from the loops in motorways and other roads and work was done on characterising the data using a refined version of the raw data called MIDAS Gold. The need identified from discussions with local authorities, the Highways Agency and following a visit to the National Traffic Control Centre in Birmingham was for a system which would take data from SCOOT loops, sensors in local authority areas, and from MIDAS loops, and provide a means of seeing movement of vehicles from a MIDAS area into a SCOOT area.

The NTCC will route vehicles onto a non-HA road if there is a road closure and no system existed to automatically alert a local authority that this was happening, but the NTCC would typically phone the local authority to warn them.

Work on understanding the MIDAS data involved the project in looking at the quality of data coming from the loops, and work on characterising the data according to characteristics of the devices themselves and their operation, and which has resulted in a paper setting out key issues and pointing the way to future work (see Appendix IV).

Attempts to find a location where a local authority SCOOT system and a MIDAS system were next to each other were not successful but the project gained access to the Cambridgeshire SCOOT data with help from colleagues in the Computer Laboratory at Cambridge University and it was decided that a simple but elegant exemplar of combining novel data sources to produce valuable results would be to take weather data for Cambridge and combine it with data from the SCOOT system in the city. The results can be seen in the prototype.

It was planned to incorporate into the prototype a demonstrator application using GPS data negotiated with a Cambridgeshire bus company. A simple combination of weather data with a GPS map of bus movements could have produced non-inutitive results on how bus movements could be affected by rainfall, for example. Work on incorporating GPS data into the prototype was not completed and the re-design of the system from the summer of 2007 meant it had to be dropped.

Two other demonstrator applications were envisaged, one to demonstrate patterns of road accidents on a map of the City of Cambridge, and a further one to make use of travel to work data from the Cambridgeshire Travel to Work Partnership to calculate CO2 footprints for individual employees. This has been trialled by Transport Direct and may go forward as part of the My Travel Footprint initiative, but is not in the prototype system.

Rail Use Case

The original rail use case was defined very broadly to include rail asset, performance and demand data. However, following a series of discussions at the start of the project with representatives from the rail industry and relevant government agencies, it was decided that the rail use case should focus more specifically on exploring issues providing access to the National Rail Travel Survey (NRTS). This reflected both the need to define a more manageable scope of work and

concerns regarding commercial sensitivities within the rail industry regarding rail asset and performance data.

The NRTS is a new survey of rail passenger travel in Great Britain, designed to provide information about how passengers use the national rail network. It is the first such survey to provide detailed information on the travel behaviour of rail passengers. The principal applications envisaged for the NRTS data are the analysis and modelling of travel demand, both for public policy purposes and as an input to the rail franchise bidding process. The NRTS is thus relevant to a wide range of industry, profession and academic users, This naturally lead to the definition of the goal of the rail use case as being to provide the means to facilitate the maximum use of the NRTS data both by parties within the rail industry and those outside. This entailed developing a metadata characterisation of the NRTS data and making this metadata, together with the NRTS data themselves available via the NTDF web site.

Although the development of a metadata description for the NRTS data themselves was reasonably straightforward, aided by considerable cooperation from the DfT and their Consultants, it quickly became apparent that a comprehensive description of the data required much more than a conventional data dictionary approach. The NRTS, in common with many large and complex sample surveys involves a series of transformations of information, starting with the real-world state of interest (in this case, all weekday passenger rail trip in GB) and ending with a specific dataset. The comprehensive characterisation these data requires the characterisation of these processes. This realisation was closely linked to the growing importance within the project of data quality, and is discussed in more detail in the next section.

In the event, towards the end of the project, we were informed by the DfT that delays in the processing of the NRTS data meant that it would not be possible for the Department to make the available to the NTDF.

Car Park Assets Use Case

This Use Case was proposed following discussions with Transport Direct on the issue of collection of data on car parks in the UK, where Transport Direct had commissioned the creation of a baseline national car park inventory for major cities. These data would need to be kept up to date, however, as car park operators make changes to the location, scale, pricing and operating regimes of their assets. Resources are not available to repeat the survey so the updating would depend on car park operators voluntarily providing information to Transport Direct. It could be possible, however, to link operators' databases to a central system in such a way that changes to some key indicators could automatically be propagated into the central database.

This is a valuable but complex project for which there was not enough time for development. Discussions were begun with the company collecting the baseline data but a data owner was not identified as the project concentrated on completing the system.

Data Quality (see also Appendix IV)

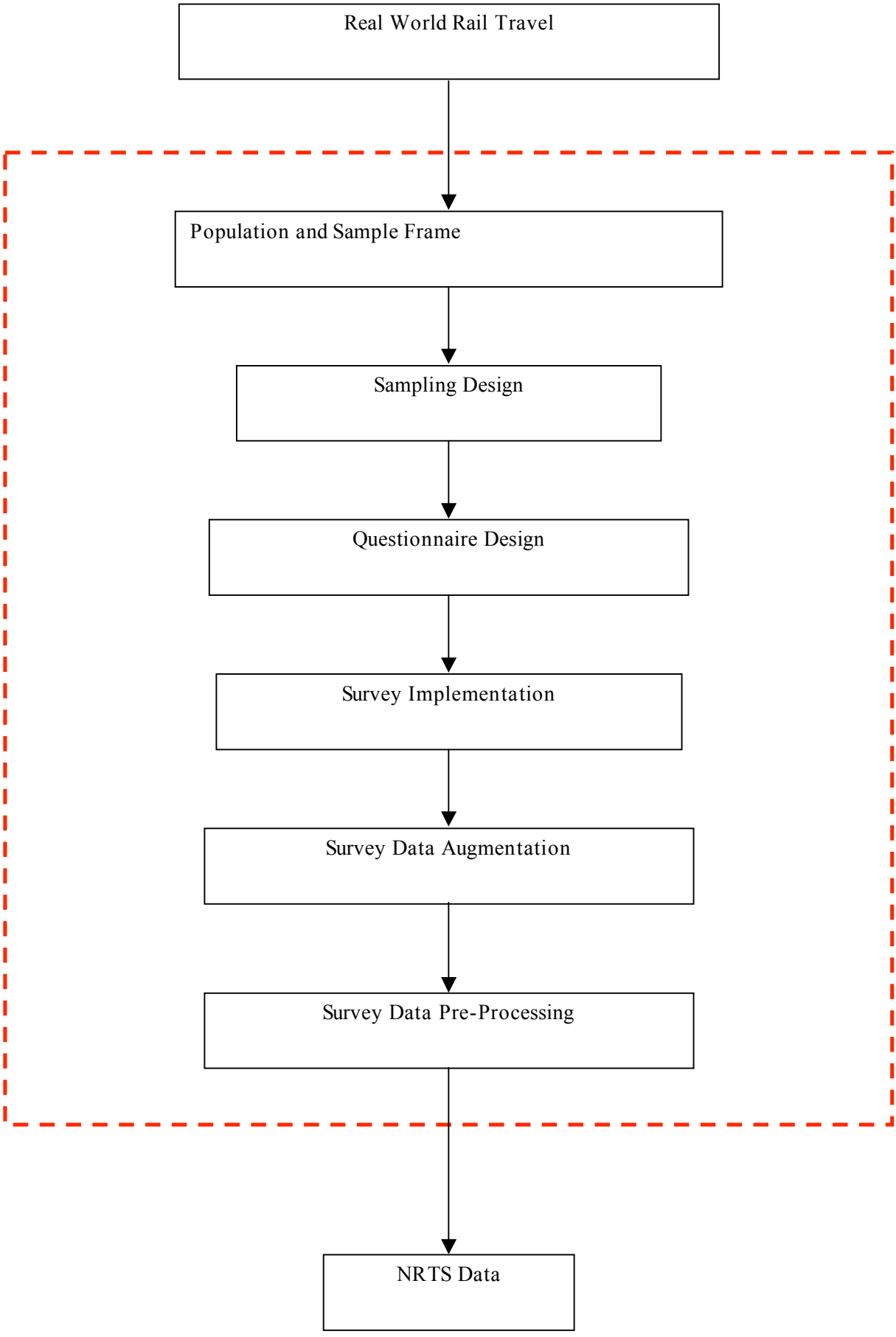
Although not explicitly emphasised in the original work programme, an issue that emerged very powerfully in our initial discussions with all data owners and data users was data quality. There was a widespread perception that as data become more plentiful (e.g., through new forms of data collection system), and as the value of sharing and exchanging data becomes more widely appreciated (e.g., through the growth of web 2.0 technologies) the need will grow for new and more flexible ways of characterising the quality of data. In particular, treatments of data quality will need to acknowledge that a given item of data may be put to quite different uses than those that originally motivated its collection. Although such re-use may stimulate beneficial innovation, there was considerable concern that if improved treatments of data quality are not developed, then extensive re-use of data might also propagate and amplify any weaknesses in the quality of the underlying data sources.

Thus it was decided to create a work stream focused explicitly on data quality, with the aim of exploring the adequacy of existing concepts of data quality and identifying suitable concepts for development in the future. This work stream involved a comprehensive literature review of the treatment of data quality in transport and in related domains such as information systems engineering, geographic information systems, economic and finance and health care, followed up by more detailed consideration of the characterization of data quality in two specific and quite different transport contexts – inductive loop detector data and the National Rail Travel Survey.

This review found that (notwithstanding predictable differences in terminology) there was a substantial level of agreement across the different fields in terms of the importance attached to concepts of accuracy, completeness, meaningfulness and lack of ambiguity. However, important differences exist between approaches adopted in different fields in terms of the extent to which they view data quality as an inherent property of the data (usually judged relative to some notion of ground truth) or as a characteristic of how the data are generated and used. In the context of the objectives of the NTDF, we argue strongly that the latter view is the most appropriate. That is, we believe that the most salient features of quality are those that allow analysts to take informed decisions regarding how data might be most appropriately used.

This argument leads to our proposal regarding the data supply chain. We conceive of the data supply chain as the sequence of processes of abstraction, transformation and computation that take place as one moves from a particular real world concept of interest (e.g., the traffic flow on a road link or the number of passenger rail trips along a particular corridor) to its measured counterpart residing in a particular dataset. From this standpoint, the characterisation of data quality of a particular data source essentially becomes the process of fully describing the relevant data supply chain. The concept is based on what we believe to be two important truths. First that the fitness of data for a particular application can only be judged in relation to that application – that is, data quality can only be judged in the context of a proposed use and secondly, that both now and increasingly in the future, data will be used for purposes other than those originally conceived. Thus the critical issue increasingly becomes how to support the intelligent analyst seeking to establish the relevance and fitness for use of a particular datasets relative to a specific analysis objective. What such an analyst requires is as complete a description as possible of the data in hand and how it was produced.

The feasibility of fully documenting the data supply chain was explored in the context traffic data from inductive loop detectors and data on patterns of rail passenger demand from the National Rail Travel Survey. It was found that the supply chains for these two types of data are quite different and pose different challenges. Moreover, while existing technology provides good means of charactering some stage of the supply chain others can be more problematic. For example, the figure on the next page shows in outline the data supply chain for the NRTS.


Real World Rail Travel
Population and Sample Frame
Sampling Design
Questionnaire Design
Survey Implementation
Survey Data Augmentation
Survey Data Pre-Processing
NRTS Data

Existing tools can deal well with the stages associated with population and sample frame development, sampling design and questionnaire design but are seriously under-developed in the areas of survey implementation and survey data pre-processing.

Overall, the concept of the data supply chain places emphasis on data collectors, data owners and/or metadata creators to encapsulate and publish aspects of their practice that have not commonly in the past been the subject of active dissemination. As such, it may encounter some institutional resistance, especially from those wishing to maintain their privileged access to this information. However, without a move in the direction of properly documenting the data supply chain, it is difficult to see how the new problems of data quality raised by systems such as the NTDF can be properly addressed.

Access to and Protection of Data Sources

Most data owners are cautious about who they release their data to, and for what purpose. There may be hidden or overt commercial value in a dataset, which a company wants to exploit. There may be sensitive information, for example concerning individuals, which an organisation may not anyway legally release. Those requesting use of the data may not be a trusted entity, and reassurances on security and access control (see Appendix III) may not satisfy the data owner. Legal obstacles in the form of agreements may exist relating to particular datasets which are problematic to change.

The NTDF project did achieve access to data sources in certain cases where such access was the academic use; in other cases where the data was anyway publicly available or being collected by a local authority.

Demonstrating the value of a metadata repository and persuading data owners simply to submit a description of their data to it proved challenging: doing so was not obviously of value to a data owner. The push for such submissions will need to be provided by high level policy in the Department or government if a ‘library of libraries’ of metadata is to be achieved.

Work on providing strong access control (see Appendix III) and security in the NTDF prototype did not advance as far as planned, and the development of this will be an important aspect of any such system. Even if the data were to be made freely available via such a system as the NTDF, data owners will need a means to ensure that what they submit carries the conditions of sharing, even if these are complete freedom for any user.

In the NTDF it was hoped that by demonstrating the value of submitting data for metadata description, data owners would become involved. But there needs to be a very powerful impetus to get data owners to do this: commercial value is one (and to some extent the National Transport Information Incubator will be able to demomstrate this), public interest or government interest are others.

Spokes from the NTDF Hub

NaTII

NaTII will be the major application to come out of the NTDF. It is a process for organisations to experiment with data combination (or mashing) to test its value, either public benefit or commercial. Partners in the DfT-funded NaTII pilot (from September 2007 to April 2008) are

the University of Cambridge, which holds the contract with Peter Landshoff as lead, Deloitte Touche, Lockheed Martin, and Thales Research and Technology.

In September 2006, Peter Landshoff, Michael Simmons and John Patman of CMI, working with project colleagues, called a meeting of carefully-targeted organisations to discuss the question of how to get more value from data. Attendees at this meeting are listed in Appendix V. On 8th December, 2006, a key meeting which included many of the September attendees and some others was held at the Royal Society. Attendees at this meeting are also listed in Appendix V. These lists are of organisations with an interest in data value which can be contacted in future.

Three proposals for applications emerged. All three are potential applications arising out of the NTDF concept and are being taken forward:

- ***My Journey***. This is the key case for the NaTII pilot. With BAA Stansted as the 'client', the aim is to produce a demonstrator system which will enable arriving passengers to navigate through the terminal and on to their onward destinations. A subsidiary client is National Express.
- ***My Travel Footprint***. Some work for this will be shown in the NaTII pilot, and a 'mini-application' using travel-to-work data created for the NTDF working with Transport Direct. The aim of My Travel Footprint is to show how travel-to-work data can be used to produce valuable CO2 footprint data for individual employees and to give organisations a detailed view of this and other travel activity.
- ***My Event***. Spurs football club expressed interest in the idea of providing a system to enable spectators to get to and from their stadium more expeditiously. Discussions had begun with the DCMS to use a Spurs project as a pilot for a more ambitious one for the Olympics, when Spurs withdrew because of internal problems. Discussions have begun to replace them with the O2 Dome.

