Towards a future European Statistical Information System

Dave Sloggett[1], Chris de Vaney[2] and Steve Darroch[3]

[1] Centre for Coastal Zones Research, University of Dundee, DUNDEE DD1 4HN, UK daveslogget@compuserve.com
[2] Public Sector Division, Anite Systems, 13 rue Robert Stumper, L-2557 Luxembourg chris@anite-systems.lu
[3] ENVSTAT Project, Eurostat F3 (Environment), B4/066 Batiment Jean Monnet, L-2990 Luxembourg sdarroch@trier-online.de

Summary

Eurostat and the Member States of the European Union have decided to establish the European Statistical System (ESS) to co-ordinate, harmonize, and integrate their joint statistical effort. The ESS will operate in a de-centralized manner, emphasizing the work of the member states. This paper outlines the key features of the underlying metadata model behind a complete, generic, distributed and multi-lingual statistical information system capable of providing the necessary supporting informatics infrastructure for the ESS. The system is currently under development, and a proof-of-concept prototype will be demonstrated.

Keywords: Metadata, OMG, XML, CORBA, European Statistical System, distributed workflow

Introduction

Over the past ten years it has become obvious that statistical information systems, both official and commercial, need to integrate different statistical information sources and tool-sets to achieve a coherent framework for production, reference and dissemination activities.

In addition, globalisation has led to an increasing need for official statistics at a supra-national level. The information revolution has raised the perception among both the general public and policy makers that statistical data and supporting information on an increasing range of topics is, or should be, readily available. To meet this demand, national statistical institutes need to work closely together.

This paper describes the metadata model behind a proof-of-concept prototype designed to meet two objectives:

- *To promote integration between the various processes in the statistical process chain*
- *To promote harmonisation of metadata constructs between statistical organisations at the regional, national and international level in order to facilitate the flow of statistical information*

Meeting these aims offers significant benefits to statistical institutes:

- *Greater sales opportunities*
- *Reduced time to market*
- *Improved quality*
- *Greater transparency and traceability*
- *Reduced IT costs*

Integration of metadata across the statistical process chain

Within most statistical institutions the whole statistical process is divided into a series of discrete processes. Some of these may have been automated through the use of software tools such as Blaise, SAS, Excel etc. Often, metadata is used either to drive these processes or is created as a result of them. Typically metadata is held on paper, in separate databases or simply in experts' heads. The difficulty is to ensure that data and metadata are kept in step at each stage in the chain.

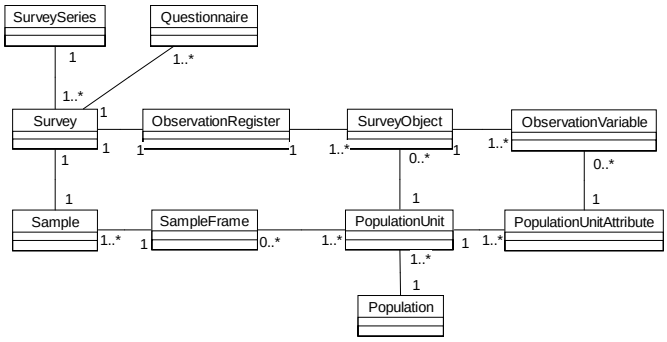
Distributed Workflow

Within the international context, statistical data typically flows from the regional to the national and finally to the international level. Metadata in contrast is typically created at the national level, sometimes with reference to the international level, and flows down to the local level where corresponding data is captured. Summaries and/or aggregates then flow back up to the international level with the metadata.

Efforts have been made to automate this process, notably centered on EDIFACT standards such as GESMES, CLASET and RDRMES, or architectures such as Lotus Notes. GESMES, CLASET and other EDIFACT standards provide excellent data models for statistical data exchange. However, difficulties in

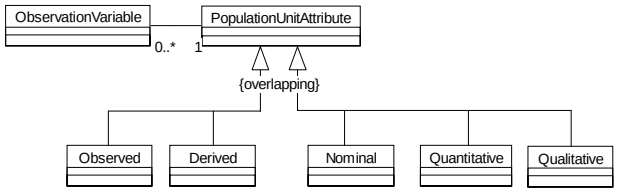
integrating these messages with legacy software have hindered their more widespread adoption by statistical agencies.
Recent advances in Internet protocols, notably XML and CORBA, offer the opportunity to address these difficulties. The standards setting body for CORBA is the Object Management Group (OMG) who have recently established a Statistical Interest Group. This group issued a request for information (RFI), closing during October 1998. A closing date for an RFP (Request for Proposals) is currently scheduled for the first half of 1999.
The metadata model described below is intended as an initial step in this direction.

