OPUS

Optimising the use of Partial information in Urban and regional Systems

Project IST-2001-32471

WP3: Development of Metadata Methods

Title : **Specifications for the Extension of the LATS Database System for the Transport Domain**

Creator (Author):	Andrew Westlake	Survey & Statistical Computing
Contributor :	Saikumar Chalisani	ETH
	Mike Collop	Transport *for* London
	Miles Logie	Minnerva

Identifier :	Deliverable D3.2 of project IST-2001-32471
Status :	Draft version of programme deliverable
Type :	Report
Version :	1.2 – Final deliverable
Date.Created :	29 April 2005
Date.Modified :	17 May 2005
Date.Next version due :	20 May 2005
Submission Date :	
Subject.Category	
Subject.Keyword	
Source	
Relation	This header section draws on the e-GMS structure for document metadata developed by the UK e-Gov initiative.
Rights.Copyright	The Opus Project
Contract Date :	April 2003
Publisher (Project Coordinator) :	Imperial College London
Contact Person :	John Polak

Address :	Centre for Transport Studies Department of Civil and Environmental Engineering Imperial College London South Kensington campus London SW7 2AZ United Kingdom
Telephone :	+44-(0)20-7594.6089
Fax :	+44-(0)20-7594.6102
e-mail :	j.polak@imperial.ac.uk
Consortium :	CTS, TFL, KATALYSIS, ETHZ, FUNDP, PTV, SYSTEMATICA, WHO. MINNERVA, SURVEY & STATISTICAL COMPUTING, OXFORD SYSTEMATICS

TABLE OF CONTENTS

TECHNICAL ABSTRACT

This deliverable D3.2 is a result of Work Package WP03 of the OPUS project. Work Package WP03 has as title: "Development of Metadata Methods". The specification of this Work Package from the project proposal is as follows.

Objectives

To extend current initiatives on statistical metadata (particularly metadata for statistical processes) to include the representation of statistical models, both in their underlying structure and as used for particular analyses or data syntheses.

In particular, to provide a representation of algebraic model forms as an extension of existing expression representations, records of input data sources (including their versions), parametric assumptions used in model invocations (including distribution (Bayesian) assumptions for parameters) and links to synthesised information from model invocations. It will also draw on work for generic version control and audit trails.

Description of work

The work package will draw on existing (and ongoing) work by others on statistical metadata (particularly on process metadata), and on the related implementation activities for metadata within the LATS transport database project being undertaken in London. It will also draw on concurrent work in WP2 to identify the form of models to be considered. The work will comprise three interrelated activities.

1. *The development of a representation of the underlying form of statistical models as metadata, in a way that is accessible for review by people and execution by software.*
   This will include algebraic expressions and relationships, distributions, variables and parameters, and will not assume that the model can be represented as a single component.
2. *The development of a representation of the way in which a model is used in the context of a statistical database.*
   This will include recording the input information sources used, together with prior settings or assumptions about parameters (including vague assumptions in the form of distributions and dependencies). The representation should allow the use of a model to be reviewed, revised and re-executed. Issues of version control will need to be addressed, since in general the input sources will be dynamically updated.
3. *The development of a representation of the results of using a model.*
   This will include estimated parameter values (with suitable posterior support or precision information), and also synthesised information generated from the fitted model. The main body of the latter will be stored in standard structures (as for real data), but it is essential to retain the link to the model and the generation process. Issues of version control will arise again.

Deliverable D3.1 discussed general issues about the nature of statistical modelling and about statistical metadata and developed a proposal for a (metadata) structure for the representation of statistical modelling (including the model calibration process). The current document discusses the requirements for and implementation of a system for providing user access to information about the provenance and reliability of results from models that have been calibrated using the Opus Methodology.

The Opus Methodology is Bayesian, so all the information about a model lies in the model specification plus the posterior distributions of the parameters. In theory it is sufficient to present just this information to users. In practice, this will be too complex or impenetrable for most users, so, as with most statistical analyses, other forms of interpretation and presentation will be needed.

We anticipate four forms of presentation of information derived from a model.

1. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.
2. Presentation of the posterior distributions of the parameters. This can be done in terms of summary statistics (particularly means and standard deviations) of the posterior distributions, or of complete distributions as histograms or multivariate contour plots. Population parameters of direct interest to users (for example, in decision making) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process.
3. Information about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.
4. Synthetic data. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.
   The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences, and the solution will draw on the previous three types of information.

In this report we elaborate these presentations, and discuss how they can be implemented alongside (or within) the Nesstar architecture. The details of the structures and functionality, and their implementation, will be the concern of work package 6.

EXECUTIVE SUMMARY

This document is Deliverable D3.2 of the Fifth-Framework project (FP5) OPUS. The OPUS project aims to develop and demonstrate statistically sound methods of combining datasets that each provide partial information on a single complex of underlying variables.

The expected practical result of application of the OPUS methodology is a calibrated probabilistic model of the problem domain at hand, with which it is possible to calculate the most likely values of missing, unobserved, or unobservable quantities of the object system under study, with potentially important savings of time and resources.

This report discusses the requirements for using the results of the Opus Methodology in the test cases in London and Zurich. It has different objectives from those originally conceived, because the nature of the systems that will be used for the test cases has changed, particularly in London.

After discussing the current context for the test cases, the report addresses the issue of providing information to support the use of the results from applying the Opus Methodology to the test cases. The task is to provide the user with information to inform and support their use of results obtained from a calibrated model. This focuses on information about the provenance of the model, that is, how it was constructed and how it was calibrated, and the reliability of the estimates obtained from it, which relates to the posterior distributions of the parameters of the model.

The tasks to be addressed are discussed, and a broad vision for the nature and architecture for the solution is presented. The detailed specifications and implementation will follow in work package 6.

1. INTRODUCTION AND FRAMEWORK

1.1 About OPUS

1.1.1 Background

OPUS is a large information management research project, supported by Eurostat as part of the European Commission's Information Society Technologies (IST) Programme. The overall aim of the OPUS project is to enable the coherent combination and use of data from disparate, cross-sectoral sources, and so contribute to improved decision making in the public and private sector within Europe. The research is focused on developing an innovative methodology, incorporating statistical and database systems. Transport planning is a prominent example of a topic that uses multiple sources of data, and will be the main test case for OPUS, but the cross-sectoral nature of the research will be demonstrated through the inclusion of an application in the field of health information as another example.

To meet the needs for comprehensive information on socio-economic systems such as urban and regional transport planning, and in the health services sector, data from diverse sources (e.g. conventional sample surveys, census records, operational data streams and data generated by IST systems themselves) must be combined. There is currently no appropriate developed methodology that enables the combination of complex spatial, temporal and real time data in a statistically coherent fashion. The aim of the project is to develop, apply and evaluate such a methodology. OPUS will develop a general statistical framework for combining diverse data sources and specialise this framework to estimate indicators of mobility such as travel patterns over space and time for different groups of people. The project will undertake pilot and feasibility study applications in London, Zurich, Milan, and on a national level in Belgium. Methods for extending the framework to information aspects of the health domain will also be investigated.

The benefits of OPUS will be:

- Improved estimation of detailed travel demand, using all available information;
- Avoidance of simplified combination of data that can give erroneous estimates;
- Indicators of data quality, to provide guidance for new data collection;
- A framework for managing data from rolling survey programmes;
- Better understanding of the role of variability and uncertainty in results and models;
- Avoidance of confusion from different, apparently conflicting, estimates of the same quantity;
- A generalised methodology for other domains of interest.

The participants in the OPUS project are as follows:

Research Organisations

- CTS (Centre for Transport Studies, Department of Civil and Environmental Engineering, Imperial College London), United Kingdom – Lead Partner
- DEPH (Department of Epidemiology and Public Health, Imperial College London), United Kingdom
- ETHZ (Institut für Verkehrsplanung, Transporttechnik, Strassen- und Eisenbahnbau), Switzerland
- FUNDP, Transport Research Group (Facultés Universitaires Notre-Dame de la Paix), Belgium

Practitioners

- Minnerva Ltd., United Kingdom.
- Survey and Statistical Computing, United Kingdom.
- Katalysis Ltd., United Kingdom.
- PTV AG, Germany
- Systematica, Italy.
- Oxford Systematics, Australia: Peer Reviewer

Public Bodies

- Transport for London (TfL), United Kingdom.
- World Health Organisation (WHO), Italy.