Further projects are being developed to put through NaTII once it is created.

MESSAGE

The Mobile Environmental Sensors Across a Grid Environment (MESSAGE) project was initiated by Peter Landshoff, then of CMI, in 2005. This DfT/EPSRC-funded project is for three years, starting in autumn 2006, to develop the means to get better data on pollution across three urban environments: London, Cambridge and Newcastle (and using facilities in Leicester). The project's Principal Investigator is Professor John Polak, and the partners are Imperial College London, and the universities of Cambridge, Newcastle, Southampton and Leeds.

MESSAGE is part of the NTDF concept in that several institutions and companies have formed a consortium to get more value from data and MESSAGE came about as a result of discussions with universities about the NTDF. Company supporters include Logica CMG, IBM, O2, Thales, Symbian and Nokia.

Conclusions

- The demand from users to find data is increasing at a great rate with a concomitant increase in the collection of data. In order to handle this, repositories describing data sets need to be created. Existing repositories should be linked under one overarching system.
- We have shown in this project what such a metadata system could be like

- The development of such a system is more difficult than we had envisaged but the lessons learnt should help guide future research and developers.
- There are institutional and cultural barriers to getting data owners to allow their data to be described and made accessible, or potentially accessible
- These institutional and cultural barriers are such that the approach of the NTDF, in trying to persuade owners to submit their data by showing them the value of the system, will benefit from high-level direction in government
- The value of participation was not always clear to data owners, and efforts should continue to market the value; but ultimately, in government, a directional approach is called for.
- The semantic web has value in clearly-defined fields, but is less appropriate for very disparate data from a variety of fields, as the data must be manually encoded for the most part. The cost-benefit ratio of the semantic web applied when there is a very large amount of data in a relatively limited field.

NTDF Potential

The amount of data being collected is rising at a dramatic rate and with it the need to be able to catalogue and find the data. In government, like in most large complex organisations, data is located in silos and is not generally centrally catalogued and therefore searchable. For government, it is highly desirable to have one central system to which owners must submit key data details for cataloguing as metadata. The NTDF prototype could form the basis for such a system.

An example of how an overarching metadata repository would benefit users in transport is the lack of knowledge among most users of the existence of individual road leg journey predictions for the M11. Tracking down such a dataset currently would be time-consuming; the existence of a system where such datasets were logged as metadata and fully searchable would be a major benefit.

NTDF Prototype Development

The main conclusions are:

- The design of NTDF, using semantic web technology to describe the metadata (which allows descriptions of diverse data sets to be stored and later searched for and accessed) will improve access to a variety of transport and other relevant data; originally it was thought that the semantic web would be applicable more widely to actual original data sets but the work to characterise datasets in semantic web languages is very time-consuming
- The NTDF prototype has achieved the basis for deriving useful information in real time and to analyse archival data from disparate sources, but more time is needed to develop it further;
- The NTDF prototype provided the basis for a Web service that provides common access to relevant data sources for a variety of users with an interest in transport;
- The NTDF prototype system cannot be described as ‘programmable’ but there is ample scope for users to develop applications on the open architecture provided;

- A key feature of the NTDF prototype is the distributed responsibility for data, with the system accessing data held elsewhere; data owners need to have strong motivators to participate, however
- The NTDF is a web-based resource which is currently available to users who register.

NTDF Prototype Applications

The main conclusions are:

- The National Transport Information Incubator (NaTII) is the major spoke to develop from the NTDF catalyst. When fully established, NaTII will then provide a process for users to test out scenarios for data federation and services;
- Specialist tools will be developed, or may be developed by others in the NTDF either as part of its functions as a back office system, possibly through Transport Direct, or as part of the process in NaTII

NTDF Prototype Evaluation

The NTDF was evaluated in a project deliverable, a preliminary evaluation of the NTDF concept and methods, in June 2007 by Professor John Polak, who found that the principal conclusion to emerge from the evaluation exercise was that the basic architectural approach proposed seems to be regarded by a wide group of technical professionals as appropriate. The principal insight gained was to confirm the importance of adopting a flexible and extensible approach to the development of metadata.

In October and November 2007, a number of potential users were invited to give us feedback on the prototype functionality and usability. The users fell into two categories, expert and non-expert. All were chosen because they had the potential to use the prototype as a way of cataloguing and retrieving data sources. Those invited to give feedback included: Transport Direct, Highways Agency, ITS Registry, Kizoom, Thales Research, Lockheed Martin, Deloitte Touche, East of England Development Agency, Cambridgeshire County Council, Cambridge-MIT Institute.

The website was evaluated by Professor Marcus Wigan near the end of the project. The system is a prototype, and many of these recommendations should be implemented in the next stage of development.

The key recommendations were as follows. The NTDF project responses follow each recommendation.

1. It should be made clear on the NTDF home page what the system does, why a user should register and what to expect when you log in.

NTDF project response: implemented

2. There is an overspill of user-id in the side bar. This is probably due to the user's browser.

NTDF project response: changed

3. 'applications' should be changed to 'use cases'

NTDF project response: we do not agree, these are exemplar applications and distinct from the earlier use case system documentation.

4. The access control functionality should be moved to the general metadata page

NTDF project response: we have looked closely at this and it is clearer to keep these separate

5. 'add metadata' should be changed to something more descriptive

NTDF project response: changed

6. A controlled vocabulary of keywords from which the user can select when adding metadata should be created

NTDF project response: we recommend that this be implemented in the development of the NTDF beyond prototype stage

7. There should also be a way of adding to this vocabulary

NTDF project response: we recommend this for future development of the NTDF

8. There should be a third search interface using drop down menus for the vocab

NTDF project response: we recommend this for future development

9. A description of the SCOOT system should be added to that application/use case

NTDF project response: done

9. A link to a kmz file for the road safety application/use case should be added so that Google Earth opens automatically

NTDF project response: a link to the help file for Google Earth has been added which informs a user how to add Google Earth (if it is not already on their system). Installing Google Earth will vary according to the system and browser being used.

10. In general, applications/use cases need to be better documented

NTDF project response: feedback on this was generally very positive, we believe the documentation is sufficient for the prototype

11. Whether metadata is open or closed access should be better flagged

NTDF project response: implemented

12. The content of the 'help' file could be moved to the web site

NTDF project response: feedback from users did not raise this as an issue. We have created the help file in a platform-independent format (.pdf).

Recommendations

We recommend the development of a repository for transport metadata, led by the Department for Transport, to include existing repositories and approaches of which the NTDF may be the model for the overarching system. The executive agency tasked with this will need the power to require data owners to co-operate, to the extent that the initiative may need to be supported across government and at the highest level. We recommend speedy implementation of this system to enable the cataloguing and means of discovery and retrieval of data sources within the Department.

The issue of whether data held by government should largely be made available at no or little cost to stimulate the development of new businesses is a critical one. At present, however, there is no catalogue of data across government and no coherent way to find it: the NTDF should be implemented as part of the essential move to provide such a catalogue.

We recommend that the NTDF continue to be developed while converging with existing repositories and other approaches such as GovTalk and the ITS Registry. Specific areas for research and development are access control and security (see Appendix III) and data quality (see Appendix IV) characterisation and controlled vocabularies.

The Car Park Assets Use Case identified a very powerful exemplar for automatic updating of a central resource through agreed links to a data owners' databases and should be taken forward.

The Rail Use Case also proposed some work of value to the Department and users and should be taken forward.

Resourcing

There is a shortage of qualified candidates for software development posts of the type that were needed for this project. The two software developers that we recruited appeared to be well qualified, but in future we will ask candidates to provide software coding examples.

Research work on issues of survey characterisation and data quality for transport data, brought out in this project, should be considered; see Recommendations (and also Appendix IV).

Sustainability

Transport Direct is interested in the possibilities of the NTDF prototype as part of the drive to provide more integration and better access to transport data sets for users. The existing NTDF web site will continue to be hosted by the Cambridge eScience Centre until some other home is found for it. NaTII will make use of the NTDF concept.

Dissemination

Presentations of NTDF have been made at

- eScience All Hands Meetings, Nottingham;
- Data Grand Challenge Meetings;
- International Transport Congress;
- Transport Direct meetings on standards;

- 12th September and 8th December 2006 'industry' meetings in London;
- Inform UK (informatics) conference, York, November 2007

The two software developers are expected to produce papers for publication.

Lessons Learnt

Persuading data owners to allow access to their data is difficult. Data owners need powerful drivers to encourage them to allow access.

Most of the issues around the development of a system like the NTDF are human rather than technical, although there are technical challenges described elsewhere in this report. To sum up, the core human issues for the creation of a metadata catalogue are:

- People do not know what data is there, so there is a pressing need for a repository
- There must be a trusted access system
- There should be useful tool to link together various datasets
- Data owners should retain responsibility for their own data but allow it to be described
- There is a great advantage for diverse data sets to have metadata descriptions in a standard form, which we believe should be the semantic web language RDF, so that data sets can be searched for and found effectively using the specialist search language SPARQL.

Appendices

APPENDIX I: USE CASES

NTDF Urban Road Network Performance Monitoring

Version: v1

Author: John Polak / Imperial College London

GOAL
The goal with this Use Case is to enable an urban road network operator to develop improved systems for monitoring of the performance of their network. This is achieved by providing integrated and coherent access to a number of ITS data sources that are currently fragmented. These data sources include:

- Inductive loop detector data from urban traffic control (UTC) systems such as SCOOT;
- GPS-based vehicle tracking data from GPS service providers;
- Data from beacon/transponder based bus tracking and location systems;
- GSM-based terminal tracking data from mobile telephone network operators;

- ANPR camera data from both public and private sector data providers;
- Traffic surveillance camera data.

These ITS data sources should be considered in conjunction with existing sources of network performance information such as data from speed measurement loop detectors and moving car observer surveys.

Each data source provides a measurement of some feature of the network state at a particular point in the network or along a particular trajectory in the network, at a particular time. By appropriately combing these measurements we can estimate certain relevant underlying network state parameters, such as space mean speed.

The method of combination will require inter alia:

- For each data source, knowledge of the relationship between the measured network state features and the relevant underlying network state parameters (structural knowledge);
- For each data source, knowledge of relevant aspects of the measurement process associated with the measured network state features e.g., typology and magnitude of sampling and non-sampling variation (measurement knowledge);
- The alignment of local time measurement in each data source (temporal rectification)
- The projection of the measurements from each data source onto a common transport network description (spatial rectification).

The benefits for the road operator include:

- Increased temporal and spatial resolution in the measurement of network performance;
- Ability to develop improved indicators of network performance;
- Improved understanding of the effects of traffic management interventions.

Note: This Use Case deliberately does not consider real-time state-estimation and information dissemination, since this involves a significant number of additional complications. This situation will be considered in a separate Use Case in due course.

ACTORS
The actors involved include:

- Road network operator;
- Traffic management operator;
- Central government;
- UTC system provider;
- GPS data providers;
- GSM data providers;
- ANPR data providers;
- Traffic surveillance camera data providers;
- Traffic data consolidators and resellers.

PRE-CONDITION
The road network operator should have access to some (though not necessarily all) of the relevant ITS data sources.

POST-CONDITION
An estimate is produced of the performance of a particular network element at a particular time.

MAIN-FLOW

1. The road operator acquires relevant ITS data and metadata for each relevant data source and analysis period.
2. The information from each data source is temporally aligned.
3. The information from each data source is mapped to a common transport network representation.
4. State estimation is performed using structural and measurement knowledge and appropriate data combination method(s).

EXCEPTIONS
<<none yet>>

Dependency-UCs

<<none yet>>

National Rail Travel Survey (Rail Use Case)

Author: John Polak, Imperial College London

GOAL
The goal with this use case is to provide the means to facilitate the maximum use of the NRTS data both by parties within the rail industry and those outside. This will ensure that the benefits of the substantial investment made in the NRTS are maximised. This goal is achieved by providing a comprehensive metadata characterisation both of the NRTS data themselves and of key aspects of the survey and pre-analysis process.

The system developed under this use case must provide a range of functionality including:

- Comprehensive metadata creation and updating tools for the range of travel survey related data types encountered in the NRTS;
- Process metadata (and associated creation and updating tools) for the characterisations of the processes of sampling design, survey execution and data aggregation, imputation and weighting involved in the collection and pre-analysis of the NRTS data;
- Tools to support data discovery;
- Tools to enable users to flexibly extract relevant subsets of the NRTS data defined in terms for example of geographical, functional or travel behaviour dimensions;
- Tools to prevent the execution of unsuitable queries e.g., those that disclose personally or commercially sensitive information, those for which confidence intervals on key parameters are too wide;
- Tools for data export;
- Tools for access control both at the level of the entire NRTS collection and at the level of individual queries or retrievals.

Additional elements of functionality may emerge in later discussions.

The benefits for the data holder (DfT) include:

- Maximise use of the NRTS data;

- Minimise nuisance enquiries from ill-informed potential users;
- Improved understanding of the NRTS data and hence of the UK rail system.

ACTORS
The actors involved include:

- Department for Transport;
- The consultants directly involved in the NRTS survey design and data collection;
- Network Rail;
- Office of the Rail Regulator;
- Train Operating Companies;
- Other transport operators (e.g., bus operators, BAA);
- Local Authorities;
- Other consultants;
- Academic researchers.

PRE-CONDITION
A potential user formulates a question to which the NRTS is relevant.

POST-CONDITION
The user's query of the NRTS is satisfied.

MAIN-FLOW

5. The user is guided to the relevant parts of the NRTS (discovery).
6. The user is guided in the formulation of his research question as a suitable query to the NRTS.
7. The query is executed.
8. The data generated by the query is conveyed to the user.

EXCEPTIONS
It is possible that a query may not be able to be executed (due e.g., to confidentiality or statistical reasons)

Dependency-UCs: none yet

NTDF Car Park Asset Inventory

Author: John Polak / Imperial College London

GOAL
The goal with this use case is to enable Transport Direct to maintain an up top date inventory of car park assets in the UK. This information will play a vital role in planned extensions of Transport Direct's trip-planning services to include guidance to destination specific destination car parks.