1.1.2 Objectives of the OPUS project

To meet the needs for comprehensive information on socio-economic systems such as urban and regional transport planning, and in the health services sector, data from diverse sources (e.g. conventional sample surveys, census records, operational data streams and data generated by IST systems themselves) must be *combined*. There is currently no appropriate developed methodology that enables the combination of complex spatial, temporal and real time data in a statistically coherent fashion.

The overall aim of the proposed project is to develop, apply and evaluate such methodologies, taking as a specific case study the transport planning sector. The specific objectives of the study are:

- To develop a generic statistical framework to enable the optimal combination of complex spatial and temporal data from survey and non-survey sources. This framework will specify how to optimally estimate the underlying population parameters of interest taking into account the structural relationships between the different measured data quantities and the sampling and non-sampling errors associated with the respective data collection processes. It is envisaged that the framework will be broadly Bayesian in nature. The framework will make no specific assumptions regarding the particular structural and sampling/non-sampling errors and will thus be relevant to a wide range of application domains.
- To apply the generic framework within the field of urban and regional transport planning. This will involve the definition of specific structural relationships amongst measured quantities and the characterisation of sampling/non-sampling errors, based on domain knowledge from the field of transport planning.

- To develop the necessary database and estimation software to enable the application of the statistical framework in a number of case study areas.
- To undertake a major pilot application study in London, focusing on the derivation of indicators of the mobility and the performance of transport policy measures.
- In parallel, to investigate the feasibility of applying the framework and methodologies developed both in other transport planning contexts and in other proximate domains, specifically environmental management and social statistics.
- Based on the experience gained in the pilot application and the feasibility studies, to evaluate the performance of the proposed methods and to define the scope and approach for wider applications in relevant domains including environmental management and health care.
- To disseminate the results to the relevant academic and practitioner communities.

1.1.3 Motivation

OPUS addresses the situation in which the analyst must combine data from a variety of different data sources to obtain a best estimate, or a fuller understanding, of a system. Such a situation can arise for a number of reasons including:

- No single source contains sufficient information by itself; or
- Multiple sources naturally arise (e.g. through observations at different levels of spatial or temporal aggregation or by means of different survey methods), resulting in a need to reconcile potentially conflicting estimations; or
- The need to update or transfer an existing set of data and parameter estimates when additional information becomes available.

Problems of combining data from different sources to produce consistent estimates of underlying population parameters arise in many fields of study including environmental monitoring, epidemiology and public health, earth observation, geographic information and navigation systems, transport and logistics, and economic and social statistics. Although the risks of using *ad hoc* combination rules and procedures are well understood, there are nevertheless many examples from practice in which just such approaches are still used. This reflects the fact that, although relatively straightforward methods exist for simple cases, there does not exist a coherent and well developed set of applicable methods capable of dealing with the full range of data combination problems, including factors such as:

- Data sources that provide both direct and indirect information on the relevant population parameters
- Data that are presented at different levels of aggregation
- Data sources with differing levels of statistical precision or user confidence
- Data that overlap, but that may provide different or conflicting information
- Gaps in the data observations
- The issues raised by the aging of sample survey data and the consequent need for updating
- Accommodating the updating sources
- The effect of sampling and non-sampling errors (including survey non-response and other sources of missing data)
- The opportunities presented by new data streams from IST systems

The key scientific objective of the project is to develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey and non-survey sources. The framework will be sufficiently abstract to be applicable to a wide range of potential domains.

Associated with this overall objective is the need for a suitable representation of the statistical metadata that is used for the specification and application of such a framework. That is the immediate objective of this report.

1.1.4 Subject areas

OPUS provides a generic approach but, in each case, it is necessary to make this approach specific to the particular area of interest (whether the area is geographical or topical in nature). A particular test-bed is transport in London, but studies will be made for transport in Belgium, Switzerland, and Italy, as well as health studies.

1.2 OPUS Project Work Package WP3

This section summarises the specifications for this work package.

1.2.1 Objectives

To extend current initiatives on statistical metadata (particularly metadata for statistical processes) to include the representation of statistical models, both in their underlying structure and as used for particular analyses or data syntheses.

In particular, to provide

- a representation of algebraic model forms as an extension of existing expression representations,
- records of input data sources (including their versions),
- parametric assumptions used in model invocations (including distribution (Bayesian) assumptions for parameters) and
- links to synthesised information from model invocations.

It will also draw on work for generic version control and audit trails.

1.2.2 Description of work

The work package will draw on existing (and ongoing) work by others on statistical metadata (particularly on process metadata), and on the related implementation activities for metadata within the LATS transport database project being undertaken in London. It will also draw on concurrent work in WP2 to identify the form of models to be considered. The work will comprise three interrelated activities.

1. The development of a representation of the underlying form of statistical models as metadata, in a way that is accessible for review by people and execution by software.
   This will include algebraic expressions and relationships, distributions, variables and parameters, and will not assume that the model can be represented as a single component.

2. The development of a representation of the way in which a model is used in the context of a statistical database.
   This will include recording the input information sources used, together with prior settings or assumptions about parameters (including vague assumptions in the form of distributions and dependencies). The representation should allow the use of a model to be reviewed, revised and re-executed. Issues of version control will need to be addressed, since in general the input sources will be dynamically updated.
3. The development of a representation of the results of using a model.
   This will include estimated parameter values (with suitable posterior support or precision information), and also synthesised information generated from the fitted model. The main body of the latter will be stored in standard structures (as for real data), but it is essential to retain the link to the model and the generation process. Issues of version control will arise again.

1.2.3 Deliverables

D3.1 Proposals for Metadata for Generic Support of Statistical Modelling in Statistical Databases

D3.2 Specifications for the Extension of the LATS Database System for the Transport Domain

1.3 Objectives of Deliverable D3.2

The key scientific objective of the Opus project is to develop a generic statistical framework for the optimal combination of complex spatial and temporal data from survey and non-survey sources. The framework will be sufficiently abstract to be applicable to a wide range of potential domains.

Associated with this objective is the need for a suitable representation of the statistical metadata that is used for the specification and application of such a framework. That is the objective of deliverable D3.1. In the current report we discuss the needs and requirements of the LATS group for the enhancement of their statistical dissemination system to make use of the Opus methodology, and the implications of those requirements in terms of functionality and implementation strategy. This report acts as input to work package 6, which elaborates the work of WP3 and leads through to implementation.

1.4 Relation with the OPUS Life-cycle

The objective of the opus project is to develop a methodology for the integration of multiple datasets in complex situations. Work package three supports this, through the development of models for the data structures and processes that are needed to support the implementation of the generic methodology of the project. This work package is dependent on the project is a whole, and particularly on work package two, for specifying the types of methodology and statistical method that need to be covered as metadata. The development of methodology is not of itself dependent on work package three, though it is expected that consideration of data structures at an

abstract level will contribute to the thinking about the generic methodology. However, the impact of this work package is expected to be felt most in the implementation efforts that will be part of the trials that follow development of the generic methodology.

1.5 Structure of the Deliverable

Chapter 2 presents background information, describing the current situation with respect to the existing statistical dissemination systems in use in London (at T*f*L for the LATS data) and in Zürich and setting out the basic objectives for WP6. Chapter 3 describes the requirements, focussing on the support of the use of synthetic data. Chapter 4 presents initial ideas about the architecture that will be used to build a usable solution that can be implemented in WP6 and used in WP8 and 9.

2. BACKGROUND

2.1 The LATS Database environment

The proposal for the Opus project was written shortly after the completion of a design report for a Statistical Database for the results of the 2001 London Area Transport Survey (see [West01]). This report envisaged the construction of a specialised database system:

> *The LATS database system is intended to be a dynamic resource containing information about travel in London. It will contain information about demand, use and attitudes and will cover all modes of transport. It is complementary to various other databases about transport facilities in London, and will have facilities to co-operate with them.*
>
> *This document presents various issues relating to the design and use of such a database. It considers the objectives of the database, the use and users of the database, the main requirements for features and functionality (with considerable detail in some areas), and some technologies relevant to the implementation. It is the main output from a design study to investigate the architectural and functional characteristics required for the database.*

After a more detailed investigation T*f*L decided that the full implementation of a new system was neither economical nor practical, and instead have constructed a system (called *Romulus*) which is based on the dissemination package Nesstar[1]. This is the system that will be used to host suitable results produced by WP8 (the London Test Case), and so is the target system for WP6. WP9, the Zürich Test Case is also to be used with a system constructed using Nesstar, so the results of WP6 will be usable with both test cases.