A large-scale national survey of car park assets has recently been undertaken by Transport Direct. This establishes a baseline national car park inventory. However, this will need to kept up to date as car par operators make changes to the location, scale, pricing and operating regimes of their assets. Resources are not available to repeat the survey so the updating will depend on car park operators

voluntarily providing information to Transport Direct. Some car park operators may be reluctant to provide this information due to concerns regarding the direct costs of doing so and/or the leakage of commercially confidential information.

The system developed under this use case must provide a range of functionality including:

- A web service to enable the manual creation, correction and updating of information regarding individual car parks, including their type, location, access arrangements, number of spaces, tariff structure, surcharges, provision of those with special needs etc.
- Capability to (if and when permitted) automatically access the car park operators' own asset databases to retrieve this information when it changes ;
- Robust tools to validate change information;
- Robust tools to manage the propagation of change information;
- Robust tools for security and access control both at the level of operator and individual car park;
- Integration with the operational procedures of Transport Direct.

Additional elements of functionality may emerge in later discussions.

Although the focus of the use case is car part assets, the nature of the problem is generic and similar requirements may exist in many other domains.

The benefits for Transport Direct include:

- Increased quality of service to end users;
- Improved relations with car park operators;
- Avoidance of future survey costs.

The benefits for the car park operator include:

- Potential for increased demand and revenue due to increased exposure via Transport Direct;
- In the case of Network Rail, scope for increased multi-modal integration;

ACTORS
The actors involved include:

- Transport Direct (currently DfT)
- Related and dependent trip-planning information providers (e.g., various local authority services);
- Competing trip-planning information providers;
- Car park operators;
- Network Rail;
- Train Operating Companies;
- Other transport operators (e.g., bus operators, BAA);
- Local Authorities;

PRE-CONDITION
A car park operator makes some change to the features of a car park.

POST-CONDITION
This change is fully and correctly reflected in the information provided to travellers by Transport Direct.

MAIN-FLOW

9. The car park operator signals (manually or automatically) a change to the features of a car park.
10. The relevant change information is captured and validated.
11. The change information is propagate to the relevant Transport Direct operational databases.

EXCEPTIONS
<<none yet>>

Dependency-UCs
<<none yet>>

APPENDIX II: UPDATE ON USE CASES, MAY 2007

Dr. Hongqian Liang
Prof John Polak
Centre for Transport Studies
Imperial College London

Amine Tafat
Michael Simmons
Centre for Mathematical Sciences
University of Cambridge

May 2007

1 INTRODUCTION

The National Transport Data Framework project is currently working on three use cases as a means of focusing the development of its core technologies. These use cases are:

National Rail Travel Survey
Road Network Performance
Car Park Inventory

The objective of this deliverable is to present a brief update on each of these use cases, highlighting the overall approach adopted, the work completed to date and the work planned for the next three months.

2 THE NATIONAL RAIL TRAVEL SURVEY USE CASE

2.1 Background

An important class of transport data is data describing patterns of travel demand. These data are typically collected via sample surveys. The manner in which such sample surveys are designed and executed has an important bearing on the use that can be made of the data that are collected. It is highly desirable that information on survey design, execution and pre-processing be captured in a consistent and transparent fashion.

This use case concentrates on addressing this issue in the specific context of the National Rail Travel Survey, which is a major survey of rail demand undertaken on behalf of the DfT.

2.2 Objective

The objective of this use case is to provide a comprehensive XML based metadata description framework for all features of a travel survey including its statistical design, practical implementation, questionnaire design (question types, routing, checking), pre-processing (imputation, weighting, integrity) and data quality.

2.3 Approach

A set of XML elements for the description of travel surveys will be developed. This Travel Survey Set (TSS) will comprise a subset of the DDI Lite elements. TSS can be divided into 8 subsets.

- Statistical Design
- Survey Implementation
- Questionnaire Design
- o Question Types
- o Routing
- o Checking
- Pre-Processing
- Imputation
- Weighting
- Integrity
- Data Quality

A metadata framework has been developed to represent the whole process of the NRTS, including not only questionnaire itself, but also survey statistical design, implementing process, data pre-processing including imputation and weighting. This framework draws on a number of streams of existing work including:

Where possible, we have used XML tags from DDI Lite, a subset of DDI.

The description of questionnaire structure draws on the approach adopted in a number of commercial CAPI and CATI tools, including Blaise, SurveyCraft, Visual QSL and Quancept.

We have drawn on certain features of AskML and SuML to develop a characterisation of the survey implementing process.

We also created a metadata structure. For example, we have created XML elements for describing certain aspects of data quality and statistical processing.

2.4 Work completed to date

The following elements of this framework are currently available:

Questionnaire Description

In this part, we give XML tags for describing the questionnaire content, for example:

<question>

<question id></question id>

<question no></question no>

<question type></question type>

<question content></question content>

<answer type></answer type>

...

</question>

Data Collecting Quality Description[Richardson and Pisarski 1997]

The XML tags are available to describe missing, invalid, illogical data collected from respondents.

<answer>

<answer id></answer id>

<answer no></answer no>

<answer content></answer content>

<answer quality></answer quality>

...

</answer>

Data Statistical Analysis Description

In this part, we give XML tags for describing the conclusion from statistical analysis, such as percentage, mean, standard deviation, etc. For example, for multiple choice questions, we use the following tags to describe the percentage for each choice.

<choice>

<choice id></choice id>

<choice content></choice content>

<choice percentage></choice percentage>

...

</choice>

Questionnaire Design Basic Description

We use tags like <font>, <font size>, <style>, <title>, <subtitle>, ... etc to give a proper description for all the questionnaire layout. People may use this metadata to redesign or modify the questionnaires.

Survey Implementation Process Basic Description(including data collecting)

Tags are used to describe the properties like data collecting method, style etc. For example,

<data collecting>

<place>

<place id></place id>

<place name></place name>

<collecting time></collecting time>

<responding>

<success rate></success rate>

<failing rate></failing rate>

<invalid rate></invalid rate>

</responding>

...

<place>

....

</data collecting>

2.5 Future work

The work to be undertaken in the next period includes the following:

- Questionnaire Design General Description (including design, review)
- Survey Implementation Process General Description (including sampling, conducting,

- processing, data confidentiality)
- Survey Statistical Design Representation (including population of interest, sampling
- size, sampling design)
- Survey Pre-processing Representation (including imputation, weighting, integrating)

3 THE ROAD NETWORK PERFORMANCE USE CASE

3.1 Background

There are a wide variety of different data sources that provide information on the status and performance of the road network. These data sources are disparate in nature and are sourced from different organisations and suppliers. It is desirable to be able to provide the means of drawing together these different data sources to provide a more comprehensive and complete representation of the underlying status and performance of the road network.

This use case concentrates on addressing these issues in the context of a number of specific data sources including SCOOT and weather data

3.2 Objective

The objective of this use case is to develop a prototype that demonstrates the capability to flexibly discover, retrieve and combine data from the SCOOT system in Cambridge.

3.3 Approach

For each dataset (e.g., traffic and weather) a semantic description was generated using OWL (based on XML), which is stored in the NTDF repository. Mechanisms are provided to enable the user to search this metadata and request the data needed. Interfaces are provided to enable data that are so identified to be retrieved, independently of the details of their physical storage or location. Further tools are provided to enable the simple visualisation of the retrieved data.

A key functionality is the ability to submit a query that retrieves information from several separate data sources and combines this information in a meaningful way.

3.4 Work completed to date

An early prototype has been developed drawing on two data sources: (a) traffic flow data from the Cambridge SCOOT system and (b) data on wind speed, wind direction and rainfall is taken from a local weather monitoring station.

This prototype is currently being tested in a number of small exercises to explore to relationship between traffic flow and weather conditions.

3.5 Future work

Further work will focus on the refinement of this prototype, in two specific directions.

- Extending the range of data sources included, in particular, to include GPS trace data and air quality data.
- Enhancing the scope of the system to deal with live real time data feeds.

4 THE CAR PARK INVENTORY USE CASE

4.1 Background

A common requirement in many transport contexts is to create and update an asset inventory. This can be a complicated process, especially when ownership and control of the assets in question is distributed between many different organisations. These complications can be further exacerbated when commercial or other competitive pressures make organisations sensitive to the risks of inadvertent disclosure of information to competitors.

This new use case, which is in addition to those originally identified in the NTDF proposal, is concerned with addressing these issues in the context of recent work undertaken by Transport

Direct to establish an inventory of car parks, as part of enhancements planned for its travel planning service.

4.2 Objective

The objective of this use case is to enable Transport Direct to extend the coverage and scope of the car park inventory information recently collected by its contractor, Landmark Services and the enable this information to be kept up to date in the future.

4.3 Approach

Although discussions are still at an early stage, there seems to be considerable potential in developing an approach based on a user contribution model of data provision.

4.4 Work completed to date

This use case is still at an early stage of development. Initial discussions have been had with Transport Direct and further discussions with them and with Landmark Services are planned. As these discussions mature, a suitable work programme will be developed over the coming weeks.

APPENDIX III: ACCESS CONTROL

The eScience programme and Grid technologies in general have traditionally used X.509 certificates to identify users and network services. The EGEE particle physics Grid and the National Grid Service use the Virtual Organisation Management Service (VOMS) [2]. This was evaluated as a potential component of the NTDF but was found to be difficult to extract as a stand-alone module without breaking dependencies on the rest of the EGEE software stack. In addition, while the use of X.509 certificates for network services is a well-established practice (e.g. for web sites secured by SSL), the use of certificates by end-users is known to be problematic. Maintaining the "chain of trust" back to a central certificate authority entails a heavyweight procedure for obtaining a certificate. For example, Grid users are required to take photo-id, in person, to an institutional authority. This does not scale well. The revocation of comprised

certificates is also problematic: sites may be slow to update their list of revoked certificates and this is often a manual process.

It was clear that a more scalable, lightweight and "semantic-web-friendly" approach to access control was required for the NTDF.

There are many access control technologies in use on both public and private networks. Kerberos [3] is particularly popular. (It is the default means of authentication in later versions of MS Windows.) It is well suited for authentication across local area networks where there is a central server capable of acting as an authoritative database of client credentials. However, this is not a desirable feature of a loosely coupled, distributed network of data sources. In addition, Kerberos provides authentication but no functionality for authorisation.

As part of the metadata held in the NTDF, it would be useful to describe the access control policies of the various data providers. XACML is an XML dialect for described access control policies with a processing model for the interpretation and implementation of these policies. XACML v2 was adopted by the standards organisation OASIS in February 2005. [4] Work continues on v3 of the protocol. Sun have produced an open source Java implementation, which we use in the NTDF. As XACML is expressed in XML, policy descriptions may be referenced in RDF triples as web resources with a URL. It is therefore consistent with the general architecture of the NTDF.

In XACML users may be identified by most common methods, including email address, LDAP entry and HTTP user-id. If it possible to express a user-id system as an XML schema, it can be used by XACML. The NTDF uses HTTP basic authentication with a user's email address as their user-id.

XACML policies usually apply to web accessible resources with a URL. (Although they may also refer to local files.) Access is commonly restricted by user-id (or a group of user-ids, often with wildcards) and source IP address of the request. Any criteria expressible as an XML schema will also work, for example current date and time, membership of an organisation or physical location.

In the XACML processing model, requests for access to a resource are received by a Policy Enforcement Point (PEP). The PEP will express a request as XACML and forward it to a Policy Decision Point (PDP). The PDP will evaluate the request against a set of policies and return a result to the PDP. The PDP may be physically remote from the PEP. This distributed model is consistent with the NTDF architecture: the PEP may be part of the NTDF web interface while the PDP would reside with the remote data source.

As an illustration, we have implemented a tool to define XACML policies for resources held in the NTDF (see www.ntdf.org.uk/accesscontrol).Users are identified by their NTDF user-id and access control to these data sources is provided by a PEP and PDP which, for the purposes of this prototype, both reside on the NTDF server.

We also provide a tool with which simple XACML policies may be built for user-contributed data sources. Policy files are stored on the NTDF web server for access by remote PDPs.

In summary, XACML is a well established standard protocol for describing and enforcing access control policies. The open source implementation provided by Sun lags a little behind the current published standard, although the SICS implementation[b] and the commercial product from

Axiomatics[c] are fully backwards compatible. XACML has been adopted by the Fedora[d] digital repository project. The Google Code project has recently released an independent Java implementation of XACML: see http://code.google.com/p/enterprise-java-xacml/ The whole area of distributed, fine grained access control at the level of sophistication provided by XACML is still relatively unexplored and under-utilised. As the "web of documents" evolves towards a "web of data"[e] it will be necessary to build distributed access control models for data that is not meant for the public domain. XACML and its use in the NTDF is a first step in this direction. A demonstrator application for access control has been put into the system whereby registered users can limit access to the metadata they add to other users with email addresses from a specific domain.

[1] Security Research Challenges for e-Science:

http://www.nesc.ac.uk/teams/stf/documents/SecurityResearchAgendaMay05.pdf

[2] http://hep-project-grid-scg.web.cern.ch/hep-project-grid-scg/voms.html

[3] http://web.mit.edu/Kerberos/

[4] http://www.oasis-open.org/committees/xacml/

[5] http://www.sics.se/spot/xacml_3_0.html

[6] http://www.axiomatics.com/

[7] http://fedora.info/

[8] http://www.w3.org/DesignIssues/LinkedData.html

APPENDIX IV: DATA QUALITY

Imperial College London

National Transport Data Framework

The Characterisation of Transport Data Quality:

Concepts, Methods and Applications

Professor John W. Polak

Hongqian Liang

Centre for Transport Studies

Imperial College London

j.polak@imperial.ac.uk

February 2008

1. INTRODUCTION

Sound transport research and policy making depends upon the availability of appropriate, high quality and up-to-date information. Thus, the means by which transport-related data are collected, stored, processed and made available are of central importance to the outcome of research and practice. The transport research and policy questions facing the UK are more complex than ever before and are making growing demands on underlying data systems – requiring both demand and supply data of ever greater spatial and temporal precision and scope, and socio-economic disaggregation. At the same time as these demands are growing, a wide range of new forms of transport data are becoming available, current examples of which include spatio-temporal trace data from a number of different forms of positioning and navigation system, transaction data from smartcard, tag and related payment systems, and large volumes of data arising as a by-product of the operation of network management and traffic control systems. The volume and range of these data are likely to increase substantially over the coming decades.