2.2 Statistical Database Enhancement

Nesstar is not a partner in the Opus project. While the project has some aspirations that aspects of the Opus methodology might be integrated with Nesstar (and other, similar systems), this would be a matter for an exploitation phase subsequent to the completion of the Opus project.

We thus do not expect to integrate any parts of the Opus methodology directly into the existing statistical dissemination databases during the lifetime of the project (though it may happen later). In particular, we do not expect to implement any of the Opus model fitting methodology within the systems. Instead, the model fitting will be done externally. Datasets (and the related metadata) will be used from the statistical databases (possibly later by direct extraction, but certainly initially by taking copies), and new information obtained from the modelling will feed back into the database.

---

[1] The Nesstar product was developed under a number of EU framework projects, including the Nesstar and Faster projects. Commercial exploitation is being undertaken by Nesstar Ltd, a spin-off company hosted at the University of Essex – see www.nesstar.com.

Because any information that is fed back into the statistical system will be a result of the Opus methodology, we have decided to focus our efforts in WP6 on supporting the use and understanding of this information. This will be done alongside the existing statistical system, using the same technology where appropriate, and implemented in such a way as to appear as seamless as possible for the user of the statistical system. A specific objective of this enhancement will be to implement new facilities in a way that demonstrates their usefulness and facilitates any later implementation within the Nesstar system.

In consequence, this report is not a specification for the inclusion of modelling methodology in statistical databases, but is rather an anticipation of the work in WP6 on making the results of modelling available in a statistical database context. Its focus is on the broad requirements for making use of the results of the Opus methodology in association with those statistical databases, and on the overall architecture needed to support those requirements.

We refer to this problem area as *Provenance and Reliability* with the implications of this term being expanded in following sections.

2.3 Results from the Opus Methodology

The Opus methodology is Bayesian, so all the information about a model lies in the model specification plus the posterior distributions of the parameters. In theory it is sufficient to present just this information to users. In practice, this will be too complex or impenetrable for most users, so, as with most statistical analyses, other forms of interpretation and presentation will be needed.

We anticipate four forms of presentation of information derived from a model.

1. Summary reports which provide interpretations of the fitted model, based on the experience and judgement of the author. These will be largely textual, but will include illustrative material and links back to the model.
2. Presentation of the posterior distributions of the parameters. This can be done in terms of summary statistics (particularly means and standard deviations) of the posterior distributions, or of complete distributions as histograms or multivariate contour plots. Population parameters of direct interest to users (for example, in decision making) will be the primary focus, but these are generally dependent on internal (hyper-) parameters, which are the ones directly adjusted by the fitting process.
3. Information about the model fitting process (the audit trail). This includes information about the fitting methodology (which will apply across a set of related models), together with the datasets used at the various fitting stages and the contribution of each such stage to the final fit. The latter is particularly important in terms of understanding how well the posterior distributions of parameters have been determined by the fitting process.
4. Synthetic data. Given the model specification and the posterior distributions, it is possible to simulate observations on data subjects. In this way, we can create synthetic datasets which have the same characteristics as the model. These are much easier to analyse for people used to handling real datasets. It is also possible to

generate data for specific conditions, for example by limiting the impact of abnormal events, focussing on particular subsets of the overall possibilities, or assuming away some uncertainty in parameters.
The problem is that synthetic data is not real, and its statistical properties are not the same as those of real observations, because they come entirely from the fitted model. The challenge is to guide users to appreciate these differences, and the solution will draw on the previous three types of information.

In the following sections we elaborate these presentations, and discuss how they can be implemented alongside (or within) the Nesstar architecture.

2.4 Functionality in Nesstar

The Nesstar system manages user access to a distributed database of statistical data. It also provides basic statistical analysis facilities.

Statistical data in Nesstar is stored either as normal data records (referred to as Survey Data), or aggregated into multidimensional data cubes (called Summary Data). All datasets are described using the DDI Codebook metadata standard (see [DDI]), which covers both datasets and the variables within them. The metadata is stored in XML documents.