A key issue in the effective exploitation of both existing and emerging transport data sources is data quality. Aside from the general desirability that those using data should be fully informed of the data's properties and limitations, there are at least two specific considerations that make a current focus on data quality both timely and necessary. The first is that the technological trends alluded to above are likely to lead to situation in which in the future the range and number of entities serving as 'data producers' will dramatically increase and, in particular, increase to include many more informal and *ad hoc* sources of data (e.g., data provided by participants in social networks). Such data are likely to manifest much greater variability, both in their nature and quality, than do current data sources thus making it even more important that we have sound methods of establishing the provenance and quality of individual data items. The second consideration is that it seems likely that we are entering an era in which transport data will increasingly be re-used for purposes that are potentially widely different from those that originally motivated its collection. Although such re-use may stimulate beneficial innovation, it may also propagate and amplify the consequences of any weaknesses in the quality of the underlying data sources, hence, again, a need for renewed attention to be paid to the question of data quality. These considerations are particularly important of the National Transport Data Framework project, which is directly concerned with facilitating these types of development.

To date, the treatment of data quality in the transport literature has been fragmented, with different aspects being emphasised in different areas and little general agreement regarding how data quality should be conceptualised. The aim of this paper is to provide a more comprehensive treatment of transport data quality. We introduce the concept of the *data supply chain*, which we conceive of as the sequence of processes of abstraction, transformation and computation that take place as one moves from a particular real world quantity to its measured counterpart residing in a dataset. The characterisation of data quality then becomes the process of fully describing this data supply chain. We show how this concept can be used as a means of unifying existing treatments of data quality and how it can also be used as a basis for the development of metadata systems to support the discovery, retrieval and analysis of transport data sources. We illustrate this approach with two quite different examples – a sample survey of rail travellers, the National Rail Travel Survey and inductive loop detector data from the SCOOT traffic control system. Finally, we conclude by discussing the implications of this work for the NTDF and similar systems.

2. CONCEPTS OF DATA QUALITY

The topic of data quality is widely discussed in many disciplines and has been the subject of a great deal of academic and professional activity. For example, Wang *et al*., (1993) identify more than 200 concepts (or dimensions) of data quality current in the information systems literature alone. However, despite this volume of activity, data quality remains an elusive concept. For example, after a wide ranging review of various concepts of data quality used in information systems design, Wand and Wang (1996) conclude that:

> "Despite extensive discussion in the data quality literature, there is no consensus on what constitutes a good set of data quality dimensions and on an appropriate definition for each dimension. Even a relatively obvious dimension, such as accuracy, does not have a well established definition."

Notwithstanding the heterogeneity in approach that is apparent across different fields, one area where there does seem to be some measure of consensus is that data quality is a multi-dimensional concept (Holt and Jones, 1998). Indeed, this point is explicitly emphasised by Crito and Norwood (1997) in the context of transport data.

It is therefore of value of briefly examine the range of different concepts of data quality that have been advanced both within the domain of transport and also in cognate domains. These cognate domains include information systems engineering, geographical information systems, economics and finance and health care. These domains were identified as relevant due to similarities between the types data encountered in transport applications and the characteristics of the data types typically dealt with in these domains.

2.1 Information Systems Engineering

Wang *et al*. (1995) undertook a comprehensive review of the concepts of data quality used in the academic literature in information systems engineering. They identified several tens of different concepts of data quality with currency in the field, of which the most frequently cited are shown (in order) in Table 1. Wang and Strong (1996) observed that many of these concepts appear to overlap and this proliferation of overlapping concepts appears to have lead to confusion both amongst academics and practitioners regarding how best to approach the issue of data quality. They suggest that this proliferation is due, at least in part, of the absence of any coherent methodological foundation for the discussion of data quality.

In later work, Wand and Wang (1996) attempt to address this weakness by proposing an ontological model of the relationship between real world states and the data that represent these states in information systems, from which they derive what they term 'intrinsic' data quality dimensions, signifying inherently different ways in which the information system representation might mis-represent the real world. The four intrinsic' data quality dimensions they propose are:

- Completeness: No loss of information regarding the relevant features of the real world.
- Unambiguousness: No possibility that the data can be interpreted other than in the intended and correct manner.
- Meaningfulness: No possibility that any data item cannot be related to a corresponding real world entity.
- Correctness: No possibility of relating the data in the information system to the wrong real world state.

Table 1 Data Quality Concepts from Information Systems Engineering (Source: Wang *et al*., 1995)

Rank	Concept	Rank	Concept
1	Accuracy	14	Importance
2	Reliability	15	Sufficiency
3	Timeliness	16	Usableness
4	Relevance	17	Usefulness
5	Completeness	18	Clarity
6	Currency	19	Comparability
7	Consistency	20	Conciseness
8	Flexibility	21	Freedom from bias
9	Precision	22	Informativeness
10	Format	23	Level of detail
11	Interpretability	24	Quantitativeness
12	Content	25	Scope
13	Efficiency	26	Understandability

The framework proposed by Wand and Wong (1996) is a useful conceptual device but they do not discuss how to operationalize it in a practical setting – for example how to associate metrics with each of their intrinsic dimensions nor how to populate such metrics, once defined, with meaningful values.

2.2 Geographical Information System

The geographical information systems community has devoted considerable effort to the question of spatial data quality (Amrhein and Schut, 1990). Whilst the GIS community deals with many different types of thematic data, in so far as these data are spatially referenced, there is an underlying unity to the analysis task (in contrast for example to the diversity of analysis tasks encountered in information systems engineering, and indeed in transport).

Accordingly, the GIS community (academic and commercial) have evolved a range of standards, including standards for spatial data quality. van Ort (2005) identifies a number of spatial data standards that include elements of data quality. These include:

- US-SDTS. The US spatial data transfer standard adopted in 1992 (Department of Commerce, 1992; O'Looney, 2000)
- CEN/TC287. Technical committee 287 of the Comité Européen de Normalisation (CEN) developed the European pre-standard ENV 12656. During the process, ISO started standardisation and CEN/TC287 was subsumed into ISO/TC211.
- ISO/TC211 (2002). Technical committee 211 of the International Standardisation Organisation (ISO) has developed a number of international standards for geographic information: 19113 (Quality principles), 19114 (Quality evaluation procedures). These elements are described as part of the 19115 (Metadata) -- see ISO (2002, 2003a,b) and Kresse and Fadaie (2004).

van Ort (2005) identifies a total of eleven dimensions of data quality discussed in these various standards. These include:

- Lineage
- Positional accuracy
- Attribute accuracy
- Logical consistency
- Completeness
- Semantic accuracy
- Usage, purpose, constraints
- Temporal quality
- Variation in quality
- Meta-quality
- Resolution

Amongst these concepts, perhaps the most interesting in the current context is meta-quality, which provides information on the quality of the quality description. For example if the positional accuracy is estimated from a smaller sample size, then that estimate is of lower quality. The three standards US-SDTS, CEN/TC287 and ISO/TC211 treat meta-quality as a part of the other elements and require it to be documented if possible.

Overall, amongst these three standards, the US-SDTS seems to provide the most comprehensive treatment of data quality, formally defining five concepts of data quality based on ideas of accuracy, consistency and completeness (see Table 2). These concepts bear marked similarities to idea of intrinsic quality dimensions put forward by Wand and Wang (1996).

Although the GIS community appears to have well developed data quality concepts and standards, van Ort (2005) notes however, that not all the standards cover all these dimensions and that there are contradictions in the definition and implementation of the same concepts in different standards.

Table 2 Five Categories for Data Quality in the Spatial Data Transfer Standard (Source: O'Looney, 2001 and Turner, 2002)

Category	Definition	Example
Positional Accuracy	The degree of horizontal and vertical control in the coordinate system.	The available precision or detail of longitude and latitude coordinates.
Attribute Accuracy	The degree of error associated with the way thematic data is categorized.	The degree to which a soil description is likely to vary from a soil measurement taken from the corresponding location.
Completeness	The degree to which data is missing and the method of handling missing data.	The ability to estimate crime rates in specific areas may be compromised if data is not available for specific areas.
Logical Consistency	The degree to which there may be contradictory relations in the underlying database.	Location data on some crimes may be based on the place where the crime occurred, while for other crimes the location might be the place where a crime report is taken.
Lineage	The degree to which there is a chronological set of similar data developed using the modelling and processing methods.	Population estimates may not be available for all years; may be estimated on different days of the year; or may be estimated using different estimation techniques and data sources.

Veregin and Hargitai (1995) and Veregin (1999) provide a useful summary of what they regard as the most important dimensions of data quality considered in this field. These include:

- Accuracy: Here accuracy is understood to include spatial accuracy, temporal accuracy and thematic accuracy.
- Resolution: Resolution (or precision) refers to the amount of detail that can be discerned in space, time or theme.
- Consistency: Consistency here refers to the internal validity of the data, that is, the avoidance of circumstances in which there are internal contradictions amongst the data.
- Completeness: Completeness refers to a lack of errors of omission in a database.

Although these concepts appear to have been developed independently of the work of Wand and Wang (1996), it is again notable that there appears to be a considerable degree of overlap with what these authors proposed.

Veregin and Hargitai (1995) also provide an interesting discussion of the different styles of data quality control. They identify three styles prevalent in the GIS/spatial data community:

- **The Minimum Quality Standards Approach**: In this approach, data quality is the responsibility of the data producer and a regulator, usually but not always a government, sets minimum quality standards that producers of data must conform to. This approach is based on compliance testing strategies to identify databases that meet the required quality thresholds. An example of such an approach and associated standard is the National Map Accuracy Standards adopted by the US Geological Survey in 1946. Veregin and Hargitai (1995) observe that "this approach lacks flexibility, in some cases a particular test may be too lax while in others it may be too restrictive".
- **The Metadata Standards Approach**: In this approach it is the data user who is responsible for assessing the fitness-for-use of the data. The data producer's responsibility is to document the data sufficiently well to enable these judgements to be made in an informed manner. Metadata are used to provide this documentation. An example of this approach is the US-SDTS, which allows data producers to flexibly describe their spatial data. However, Veregin and Hargitai (1995) point out that standards such as US-SDTS are 'one-way' – they allow the producers to describe their data but they do not permit feedback from the data users to the producers.
- **The Market Standards Approach**: In this approach, there is a two-way flow of information from producers to users (regarding the content and quality of the data) *and* from users to producers (regarding specific data problems). Veregin and Hargitai (1995) cite the example of Microsoft's Feedback Wizard, a utility that lets users email reports of map errors.

2.3 Economics and Finance

In economics and financial data quality is, for obvious reasons, critically important. Particularly at the level of national accounting, in recent years there have been a number of exercises undertaken that have sought to better understand how to characterise data quality and, at a practical level, improve the quality of national statistical outputs.

One of the most prominent of these exercises has been undertaken by the International Monetary Fund (see, Carlson, 2000, 2001; Carson and and Liuksila, 2002). This exercise established a data quality framework for national economic and financial data. This framework characterised data quality in terms of five dimensions (Laliberté *et al*., 2004).

- Integrity: What are the features that support firm adherence to objectivity in the production of statistics so as to maintain users' confidence?
- Methodological soundness: How do the current practices relate to the internationally agreed methodological practices for specific datasets?
- Accuracy: Are the source data, statistical techniques, and supporting assessments and validation techniques, inclusive of revisions studies, adequate to portray the reality to be captured by specific datasets?
- Serviceability: How are users' needs met in terms of timeliness of the statistical products, their frequency, consistency, and their revision cycle?
- Accessibility: Are effective data and metadata easily available to data users, and is there assistance to users?

For each of these interrelated dimensions of quality, the framework identifies a number of specific elements and for each element a number of measurable indicators. Carson and her colleagues argue that this hierarchical structure of dimensions, elements and indicators enables the framework to combine simplicity with flexibility.

It is notable that the IMF data quality concepts place considerable emphasis on process aspects of data quality (e.g., seeking to enforce impartiality and methodological clarity) and rather less emphasis on comparisons between the data and the real world counterparts. Overall, the emphasis seems to be principally on a goal of data that are fit-for-purpose rather than in any absolute sense, correct.

A similar structure for the characterisation of data quality has been developed by Eurostat (Grünewald, and Linden, 2001), with the intention that it should be applied to all the statistical series dealt with by Eurostat, including economic and financial data. The Eurostat framework defines six key dimensions of data quality:

- Relevance: Are the data what the user expects?
- Accuracy: Is the figure "reliable"?
- Comparability: Are the data in all necessary respects comparable across countries?
- Coherence: Are the data coherent with other data?
- Timeliness and punctuality: Does the user get the data in time and according to pre-established dates?
- Accessibility and clarity: Is the figure easily accessible and understandable?

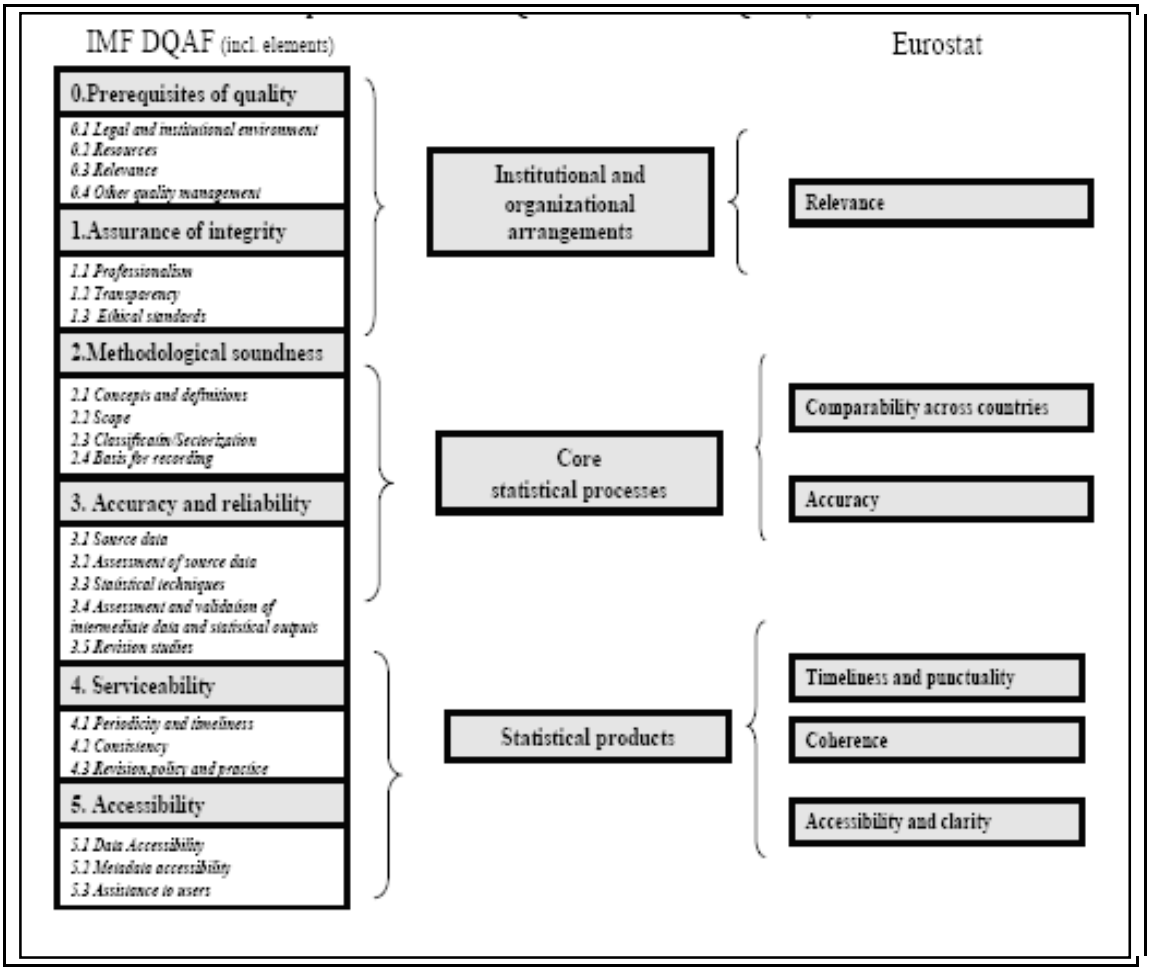
It is notable that the Eurostat data quality concepts appear to overlap in their semantic content to considerably great extent than the IMF concepts.

Laliberté *et al*., 2004 reports a recent study that attempted to reconcile these two frameworks. The interest in this study in the current context is that it abstracted from the dimensions in each framework to identify three higher level sets of constructs underlying both conceptual; approaches to data quality. These were (see Figure 1):

- Institution and organization arrangements
- Core statistical processes
- The dissemination of statistical products

These higher level concepts provide a useful context for some of our later discussions.

FIGURE 1: Comparison of IMF Data Quality Framework and the Eurostat Framework (Source: Laliberté *et al*., 2004)



2.4. Health care

In a similar manner to the finance and economics sector, the health care sector has devoted significant effort to the issue of data quality, much of which has been driven in recent years by the emergence of the idea of an electronic health record (EHR) and the more general growth in the field of health informatics (Orfanidis *et al*., 2004).

A number of recent papers have reviewed the approaches to data quality used in the health care sector (Brown and Sonksen, 2002; d'Onofrio and Gendron, 2001; NHS, 2002). These studies have identified a number of data quality concepts, which are summarised by Orfanidis *et al*., (2004) as follows:

- Accessibility and availability. Interpreted as meaning that authorized users should be able to access the HER data through fast, easy-to-use interfaces suited to the needs of both professionals and patients.
- Usability. Interpreted as meaning that EHR data should be accessible in different data formats and from different kinds of hardware and networks.
- Security and confidentiality. Interpreted as meaning that access to EHR data should be restricted to authorized users only.
- Provenance. Interpreted as meaning that EHR data should show its source and the context of data, linked to metadata about provenance of data.
- Data validation. Interpreted as meaning that EHR metadata should contain information about what validation the data themselves have been subject to.
- Integrity. Interpreted as meaning that EHR data should conform to appropriate accreditation standards preventing inconsistency and duplication.
- Accuracy and timeliness. Interpreted as meaning that EHR should reflect accurately the real world counterparts they refer to.
- Completeness. Interpreted as meaning that EHR data should when necessary indicate relevant complementary data, possibly with links to other data as appropriate.
- Consistency. Interpreted as meaning that different items of EHR data do not contradict one another.

These quality concepts again display considerable overlap and redundancy. However, they introduce the concept of data provenance and also emphasis the importance of documenting the processes, including the processes of validation that generated that data in hand.

Concepts of data quality are also of importance outside the context of EHRs. For example, Malin and Keating (2005) have recently emphasised the need for better data quality standards in epidemiological research and advocated the adoption of a practice of grading the data used in scientific research and in policy analysis according to its quality, analogous to the practice of grading medical evidence, which they argue has became standard practice in the development of clinical practice guidelines (e.g., Harris and Woolf, 2001; Lohr, 2004; West et al., 2002)

Many health care organisations have developed frameworks, based on these and related quality concepts, in order to measure the data quality of health data. One of the most prominent is that developed by the Canadian Institute for Health Information (see, Long *et al*., 2003, CIHI, 2005). The CIHI framework, which is based on initial work by Brackstone (1999), is particularly comprehensive. It defines five key quality dimensions and within each of these dimensions a number of characteristics, each of which has associated with it a number of measurable criteria. A total of 19 characteristics are defined and 58 criteria – see Figure 2.

The CIHI is particularly oriented towards the use of sample survey data (which is also a key data type in transport studies). The framework pays particular attention to characterising various feature of the data collection and survey processing. Indeed, it is notable that the quality dimension ‘accuracy’ is defined exclusively in terms of these survey process elements, rather than in terms of a comparison with an external ‘ground truth’, in marked contrast to the practice in other fields reviewed. Another notable feature of the CIHI framework is the emphasis it places on the documentation of process as being an important aspect of data quality.

FIGURE 2: The Canadian Institute for Health Information Data Quality Framework

(Source: CIHI, 2005)

Accuracy	
Coverage	1 The population of reference is explicitly stated in all releases
	2 Known sources of under- or over-coverage have been documented
	3 The frame has been validated by comparison with external and independent sources
	4 The rate of under- or over-coverage falls into one of the predefined categories
Capture and Collection	5 Practices exist that minimize response burden
	6 Practices exist that encourage cooperation
	7 Practices exist that give support to data suppliers
	8 Standard data submission forms and procedures exist
	9 Data capture quality control measures exist
Unit Non-Response	10 The magnitude of unit non-response is mentioned in the data quality documentation
	11 The number of records received is monitored to detect for unusual values
	12 The magnitude of unit non-response falls into one of the predetermined categories
Item (Partial) Non-Response	13 Item non-response is identified
	14 The magnitude of item non-response falls into one of the predetermined categories
Measurement Error	15 The level of measurement error falls into one of the predetermined categories
	16 The level of bias is not significant
	17 The degree of problems with consistency falls into one of the predetermined categories
Edit and Imputation	18 Validity checks are done for each data element
	19 Edit rules and imputation are logical and consistent
	20 Edit reports for users are easy to use and understand
	21 Imputation is automatically derived from edits
Processing and Estimation	22 Documentation for all data processes is maintained
	23 Documentation for all systems, programs or applications is maintained
	24 The processing system has been tested after the last revision
	25 Raw data are saved in a secure location
	26 The sampling bias and variance of the estimates are at acceptable levels

FIGURE 2: The Canadian Institute for Health Information Data Quality Framework (Source: CIHI, 2005) – continued

Timeliness	
Data Currency at the Time of Release	27 The difference between the actual date of release and the end of the reference period is reasonably brief
	28 The official date of release was announced in advance of the release
	29 The official date of release was met
	30 Database or registry methods are regularly reviewed for efficiency
Documentation Currency	31 The recommended data quality documentation was available at the time of data or report release
	32 Major database or registry reports were released on schedule
Comparability	
Data Dictionary Standards	33 Data elements are evaluated in comparison to the CIHI Data Dictionary
	34 Data elements conform to the CIHI Data Dictionary
Standardization	35 Data are captured at the finest level of detail as is practical
	36 For any derived data element, the original data element is also maintained on the main database
Linkage	37 Standard Geographical Classifications (SGC) can be used
	38 Data are collected using a consistent time frame
	39 Codes are used to uniquely identify institutions
	40 Codes are used to uniquely identify persons
Equivalency	41 The impact of problems related to crosswalks or conversions falls into one of the predetermined categories
	42 Methodology and crosswalks or conversions are documented
Historical Comparability	43 Trend analysis is used to examine changes in core data elements over time
	44 The extent of problems in comparing data over time falls into one of the predetermined categories
	45 Accessible documentation on historical changes to the database exists
Usability	
Accessibility	46 An official subset of microdata is defined, created, made available and frozen per release for users where appropriate
	47 Standard tables and analyses are produced per release
	48 Products are defined, catalogued and/or publicized
Documentation	49 Data quality documentation exists per annual subset release
	50 Database or registry methods documentation exists for internal purposes per annual subset release
	51 A caveat accompanies any official preliminary release
Interpretability	52 A mechanism is in place whereby key users can provide feedback to, and receive notice from, the product area
	53 Revision guidelines are available and applied per annual subset release
Relevance	
Adaptability	54 Mechanisms are in place to keep clients and stakeholders informed of developments in the field
	55 The database or registry can adapt to change
Value	56 The mandate of the data holding fills a health information gap
	57 The level of usage of the data holding is monitored
	58 User satisfaction is periodically solicited

2.5 Transport

The issue of data quality has attracted considerable attention in the transport literature (see, e.g., Richardson *et al.*, 1995). There is a large and rapidly growing literature covering data collection methodologies, the characteristics of different transport data types and the technological and research challenges and opportunities associated with transport data methods (see, Jones and Stopher, 2000; Stopher and Jones, 2003; Stopher and Stecher, 2006).

However, these treatments are very fractured, reflecting, in part, the enormous diversity of data types encountered in the transport domain and the different priorities that apply to the assessment of data quality in different areas. In very broad terms, the field can be divided into two general categories, according to the type of data being considered – those treatments concerned with sample survey data and those concerned with traffic sensor data. These are of course not completely disjoint categories but the differences are sufficient to merit keeping this distinction in mind.

At the highest level, a number of studies have been sought to identify appropriate dimensions of data quality and to develop frameworks for the ongoing improvement of sample survey data quality. One of the most comprehensive efforts of this type was produced in a major review of the US Bureau of Transport Statistics, undertaken in the late 1990s and reported in Crito and Norwood (1997). This review focused principally on sample survey data types since these were the main types of data held and managed by the BTS. The study identified data quality as a key issue for the future of the BTS and proposed a general framework within which the topic could be considered. This framework identified three main high-level dimensions of data quality:

- Comparability: Referring to both the consistency of definitions of key quantities of the data collection methods used to measure these quantities.
- Accuracy: Referring to how well the measured value of a quantity approximates the true values and touching on issues of coverage, sampling and pre-processing.
- Variability: Referring to the nature and prevalence of sampling and non-sampling errors.

In addition to identifying these three dimensions of quality, the study also emphasised the related concept of *relevance* which they defined as "the appropriateness of the concepts, definitions and measurements, the level of subject and geographical detail and the timeliness of the data from the measurement system" (p32). Considerable emphasis was also placed to the establishment of data quality standards, which were regarded as themselves having at least three distinct dimensions:

- The establishment of consistent definitions and protocols.
- The definition of minimum acceptance criteria with respect to key data quality indicators.
- The definition of standards and protocols for the reporting data quality in published data.

There is also a strong emphasis throughout on the importance of comprehensive and clear documentation of all aspects of data quality (and relevance).

Overall, the work of Crito and Norwood (1997) reflects many of the concerns and approaches discussed in the health care sector; placing emphasis on developing a detailed understanding of the process aspects of the data collection process and the documentation and dissemination of this information in appropriate form to support the informed use and re-use of the data by end users.

A different approach is taken in the work of Turner (2002), who reports the results of a study undertaken into traffic data quality on behalf of the US Federal Highway Administration. After reviewing the treatment of data quality in a number of government and other sectors, they propose the following definition for traffic data quality:

> "Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data"

And associated with this definition, they also propose a number of dimensions of data quality, which closely reflect the general practice and understanding of data quality in the information systems community:

- Accuracy – The measure or degree of agreement between a data value or set of values and a source assumed to be correct. Also, a qualitative assessment of freedom from error, with a high assessment corresponding to a small error.
- Completeness (also referred to as availability) – The degree to which data values are present in the attributes (e.g., volume and speed are attributes of traffic) that require them.
- Validity – The degree to which data values satisfy acceptance requirements of the validation criteria or fall within the respective domain of acceptable values.
- Timeliness – The degree to which data values or a set of values are provided at the time required or specified.
- Coverage – The degree to which data values in a sample accurately represent the whole of that which is to be measured.
- Accessibility (also referred to as usability) – The relative ease with which data can be retrieved and manipulated by data consumers to meet their needs.

This fitness-for-use interpretation of data quality again parallels closely the approach adopted in the health sector. It has many attractions, especially in contexts like transport where there is limited scope for the identification or measurement of an absolute ground truth. However, as framed by Turner (2002), the concept is rendered extremely inflexible by the requirement that "Measuring data quality requires an understanding of all intended purposes for that data". As we pointed out in the introduction, the major source of growth in the use of transport data in the future is likely to be associated with the novel re-use of existing data sources. In this context, it is impossible to have knowledge of all possible uses to which a certain item of data might be put.

The requirements associated with the re-use of existing data have been explicitly addressed in a number of initiatives. For example, as part of the development of the US Intelligent Transportation Systems National Architecture a set is standards covering the use and re-use of archived traffic data have been developed (ASTM, 2006, Hu *et al*., 2002). The Archived Data User Service (ADUS) is designed to enable government agencies and others to retain ITS-generated data and make them available for analysis and re-use by others. Amongst the major functions included in the ADUS standard are those covering data integrity and data validation.

Similarly in the UK, the UTMC standard a variety of quality indicators for the exchange of traffic data between systems (MVA *et al*., 1998). The overall approach to data quality adopted in UTMC is to first identify what data needs to be exchanged, the to define and agree relevant descriptors of data quality based on an underlying set of data quality parameters, which in essence attempt to characterise the measurement process associated with each data type (see Table 3).