The Nesstar system uses an enhanced web-browsing interface to present information to the user. This uses specialised components to display specialised information, but provides the flexibility to display any information that can be formatted as or linked into an HTML document. The linking is important, as it is designed to provide access to information that is not stored within the Nesstar system, but can be viewed from within the Nesstar interface.

This functionality provides the basic building blocks for the extensions to be produced under WP6.

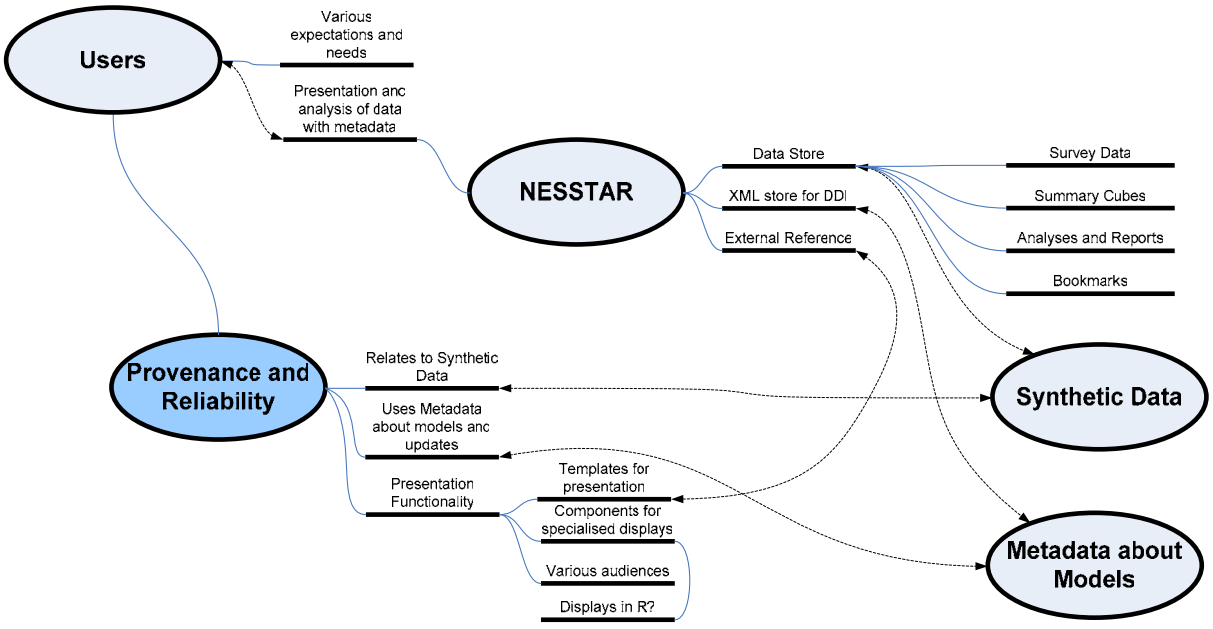
3. SYNTHETIC DATA

3.1 Overview

There is no proposal to extend either of the Nesstar implementations (in London and Zürich) to include Opus-style modelling within the dissemination systems, as was originally envisaged. However, there is the intention to use the Opus methodology outside the system, and to then transfer information back into the system, in the form of both interpreted conclusions and synthetic (simulated or enhanced) datasets. So, for the metadata and database inputs we propose to concentrate on the support of the use of the synthetic data. Some of this information can be input directly into the existing systems, and some will (at least initially) be accessible from a separate web-based system operating alongside the existing Nesstar services.

Because the synthetic data looks like real data, no special facilities are needed to add it to the existing systems, or to use it within them. However, synthetic data is different, and to use it effectively (and correctly) the user needs access to additional information (metadata) about the process by which the data was synthesised from the model, and about the form and quality of the model. The structure for metadata about statistical models already proposed in deliverable D3.1 includes (potentially) all this information (it is in effect a complete audit trail for all the specifications and stages used to produce the synthetic data), but not in a form that will be readily accessible to users. So we will need to find ways of presenting this additional information that are accessible and comprehensible for different groups of user.

The following diagram shows some of the elements involved, and how they link into existing functionality within the Nesstar system.



While it is clear that the synthetic datasets can be stored within the Nesstar data store, the diagram is not intended to imply that other components will (necessarily) be closely integrated with Nesstar. It is more a question of exploiting synergies.