FIGURE 3: Definition of UTMC Traffic Data Quality Parameters

(Source: MVA et al., 1998)

Quality Parameter	Description
measurement_standard_deviation	Accuracy of measurement in terms of standard deviation
measurement_skew	Accuracy of measurement in terms of skew
measurement_significant_figures	Accuracy of measurement expressed in terms of number of significant figures
measurement_decimal_places	Accuracy of measurement expressed in terms of number of decimal places
measurement_location_precision	Precision of where measurement is taken
measurement_location_spread	Distance over which measurement must be taken
measurement_spatial_parameters	Dimensional requirements of measurement - 1, 2 or 3 dimensions
measurement_duration_spread	Time period over which measurement is taken
measurement_regularity	Elapsed time between measurements
measurement_lifetime	Time period over which measurement is valid
measurement_age	Elapsed time since measurement taken
number_of_measurements	Total number of individual measurements required to produce single resultant value
measurement_type	Derivation of measurement eg observed, calculated, modelled or estimated
measurement_synchronisation	Synchronisation requirements between measuring equipment and other UTMC systems
source_reliability	Perceived reliability of source
standard_ratification	Degree of ratification of standard eg EU / national standard, de facto, proprietary
relationship_to_standard	Requirements for standards to be applied to measurement
verification/confirmation_time	Elapsed time between verification/confirmation
verification/confirmation_required	Requirements for verification/confirmation
message_priority	Priority of message
supplemental_advice_requirements	Requirements for supplemental data to be sent

The approach adopted by UTMC provides considerable flexibility in the characterisation of data quality (which is consistent with the ethos of flexible re-use of data) but is of course scope of the fundamental data quality parameters.

2.6 Summary and synthesis

The review of existing approaches to data in transport and cognate fields has highlighted a number of important issues.

Although there is an unhelpful proliferation of concepts and terminology associated with the idea of data quality, if one looks beyond the linguistic labels to the underlying implementations proposed, there is in fact a remarkable degree of convergence in thinking on the key dimensions of data quality. In particular, the concepts of accuracy, completeness, meaningfulness and lack of ambiguity recur throughout the different literatures considered.

An important difference exists however between approaches adopted in different fields in terms of the extent to which they view quality as an inherent property of the data (usually judged relative to some notion of ground truth) or as a characteristic of how the data are generated and used. In the context of the NTDF, we tend very much to the latter view as being the most appropriate. That is, we believe that the most salient features of quality are those that allow analysts to take informed decisions regarding how data might be most appropriately used.

To illustrate this point, consider the example of a sample survey (such as a household travel diary) that is significantly affected by unit and item non-response. One extreme approach to data quality would be to simply record the magnitude of the non-response and set a maximum threshold level beyond which the data are deemed to be of poor quality and inappropriate for further analysis. However, what matters more to an analyst considering whether or not to use these data to pursue a particular practical analysis question is the characteristics of the pattern of non-response in relation to the question at hand – for example estimates of population averages are typically rather sensitive to non-response whereas regression slope parameters are less so. Thus fitness for use can only, in general, be judged at the point of use. Moreover, different features of the data will be relevant to different potential uses; this implies that the characterisation of data quality should seek to capture as complete a representation of the features of the data as possible. To illustrate this, and extending our example, if item level non-response in, say, household income has been dealt with by imputation approach then the characterisation of data quality should document what imputation process was used, since for example mean imputation will preserve sample means but distort the sample covariance structure, impacts that may be highly relevant to certain forms of subsequent analysis.

Our view is that, in general, intelligent analysts will want to know as much information as possible regarding how the data they are considering working with has been assembled so that they in turn can make fully informed decisions regarding their own analyses.

2.7 Data Supply Chain

This argument leads to our proposal regarding the *data supply chain*, which is an idea that is implicit in a number of the treatments of data quality we have reviewed in this paper but which, to our best knowledge, has not so far been explicitly developed in the transport or cognate literature.

We conceive of the data supply chain, as the sequence of processes of abstraction, transformation and computation that take place as one moves from a particular real world concept of interest to its measured counterpart residing in a particular dataset. In the case of a sample survey such as the household travel diary example described above, this supply chain will

include the practical definition of measurement quantities corresponding to concepts such as trip, mode, purpose etc., the elaboration of a questionnaire (including routeing, exclusions, range and other checking), the development of a sample frame for the relevant population of interest, the development of a sampling design for this population and the practical implementation of this sampling design through a process of survey administration, the analysis of survey outcomes in terms of unit and item non-response, internal and external data consistency, the processes of survey data pre-analysis including weighting and imputation.

From this standpoint, the characterisation of data quality of a particular data source essentially becomes the process of fully describing the relevant data supply chain.

If this view is accepted, then the key question becomes how to most effectively characterise this data supply chain. This raises the question of metadata methods and tools, which is considered briefly detail in the next section.

3. METADATA METHODS

3.1 Background and overview

Transport metadata methods have developed rapidly in recent years (see, Axhausen and Wigan, 2003; Westlake, 2005; Wigan 2001). Westlake (2005) defines metadata as:

> "…any information that is needed by people or systems to make proper and correct **use** of the real statistical data, in terms of capturing, reading, processing, interpreting, analysing and presenting the information (or any other use). In other words, metadata is anything that might influence or control the way in which the core information is used by people or software".

Froeschel *et al*. (2003) propose a five-dimensional approach to the classification and description of metadata. These five dimensions (or facets) are:

- a *structure* facet (the "entity" dimension: what things are);
- a *view* facet (the "role" dimension: the different ways things are considered);
- a *form* facet (the "material" dimension: how things are represented);
- a *stage* facet (the "process" dimension: how and where things are used), and
- a *function* facet (the "agent" dimension: the purpose things are used for).

Westlake (2005) describes the early development of metadata methods as follows:

> "Metadata started with the first general purpose statistical packages in the 1960's, software such as BMD, XTab and SPSS. These all needed information about 'Variables': which punch card columns (or paper tape fields) contained information about which statistical measurement, had the values been scaled, what range of values was allowed, had any special codes been used (for example to indicate missing data), what is a suitable label for the measurement, what were the meanings of the codes used for classification variables, and so on? Quickly the idea of the Data Dictionary (or, sometimes, the Codebook) gained acceptance. This was intended to contain all the information about the data that was needed to perform statistical analysis, but which was not the actual data".

A range of international standards for metadata schema have been established and the most popular metadata schemas are (Chalasani *et al*., 2002):

- Dublin core (http://dublincore.org)
- AACR2 (Anglo-American Cataloguing Rules) (www.nlc-bnc.ca/jsc/docs.html)
- ISAD(G) General International Standard of Archival Description.
- (www.ica.org/biblio/com/cds/isad_g_2e.pdf)

There are numerous standards or initiatives based on these metadata schemas. Current metadata standards or initiatives include (Chalasani *et al*., 2002):

- DDI (the Data Documentation Initiative) (www.icpsr.umich.edu/DDI/) is an XML-based DTD developed within the world of social science data archives to describe archived data (mainly rectangular survey-files).
- The Cristal data object model (www.faster-data.org/Metadata/papers/Cristal.htm) is under development at CBS and the Statistical Open Source group and aimed at a generalised object model for describing micro datasets and multidimensional cubes.
- GESMES (www.ecb.int/stats/gesmes/gesmes.htm) is a Eurostat supported metadata interchange protocol.
- ISO11179 (www.diffuse.org/meta.html) is an ISO/IEC standard for description of data elements
- CWM - Common Warehouse Metamodel (www.cwmforum.org/) is an Object Management Group (OMG) effort to create a metamodel for easy interchange of warehouse metadata between warehouse tools, warehouse platforms and warehouse metadata repositories.
- BRIDGE (www.unece.org/stats/documents/1998/02/metis/7.e.pdf) is a classification model developed by several European Statistical agencies (within the IMIM project).

Among the above the most advanced, flexible and user friendly standard is DDI. DDI is an XML-based standard for the content, presentation, transport, and preservation of metadata, with an emphasis on sample survey data from the social and behavioural sciences. It and associated tools such as NESSTAR has been used by a number of researchers as a means of documenting and exchanging transport sample survey data (e.g, Axhausen, 2000, Chalasani *et al*., 2002; Levinson and Zofka, 2006).

An extension of these approaches is the Resource Description Framework (RDF). The RDF metadata model is based upon the idea of making statements about resources using first order predicate logic, i.e., a subject-predicate-object combination, called triple in RDF terminology. Although this is a very simple concept, it proves to use it to represent complex relationships.

In addition to these general purpose metadata methods and tools, there are a number of more specialised tools that are of relevance. For example, the social survey research community has developed a number of metadata tools for various stages of the data supply chain, focusing principally on the representation of sampling design and survey questionnaire design and also the inter-operability of data between systems. Notable examples include:

- AskML: (www.triple-s.org/ssstienhaara.htm)
- Blaise: http://www.westat.com/statistical_software/blaise/index.cfm
- Quancept: (www.spss.com/quancept_cati/)
- SurveyCraft (www.spss.com/surveycraft/)
- Triple-S (www.triple-s.org/)
- Visual QSL (www.pulsetrain.com/solutions/application/visual_qsl.htm)

In addition, Westlake and Krishnan (2006) describe the development of StatModel, a tool for the XML based representation of statistical models and demonstrate how it can be used to provide a portable and discoverable representation of survey pre-processing analysis.

Likewise, in the field of sensor data, there are a number of existing tools that provide data supply chain capabilities. The most notable is SensorML, developed by the OpenGeospatial Consortium (OGC), which models sensors as processes that convert real phenomena to observation data. SensorML provides standard models and an XML encoding format for describing any process including sensor metadata and the detecting process of measurement and instructions for deriving data from observations. All processes define their inputs, outputs, parameters, and methods. The metadata structure provided by SenosrML includes sensor identifiers, classifiers, constraints (time, legal, and security), capabilities, characteristics, contacts, and references, in addition to inputs, outputs, parameters, and system locations. SensorML also follows semantic web protocols and can easily be encoded for semantic web applications.

3.2 Appraisal of the capabilities of metadata tools

It is clear from this brief overview that there exist a number of useful metadata methods and tools and that progress has been made in representing certain stages of the transport data supply chain. However, there are still some critical gaps, principally relating to the detailed semantics of survey implementation and the characterisation of typical transport sensor systems. Address these gaps constitutes a significant research challenge for the future.

4. ILLUSTRATIVE APPLICATIONS OF THE DATA SUPPLY CHAIN CONCEPT

This section briefly describes two applications of the data supply chain concept to transport data sets. The aim is to illustrate both how existing metadata methods and tools can potentially be used to enable the characterisation of relevant stages of the data supply chain and to highlight areas in which further work in the development of appropriate methods and tools is required.

The two applications we consider deliberately relate to very different types of transport data – traffic data provided by inductive loop detectors (ILD) and behavioural survey data on patterns of rail travel demand collected in the National Rail Travel Survey (NRTS).

4.1 Inductive loop detector data

Inductive loop detectors (ILDs) are the most common form of traffic sensor in the UK. ILDs are loops of wire typically 2m long and 1.5m wide that are embedded several centimetres under the road surface. ILDs are usually associated with single lanes but in several applications, including in urban traffic control systems such as SCOOT, ILDs are also deployed in cross-lane configurations. ILD are point detectors, so are located at a specific position on a road link. They are connected to a power source, which applies an oscillating voltage which generates a magnetic field around the ILD. If a metallic object such as a vehicle passes over or rests on the loop, the inductance of the loop is reduced. If this decrease exceeds a certain threshold then the presence of a metallic object is signalled by the output of a binary ‘1’ bit otherwise a binary ‘0’ bit is output. This threshold value can be adjusted to enable the ILD to be tuned to detect different types of vehicle e.g., bicycles vs HGVs. Most ILDs in the UK operate at a sampling frequency of 4Hz. In countries such as the USA, where ILDs typically operate at much higher sampling frequencies (e.g., 60Hz), ILDs can also provide analogue output, effectively providing a signature for individual vehicles.

Each ILD is connected to an outstation that facilitates communication between the ILD and a control centre (Klein and Kelley, 1996; Robinson, 2005; Walmsley, 1982).

ILDs are known to be subject to a variety of physical and measurement errors affecting both the availability of data from ILDs and the quality of the data that are available (Head, 1982). Relevant factors affecting the quality of data retuned by ILDs include for example, physical damage to the ILD or its outstation or communications link, drift in ILD calibration, desensitization, electromagnetic interference and malign traffic conditions (such as stationary or parked vehicles). A number of diagnostic and remedial treatments for ILD data have been developed to deal with these problems (e.g., Chen and May, 1987; Hu *et al*., 2001; Robinson, 2005, Robinson and Polak, 2006).

Single loop ILDs are used to collect traffic flow and occupancy information and dual loop ILDs (comprising two single loop ILDs spaced several metres from one another) are used to collect speed information. The interpretation of ILD outputs in terms of fundamental traffic parameters depends on a variety of parameters that characterise the tuning and calibration of the loops (Cherrett *et al*., 2000; Robinson, 2005; Krishnamoorthy, 2008).

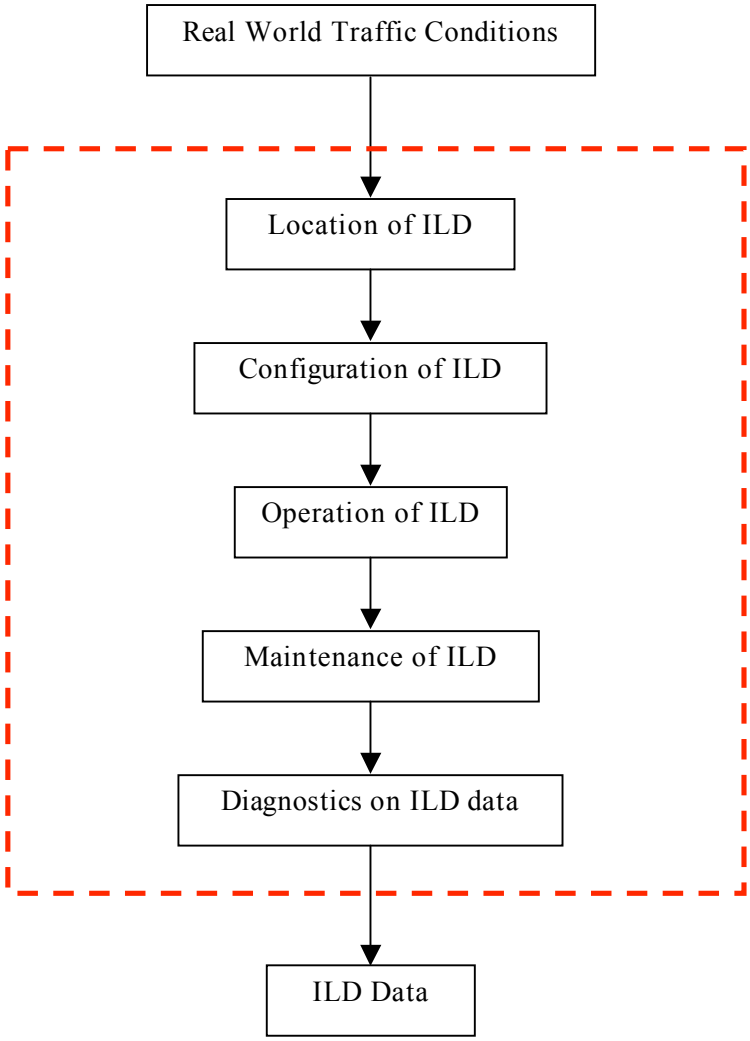
A simple version of the data supply chain for ILDs is shown in Figure 4. This figure shows the dependencies involved in the transformation from the real world traffic flow to ILD data. This transformation will depend the location of the ILD (i.e., road link and lane including, if relevant cross lane configuration and the position on the link), the configuration of the ILD (i.e., the tuning and calibration of the various internal and external ILD parameters), the operation of the ILD (i.e., polling rate, analogue to digital conversion, analogue or digital output, local data aggregation, physical and functional diagnostics etc.), the maintenance policy of the ILD (i.e., period between re-tuning, triggers for maintenance interventions) and on any diagnostic pre-processing (which might be as simple as the flagging dubious data items or as complex as the imputation of missing data) applied to the data before it is released for further analysis.

The requirement is thus for a set of methods and tools that enable these transformations to be comprehensively and consistently described. By way of illustration, one tool that can go some way towards addressing these requirements is SensorML.

SensorML is an XML-based markup language for describing sensor systems, developed by the Earth System Science Centre at The University of Alabama in Huntsville[2]. The initial objective of the development was to support applications in the remote sensing domain but it has come to be used more widely. It provides tools to enable the description of the operation of both dynamic and stationary sensor platforms and both in-situ and remote sensors. Its functions include sensor discovery, sensor geo-location and the representation of sensor operation including the logical and physical structure of the sensor and individual detectors and the processes carried out processing of sensor observations (using a sensor programming mechanism).

2 See, http://vast.uah.edu/SensorML/

FIGURE 4: Simplified Data Supply Chain for ILD Data



The following code fragment illustrates how SensorML can be used to provide a very simple description of the location aspects of the ILD data supply chain.

```
<identification>
         <IdentifierList>
            <identifier name="longName">
               <Term qualifier="urn:ogc:def:identifier:longName">Oxford Circus No 135
Inductive Loop Detector</Term>
            </identifier>
            <identifier name="shortName">
               <Term qualifier="urn:ogc:def:identifier:shortName">OC 135 ILD</Term>
```

```
            </identifier>
            <identifier name="modelNumber">
               <Term qualifier="urn:ogc:def:identifier:modelNumber">7440</Term>
            </identifier>
            <identifier name="manufacturer">
               <Term qualifier="urn:ogc:def:identifier:manufacturer">Innovation
Instruments</Term>
            </identifier>
         </IdentifierList>
</identification>
<classification>
         <ClassifierList>
            <classifier name="intendedApplication">
               <Term qualifier="urn:ogc:def:classifier:application">traffic flow</Term>
            </classifier>
            <classifier name="sensorType">
               <Term qualifier="urn:ogc:def:classifier:sensorType">single loop</Term>
            </classifier>
            <classifier name="sensorType">
               <Term qualifier="urn:ogc:def:classifier:sensorType">double loops</Term>
            </classifier>
            <classifier name="sensorType">
               <Term qualifier="urn:ogc:def:classifier:sensorType">triple loops</Term>
            </classifier>
            <classifier name="sensorType">
               <Term qualifier="urn:ogc:def:classifier:sensorType">other types</Term>
            </classifier>
         </ClassifierList>
      </classification>
      <validTime>
         <StartTime>2008-01-01</StartTime>
         <EndTime>currentTime</EndTime>
      </validTime>
```

Although considerably more work would be required to develop a complete characterisation of the ILD data supply chain, the potential of this approach is apparent.

A further consideration is that although SensorML is encoded in XML Schema, the models and encoding pattern follow semantic web concepts of Object-Association-Object. Therefore SensorML can be easily be converted into a form (such as RDF) which is compatible with the semantic web.

4.2 National Rail Travel Survey (NRTS)

The National Rail Travel Survey is a sample survey of weekday passenger trips on the National Rail system. It was designed to provide comprehensive and detailed information about how passengers use the rail network (DfT and Transport Scotland, 2007). In particular, it provides more detailed information about passenger activity on the network than is typically available from ticket sales data and more geographically and temporally extensive and complete information than is available from existing survey sources such as the National Travel Survey and the National Passenger Survey. The principal applications envisaged for the NRTS data are the analysis and modelling of travel demand, both for public policy purposes and as an input to the rail franchise bidding process.

The methodology used for the NRTS is based on the rail passenger surveys carried out as part of the London Area Transport Survey (LATS). The coverage of the NRTS included the whole of the Great Britain. Data were collected from passengers by means of self-completion questionnaires administered at the station or on-board a train. In addition, passenger counts are carried out at the same time to give details of the volume of people using each station. The survey work was undertaken during the period 2004-2005 and approximately 450,000 questionnaires were completed in total. Each questionnaire obtained various information about the respondent's current rail trip, categorised into the following broad topics:

- Rail stations used
- Time of travel
- Access and egress modes
- Origin and destination addresses
- Trip purposes
- Ticketing information
- Demographic information

The raw data collected in the questionnaire were subsequently augmented with additional geo-coded data on stations and trip origins and destinations. The raw data were also subject to extensive pre-processing involving validity and consistency checking and, where appropriate, the imputation of missing data items. After checking and imputation, the sample data were expanded to represent the population of approximately 2.7 million rail trips on an average weekday.

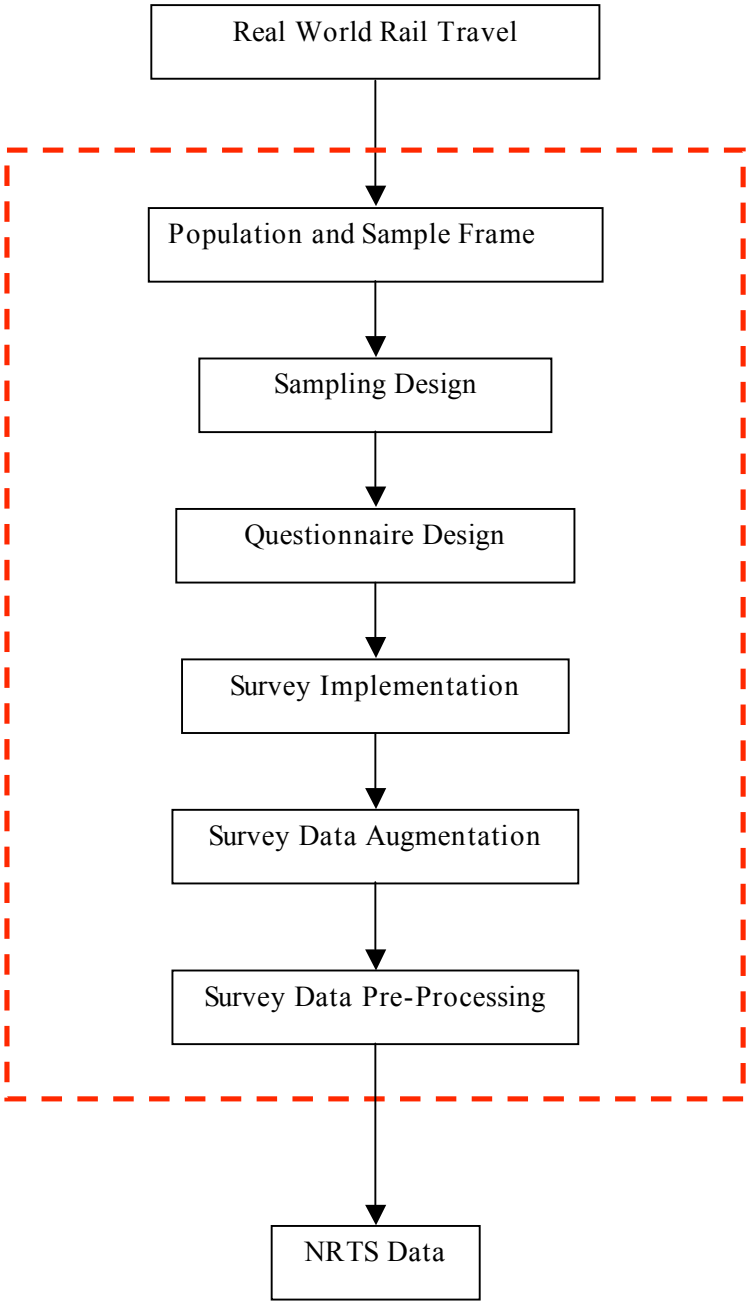
The NRTS is an excellent example of the type of complex and logistically challenging sample surveys that are common in the field of transport studies. There is an extensive academic and practitioner literature on such surveys (e.g., Cambridge Systematics, 1996; Stopher and Metcalf, 1996) which addresses various aspects of the design, and implementation of the survey and the augmentation and pre-processing of the survey data. This literature makes clear that this type of sample survey is subject to a wide range of sources of sampling and non-sampling variation (including item and unit non-response and reporting and recall errors on the part of respondents)

and also manifest strong instrument and method effects, with the result that the data one eventually obtains from such survey processes can depend strongly on the details of the organization and administration of the survey processes themselves.

A simple version of the data supply chain as it applies to the NTRS is shown in Figure 5. This figure shows the dependencies involved in the transformation from the real world rail travel to the NRTS data. This transformation will depend on a number of key stages including:

- The selection of the target population of interest (in the case of the NRTS, all weekday rail trips in Great Britain) and a sampling frame (railway stations and services)
- The statistical or sampling design of the survey (in the case of the NRTS, this appears to be a form of stratified random sampling).
- The design of the survey questionnaire including such considerations as question concept and wording, questionnaire routeing, exclusions and exceptions and internal range, logic and consistency checks.
- The practical processes of survey implementation including the protocols of respondent sampling and contact and for the management and operation of the field force of survey staff.
- The processes of survey data augmentation including the geo-coding of origins and destinations.
- The processes of survey data pre-processing including checking and imputation, sample re-weighting and expansion.

FIGURE 5: Simplified Data Supply Chain for NRTS Data



As with the earlier ILD example, the technical requirement is for a set of methods and tools that enable these transformations to be comprehensively and consistently described. However, unlike the ILD case considered earlier, in this example there is considerable heterogeneity in the types of data transformation processes that must be described, with some stages having clear and well developed ontologies and others not.

For example, as we discussed in section 3, in the case of questionnaire design there are a number of well established tools to enable the characterization of questionnaire wording, structure and routing using XML scheme. One example of such a tool is SuML. The SuML architecture was

designed by the Clinical Informatics Research Group at the University of Washington, Seattle. SuML is designed primarily for survey design, data collection and processing. SuML enables the encoding of questionnaire content and routing in XML. For example, the original purpose question in the NRTS could be encoded as follows:

```
<question name="Trip_Origin_Purpose">

        <content><b>. Where have you just come from? <xsl:value-of

                select="/sumlResultSet/question[@name='Interviewee-
                No']/answer/variable[@ident='Integer']/value"/>Home, Shopping, Normal
                workplace, Other workplace/meeting, Personal business, Visiting
                friends/relatives at home, Sport or entertainment, Other leisure activity,
                School/college/university (as student), School/college (as pupil), Taking
                someone to the airport station hotel, Meeting someone at the airport station
                hotel, Other?</b><br/> <b>(select the appropriate button)</b>

        </content>

        <answer>

                <variable ident=" Integer" type="Integer">

                <values>

                        <value code="0">Home</value>

                        <value code="1">Shopping</value>

                        <value code="2">Normal workplace</value>

                        <value code="3">Other workplace/meeting</value>

                        <value code="4">Personal business</value>

                        <value code="5">Visiting friends/relatives at home</value>

                        <value code="6">Sport or entertainment</value>

                        <value code="7">Other leisure activity</value>

                        <value code="8">School/college/university (as student)</value>

                        <value code="9">School/college (as pupil)</value>

                        <value code="10">Taking someone to the airport station hotel</value>

                        <value code="11">Meeting someone at the airport station hotel</value>

                        <value code="12">Other</value>

                </values>

                </variable>

        </answer>

 </question>
```

It is similarly straightforward to represent questionnaire routeing and checking using SuML or similar XML-based survey representation tools.

More challenging but still in principle reasonably straightforward is the encoding of the processes of population and sample frame selection and the definition of the sampling design, although we are currently unaware of any existing tools that provide this specific functionality.

However, much greater challenges are posed by the need to encode processes of survey implementation and survey data pre-processing. For survey data pre-processing, the challenge arises from the need to provide a generalised representation of the diverse and elaborate processes of checking, imputation and weighting typically undertaking in the processing of raw survey data. These processes are usually a mixture of formal mathematical and statistical procedures and rule based procedures based on experience and established practice. Although initiatives such as the StatML system developed by Westlake and Krishnan (2006) represent an important step towards tools providing this functionality, there remains a considerable amount of work to be done to develop XML and semantic web compliant ontologies and vocabularies for this area and regard this as being an important research priority.

Similarly, the details of survey implementation usually involve enormous diversity and context-dependence and if anything, pose even greater challenges to systematic codification and description than does survey pre-processing. Yet the sound understanding and interpretation of survey data can demand that analysts have access to this information. Accordingly, we also regard this as an important area for further work.

5. CONCLUSIONS

This paper has reviewed a wide range of concepts of data quality from the transport domain and from a number of cognate domains and has distilled from this literature a new concept – the data supply chain, which builds on but generalises the existing literature. This concept refers to the sequence of processes of abstraction, transformation and computation that take place as one moves from a particular real world concept of interest to its measured counterpart residing in a particular dataset. This concept is designed to support the intelligent analysts seeking to establish the relevance and fitness for use of a particular datasets relative to a specific analysis objective

We have reviewed existing metadata tools and identified aspect of the data supply chain concept which can currently be treated well and those that can not. Priority areas for further work have been identified.

REFERENCES

Amrhein, C. G., and Schut, P. (1990). "Data quality standards and geographic information systems." *Proc., National Conf. "GIS for the 1990's."* Canadian Institute of Surveying and Mapping, Canada, 918–930.

ASTM Standard Guide E2259-03a (2006) *ASTM Standard Guide for Archiving and Retrieving ITS-Generated Data* ASTM International, West Conshohocken, PA. www.astm.org.