Nesstar uses XML for the storage of the DDI metadata, and we will (almost certainly) also use XML for the model metadata. Thus we will be using the same technology to access and manipulate the metadata, but will probably not make any attempt at this stage to integrate the model metadata with the DDI in the Nesstar metadata store. However, given the characteristics of XML and DDI, this should not be too difficult to do at some later stage.

Similarly, the presentation functionality (to provide access to information from the metadata about the synthetic datasets) will be implemented using web facilities, but will be hosted separately and linked using the external reference mechanism in Nesstar.

3.2 Why is Synthetic Data Different?

A statistical model consists of a mathematical specification of relationships between variables, probability distributions that capture the variability in these variables, parameters that describe the distributions and (possibly) the relationships, and probability distributions that capture the uncertainty in our knowledge about the parameters. By sampling from the probability distributions for uncertainty and variability, and by following through the mathematics of the relationships, we can generate an observation on the set of variables (and on the parameters). If we repeat this process we can generate a set of (synthetic) data records.

With real data collection, the sample size is of central importance (along with the survey design), in that it determines the amount of information we collect about the system being observed. The same is true of synthetic data, but the relevance is different, because the system being observed then is the model, not the underlying reality. The sample size for synthetic data is arbitrary (or at least only limited by constraints of time and storage). A larger sample gives us better estimates about the parameters of the model, but that is not what is really of interest. What matters is the information **in the model** about the underlying reality, and that stays the same, whatever the size of the synthetic sample. We will use different criteria to determine the sample size to use. For example, we might choose to synthesise a complete set of all trips made in a particular time interval, or we may choose to synthesise a dataset that corresponds to a real survey dataset..

An estimate of a mean from the synthetic data will (probably) be an unbiased estimate for the mean of the posterior distribution of the parameter corresponding to this mean in the model. But the precision (standard error) of the synthetic mean will not tell us much about the precision with which the parameter is determined in the model. That information is in the model, and can be estimated from the synthetic data, but not with a naive approach.

In general, analysis of relationships in synthetic data will provide some insight into the mathematical structure of the model, in a way that may well be more approachable than tackling the mathematics in the model directly. But to understand precision and uncertainty we need to draw users into other ways of looking at the information in the model.

In the transport domain, data is often processed through complex (transport) models to derive measures of direct interest, and the synthetic data may be subjected to such

processing. It is difficult to conceive how to propagate information about model precision through such processes. However, if the measures derived are central to the domain, then they should be explicitly included in the model, and so information about them can be extracted directly, without the complex processing.

3.3 Issues

Users of synthetic data need access to additional metadata about the process by which the data were generated (an audit trail) and the quality of the models used, summarised in terms of the confidence that they can place in results obtained from the synthetic data. We thus face three issues.

1. Analysis of the information needed by different classes of user to support their use of synthetic data.
2. Design of suitable metadata structures to capture and hold this information.
3. Design and implementation of suitable presentation material and functionality, so that users with different interests and skills can access and use the information that they need.

A start has been made on point 2 with the metadata structure for models proposed in D3.1. This will receive further elaboration in D6.1, which will also address point 1. Point 3 is the subject of the remainder of WP6, but we have already indicated some general ideas for the architecture of this, and these are elaborated later.

3.4 Provenance and Reliability

3.4.1 Objectives

Users of synthetic data should properly be asking questions about how the data was generated and how much confidence they should have in conclusions drawn from it. We use the term *Provenance and Reliability* to refer to this area. This covers all issues to do with the understanding and interpretation of fitted models, not just those directly related to the use of synthetic data.

Different types of user will expect answers of different complexity and detail. Some answers can be generic, describing the philosophy behind the Opus methodology and Bayesian modelling, or showing (in UML diagrams?) the outline of the model fitting processes used. Other answers will need to be based on the specific components used in the model from which the data are synthesised, and further ones will make use of the detailed posterior information about the parameters. All this information will be available in the form of metadata, the top-level structure of which has been laid out in deliverable D3.1. The same information may need to be presented in different ways for different types of user. Not all reasonable questions will necessarily be amenable to being answered.

3.4.2 Model Form

For those interested in the specification of the models, we should be able to display various components at various levels of detail. This will extend from the top level

GAPM applicable to the model, right down to the details of the mathematics involved in the relationships, constraints and distributions in a particular model. Some of this should be shown in mathematical form, but graphical representations will be used wherever possible. For models that fit the Graphical Models framework, a display similar to the 'Doodle' system in WinBugs will be used.

3.4.3 Data used

The model metadata includes links to all the data used in reaching the final calibration of the parameters, so this can be shown, and the user can explore the DDI metadata for datasets already within the Nesstar system. The links between variables in datasets and those in the model are also available.

3.4.4 Parameters

The final model state includes information about all the posterior distributions, for the (hyper) parameters, for those induced for the parameters of direct interest and for the variables. These can be presented using standard displays of distributions, probably using the facilities in R.

Such displays show the precision with which parameters have been determined. The reliability and suitability of the model can be explored through the progress of the parameter distributions through the calibration processes.

3.4.5 Domain displays

It is expected that transport-related displays, such as the network and OD diagrams in Visum, can be enhanced to show some aspects of variability and classification, and they can be used to show appropriate parameter distributions, as well as distributions of actual flows. This will be explored as part of WP7.

3.4.6 Integration

Probable methods for integrating all these elements have been identified, but will need to be worked out in detail. Nesstar should be able to link across to new facilities, and R should provide the host for the integration.

4. ARCHITECTURE

Nesstar already provides a context in which statistical data and results can be viewed and manipulated. Associated with this can be metadata and interpretation, accessible both directly with the results and via hyperlinks. This functionality is implemented as a set of components which can be displayed in a browser interface (the Nesstar Explorer). A number of standard displays are available, which can be driven directly from the loaded data. Additional (more specialised) displays can be constructed using the components, and where a particular type of display is needed often, a template can be constructed and reused.

We will follow the same general structure for the implementation of the Provenance and Reliability functionality to be implemented in WP6.

Templates will provide the basic structure for all the different types of display of the different types of information from the model metadata. Templates will make use of components for the various types of specialised displays. Wherever possible, components will be sourced from existing functionality. For example, the R system has a comprehensive set of facilities for the display of uni- and multi-variate distributions, both formally and empirically determined. MathML can display mathematical formulae in a browser context, Visum can display traffic networks, and a SVG display component could be used for graphs.

The metadata for the fitted models will be generated (or otherwise captured) during the model fitting process. The full details of the structure are still to be elaborated, but the metadata storage will almost certainly be in XML documents. These can be manipulated using the standard set of XML-related technologies.

The R system (an Open Source implementation of the S language) can act as the host for integrating these various components. It is a scripting system that supports complex objects, and it is designed to communicate with other components, both as client and server.

REFERENCES

[DDI] Data Documentation Initiative, a standard for statistical metadata. See DDI Alliance, www.ddialliance.org.

[Fowl04] UML Distilled, 3rd Edition. ISBN: 0 321 19368 7. This is an excellent though terse guide to the content and use of UML, aimed at readers with some programming background.

[FWdV03] The Concept of Statistical Metadata (2003) by Froeschl, Grossmann, Del Vecchio, a deliverable from the MetaNet project, at www.epros.ed.ac.uk/metanet/deliverables/deliverables.html

[Karg03] Reference Models, developed by Reinhard Karge of Run Software under the MetaNet project, and maintained at www.run-software.com/downloads/documentation/ReferenceModel.doc

[MathML] Mathematical MarkUp Language. An XML language for the display and evaluation of mathematical expressions. See www.w3.org/Math.

[MetaNet] See www.epros.ed.ac.uk/metanet.

[R] The R project for Statistical Computing. See www.r-project.org/index.html.

[UML] See www.uml.org for information about UML 2.0. This is a standard developed under the auspices of the Object Management Group (www.omg.org).

[West01] LATS Database Design Study: Database Design Report, by Andrew Westlake. Available at www.sasc.co.uk/Projects/LATS DDR Public.pdf.

[West02] XML and Standards, by Andrew Westlake. Available at www.sasc.co.uk/Guides/XML%20and%20Standards.zip.

[West03] Models and Metadata, by Andrew Westlake. See www.sasc.co.uk/Guides/models%20and%20metadata.htm.

INDEX