Axhausen K.W. (2000) Presenting and Preserving Travel Data, **in** Jones, P.M. and P.R. Stropher (eds.) *Transport Surveys: Raising the Standards*, Transportation research Circular, E-C008, II-F-1 – II-F-19.

Axhausen, K.W. and M.R. Wigan (2003) 'Public use of travel surveys: The metadata perspective, **in** Stropher, P.R. and P.M.Jones, *Transport Survey Quality and Innovation*, Elsevier, England.

Brackstone, G. (1999) 'Managing data quality in a statistical agency', *Survey Methodology* **25**(2) pp. 139-149.

Brown P, Sonksen P., (2002) 'Evaluation Of The Quality Of Information Retrieval Of Clinical Findings From A Computerized Patient Database Using A Semantic Terminological Model.' *Yearbook of Medical Informatics* 340–51.

Cambridge Systematic, Inc. (1996) *Travel Survey Manual*, Department of Transportation and Environmental Protection Agency, Washington DC.

Carson, C.S. (2000) 'What is data quality? A distillation of experience' 9th Meeting of the Heads of National Statistical Offices of East Asian Countries, Japan.

Carson, C.S., (2001) 'Toward a Framework for Assessing Data Quality," IMF Working Paper 01/25, International Monetary Fund, Washington, DC.

Carson, C.S., and Liuksila, C. (2002) 'Further steps toward a framework for assessing data quality', paper presented to the International Conference on Quality in Official Statistics, Stockholm, Sweden.

Chalasani, V.S., S. Schönfelder and K.W. Axhausen.(2002) 'Archiving travel data: The Mobidrive example', *Arbeitsbericht Verkehrs- und Raumplanung*, **129**, Institut für Verkehrsplanung und Transportsysteme (IVT), ETH Zürich, Zürich.

Chen, L. and May, A. (1987), 'Traffic detector errors and diagnostics', *Transportation Research Record* **1132**, 82–93.

Cherrett, T., Bell, H. & McDonald, M. (2000), 'Traffic management parameters from single inductive loop detectors', *Transportation Research Record* **1719**, 112–120.

DCMI (2001) Dublin Core Metadata Element Set. DCMI- Libraies Application Profile Working Group http://dublincore.org/documents/2001/08/08/library-application-profile

CIHI (2005) *The CIHI Data Quality Framework*, Canadian Institute for Health Information, Ottawa.

Crito, C.F. and Norwood, J.L. (eds.) (1997) *The Bureau of Transport Statistics: Priorities for the Future*, Transportation Research Board, National Academy of Sciences, Washington DC.

CSDGM http://www.sdvs.uwyo.edu/metadata/vocab.html#CSDGM#CSDGM.

Department for Transport and Transport Scotland (2007) *National Rail Travel Survey: Provisional 2007 Survey Report*, Department for Transport.

d'Onofrio M., and Gendron M. (2001) 'Data quality in the healthcare industry', *Journal of Data Quality* **7** (1)

Data Documentation Initiative (DDI) (2001). http://www.icpsr.umich.edu/DDI/intro.html

Department of Commerce (1992) Spatial Data Transfer Standard (SDTS) (Federal Information Processing Standard 173), Department of Commerce, National Institute of Standards and Technology, Washington DC.

Eppler M. and Wittig D. (2000) Conceptualizing information quality: a review of information quality frameworks from the last ten years. Proceedings of the 2000 Conference on Information Quality.

Eurostat (2002a) *Definition of Quality in Statistics*, Document n° Eurostat/A4/ Quality/02/General/ Definition.

Eurostat (2002b), *Standard Quality Report*, Document n° Eurostat/A4/Quality/ 02/Standard/Report.

Froeschl, K.A., Grossmann, W. Del Vecchio, V. (2003) *The Concept of Statistical Metadata*, METANET Deliverable D3 to the European Commission, Centre for Educational Sociology, University of Edinburgh

Grünewald, W. and Linden, H. (2001), Quality measurement - Eurostat experiences, Symposium on Achieving Data Quality in a Statistical Agency. Statistics Canada, Ottawa.

Harris, R.P. and Woolf, S.H. (2001) 'Current methods of the US Preventive Services Task Force: A review of the process', *American Journal of Preventative Medicine* **20** pp 21-35.

Head, J. R. (1982), 'A new specification for inductive loop detectors', *Traffic Engineering + Control* **23**, 186–199.

Holt, T. and Jones, T. (1998) 'Quality work and conflicting quality objectives," paper presented at the 84th DGINS Conference, Stockholm.

Hu, P., Boundy, B., Truett, T., Chang, E. and Gordon, S. (2002) *Archive and Use of ITS-Generated Data*, Report to the US Federal Highway Administration, Center for Transportation Analysis, Oak Ridge National Laboratory.

Hu, P., Goeltz, R. & Scmoyer, R. (2001), 'Proof of concept of ITS as an alternative data resource. A demonstration project in Florida and New York data', FHWA ORNL/TM-2001/247, Final Report.

Hunter, G.J., Hope, S., Sadiq, Z., Boin,A,. Marinelli, M., Kealy A.N., Duckham, M,. and Corner, R.J. (2005) 'Next-generation research issues in spatial data quality', *Proceedings of SSC 2005 Spatial Intelligence, Innovation and Praxis: The national biennial Conference of the Spatial Sciences Institute,* September, 2005. Spatial Sciences Institute, Melbourne.

ISO (2002) ISO 19113:2002 Geographic information — Quality principles.

ISO (2003a) ISO 19114:2003 Geographic information — Quality evaluation procedures.

ISO (2003b). ISO 19115:2003 Geographic information — Metadata.

Jones, P.M. and Stopher, P.R. (2000). Transport Surveys: Raising the Standard. Transportation Research Circular E-C008, Transportation Research Board, USA.

Klein, L. and Kelley, M. R. (1996) 'Detection Technology for IVHS: Final Report', Federal Highway Administration FHWA-RD-95-100; 3B1C2021.

Kresse, W. and Fadaie, K., (2004) *ISO Standards for Geographic Information*, Springer, Berlin.

Krishnamoorthy, R. K. (2008) *Travel Time Estimation and Forecasting on Urban Roads*, PhD Dissertation, Centre for Transport Studies, Imperial College London.

Laliberté L. Grünewald, W., and Probs, P. (2004) Data Quality: A Comparison Of IMF's Data Quality Assessment Framework (DQAF) and Eurostat's Quality Definition.

Leitheiser, R.L. (2001) 'Data quality in health care data warehouse environments', Proceedings of the 34th Hawaii International Conference on System Sciences, Hawaii.

Levinson, D. and Zofka, E. (2006) 'The Metropolitan Travel Survey Archive: A Case Study in Archiving' in Stopher, P.R. and Stecher, C. Travel Survey Methods: Quality and Future Directions. Elsevier, Oxford.

Lohr KN (2004) 'Rating the strength of scientific evidence: Relevance for quality improvement programs', *International Journal of Qualitative Health Care* **16** pp 9-18.

Long J., Richards J., and Seko C. (2003) *The Canadian Institute for Health Information (CIHI) Data Quality Framework, Version 1. A Meta-Evaluation and Future Directions,* Canadian Institute for Health Information, Ottawa.

Malin, J.L. and Keating, M.L. (2005) 'The cost-quality trade-off: need for data quality standards for studies that impact clinical practice and health policy' *Journal Of Clinical Oncology* **23**(21) pp 4581-4584.

Mathieu, R.G. and Khalil, O. (1998) 'Data quality in the database systems course' Data Quality 4(1).

MVA, Birmingham City Council, Leicester City Council, Peek, Viggen Corporation (1998) *Input and Output Data Content and Quality,* UTMC 07/17 Deliverable 4. UK Department for Transport, www.utmc.gov.uk.

National Health Service (2002). NHS update: archive. http://www.nhs.uk, September 2002.

Network Social Sciences Tools and Resources (NESSTAR) (undated) www.nesstar.org

O'Looney, J. (2001) *Beyond Maps: GIS and Decision Making in Local Government*. Environmental Systems Research Institute, Inc., Redlands, California.

Orfanidis L., Bamidis, P.D. and Eaglestone, B. (2004) 'Data quality Issues in electronic health records: An adaptation framework for the Greek health system', Health Informatics 10(1) pp 23-36.

Richardson, A.J., E.S. Ampt and A.H. Meyburg (1995) *Survey Methods for Transport Planning*, Eucalyptus Press, Melbourne.

Robinson, S. (2005) 'The Development and Application of an Urban Link Travel Time Model Using Data Derived from Inductive Loop Detectors', PhD Dissertation, Centre for Transport Studies, Imperial College London.

Robinson, S. and Polak, J. W. (2006) 'ILD data cleaning treatments and their effect on the performance of Urban Link Travel Time models', Proceedings of the 85th Annual Meeting of the Transportation Research Board, Washington DC.

Stopher, P.R. and Jones, P.M. (2003). Transport Survey Quality and Innovation. Pergamon Press, New York.

Stopher, P. R., and Metcalf, H. M. A. (1996) *NCHRP Synthesis of Highway Practice 236: Methods for Household Travel Surveys.* TRB, National Research Council, Washington, D.C.

Stopher, P.R. and Stecher, C. (2006) *Travel Survey Methods: Quality and Future Directions*. Elsevier, Oxford.

Tayi, G. K., and Ballou, D. P. (1998) 'Examining data quality', *Communications of the ACM* **41**(2) pp 54-57.

Turner, S. (2002) *Defining and Measuring Traffic Data Quality*, Traffic Data Quality Workshop, Report prepared for the Office of Policy, Federal Highway Administration, Washington, DC.

Veregin, H. and Hargitai, P. (1995) 'An evaluation matrix for geographical data quality' **in** Guptill, S. C., and Morrison, J. (eds.) (1995). *The Elements of Spatial Data Quality*, Elsevier, Amsterdam.

Veregin, H. (1999) 'Data Quality Parameters' **in** Longley, P.A., Goodchild, M.F., Rhind D. and Maguire, D. (eds.) *Geographical Information Systems, Principles and Applications – Volume 1, Principles and Technical Issues*, John Wiley & Sons, London.

van Oort, P. (2005) *Spatial Data Quality: From Description to Application*, Netherlands Geodetic Commission, Delft.

Walmsley, J. (1982), 'The practical implementation of SCOOT traffic control systems', *Traffic Engineering + Control* **23**, 196–199.

Wand Y, and Wang, R.Y. (1996) 'Anchoring data quality dimensions ontological foundations' *Communications of the ACM* **39**(11) pp 86-95.

Wang, R., Kon, H., and Madnick, S. (1993) 'Data quality requirements analysis and modelling', Paper presented at the Ninth International Conference of Data Engineering, Vienna, Austria.

Wang, R.Y., Storey, V.C., and Firth, C.P. (1995) A framework for analysis of data quality research. *IEEE Transaction on Knowledge and Data Engineering* **7**(4) pp. 623–640

Wang, R.Y., and Strong, D.M. (1996) 'Beyond accuracy: What data quality means to data consumer," *Journal of Management Information Systems*, **12**(1), 5-34.

West, S.L., King, V., and Carey, T., (2002) *Systems to Rate the Strength of Scientific Evidence*. Agency for Healthcare Research and Quality, Rockville, MD.

Westlake, A.J. (2005) *Development of Meta Data Methods,* OPUS project Deliverable D3.1 to the European Commission, Centre for Transport Studies, Imperial College London.

Westlake, A and Krishnan, R (2006), *Generic Structures and Functionality for Support of Statistical Models in Statistical Databases – Implementation Report on Using Information from Statistical Models*, OPUS project Deliverable D6.2 to the European Commission, Centre for Transport Studies, Imperial College London.

Wigan, M.R. (2001) 'Enabling and managing greater access to transport data through metadata', paper presented at the *European Transport Conference*, Cambridge.

APPENDIX V: BUILDING A COMMUNITY, DELEGATES AT MEETINGS IN AUTUMN 2006

September 2006 meeting

Steve George Centro/MATISSE

Rory Bauer Norwich Union

Ian Nunney Norwich Union

Mary Gosden HA

Nick Ilsley Transportdirect

Colin Waugh Thales Facilitator 3

Ian Legg One/First

David Hytch Logica CMG

Peter Zieminski British Transport Police

Ian Hawthorne DfT

Ray Browne DTI

Ian Curran 02

Jonathan Mosedale DfT

Neal Skelton ITS (UK)

Nigel Wall Shadowcreek, L&T KTN etc

Chris Marsden Rand-Europe

Billy Denyer ORR

Martin Ballard Mobile Data Association

Graham Cattell BP

KK Ho RBS

Sean Perry-Evans SERCO

Jo Fereday Siemens

Fred Pink Principal facilitator

Peter Landshoff (CMI)

John Patman (CMI)

Mark Hayes (Cam)

Michael Simmons (Cam)

Amine Tafat (Cam)

Hongqian Liang (Imp)

John Polak (Imp)

Andy Parker (Cam)

Vic Rayward-Smith (University of East Anglia)

December 8th 2006 meeting

Jeremy Acklam ATOS Origin

Dr Christopher Barnes Trafficmaster plc

Chris Bowen O2 plc

James Bradley BAA plc

Graham Cattell BP International

Prof Brian Collins, Chief Scientific Adviser, Department for Transport

Zhan Cui BT Group plc

Ian Curran O2

Dr David Eyers University of Cambridge

Jo Fereday Siemens AG

Richard Forster Autonomy Corporation

Charles Gill Lockheed Martin

Mary Gosden Highways Agency

Ian Hamilton BT Group plc

Paul Harrison Home Office

Ian Hawthorne Department for Transport (DfT)

Mark Hayes Cambridge eScience Centre

David Hytch LogicaCMG plc

Nick Illsley Department for Transport (DfT)

Nick Knowles Kizoom

Prof Peter Landshoff Cambridge-MIT Institute (CMI)

Jon Maybury Department of Trade and Industry (DTI)

Richard S. Mills Boeing

Dr Jonathan Mosedale Department for Transport (DfT)

Dr Janko Mrsic-Flogel London Internet Centre

Ian Nunney Norwich Union

Ed Parsons Ordnance Survey

Caryll Paterson Royal Bank of Scotland

Jan Pinkerton East of England Development Agency (EEDA)

Prof John Polak Imperial College London

Kirsty Richardson BAA plc

Kathleen Salyan Lockheed Martin STASYS Ltd

Tom Scampion Deloitte & Touche LLP

Michael Simmons University of Cambridge