The metadata model
The main features of the model, showing a simplified view of the process chain from survey to dissemination, are presented below. Many features, necessary for the output to be statistically valid, are omitted for the sake of clarity. Details of the full model can be found at http://www.anite-systems.lu/dmwce/ESIS
First of all we focus on surveys:
Each survey is performed on a sample which may be stratified according to one or

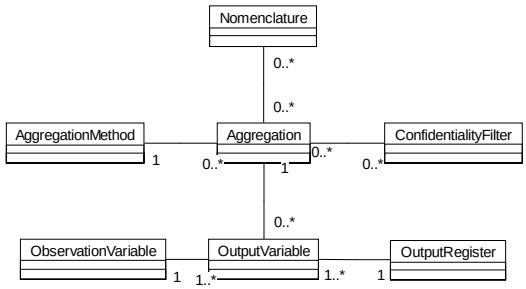


more criteria. The sample is selected from the sample frame, which is a specified list of population units drawn from a population. The selection may be undertaken on more than one criterion. For example, the population may be the residents of a town. The sample frame may be households in a particular area. The sample may be chosen at random from the sample frame.
Each survey builds an observation register. An observation register consists of respondent data on a number of survey objects; for example a census might collect data on households, individuals and their jobs. Household, person and job are survey objects.
Survey objects also allow for extended observations over defined populations. A survey object refers to a population unit, which may be used in other surveys. Population units possess attributes and may have relationships to other population units.
Each population unit posesses a set of population unit attributes. In the example given above, height, weight, age, name, address, hair colour, eye shade, income

and bank balance might be attributes of person; Number of persons resident, housing costs and household amenities may be attributes of household, while duration, permanency, income, tax paid and social security liability may be attributes of job.
A population unit attribute may be nominal, qualitative or quantitative. In addition, it may be observed, i.e. data is recorded directly from the respondents, or



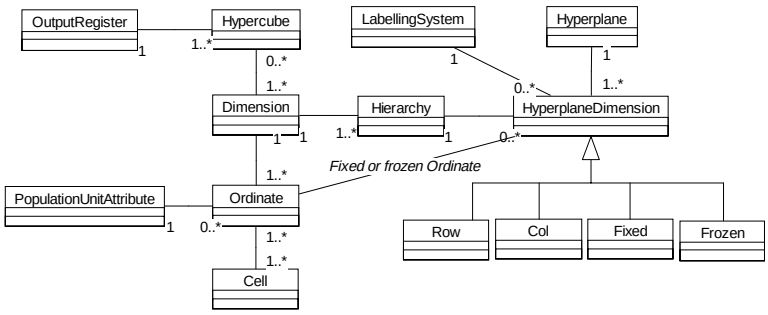
derived where it can be calculated from other qualitative or quantitative variables. Derived attributes may be qualitative or quantitative.
Each observation variable represents observations of a population unit attribute. An attribute may be observed in many surveys. An observation variable takes the attributes from its related population unit attribute. Derived variables at the unit record level may be present in the observation register.
An output register consists of a set of output variables. Each output variable is derived from one or more variables in the observation register. An observation



variable may be aggregated in different ways to create more than one output variable, according to the cross-classification and aggregation technique used.
The aggregation process consists of four stages:

1. The observation register is treated for errors arising during the survey process – non-responses, partial responses, human errors etc.
2. Records in the observation register having the same values for all qualitative variables are aggregated to create a record in the output register. The aggregation technique may vary from output variable to output variable, depending on the aggregation specification.
3. Records are aggregated up the levels in the related nomenclatures.
4. Finally confidentiality filters are applied to selectively restrict access to the aggregated data where necessary.

Each quantitative variable in the output register can be thought of as a matrix or hypercube. The qualitative variables form the dimensions in the hypercube. The



possible values of the qualitative variables, based on the related nomenclature(s), comprise the ordinates in the hypercube. Output tables are defined as hyperplanes, which in turn are selections from hypercubes. An output table is specified by defining the position of each dimension of the hypercube (row, col, fixed or frozen), along with the hierarchy (selection of ordinates) and system of labelling. The difference between fixed and frozen dimensions is that fixed dimensions may be dragged to become rows or columns in a multi-dimensional browser, while frozen dimensions are permanently fixed. This allows for hypercubes of very high dimensionality to be browsed

Nomenclatures form an important class of metadata covered by the model, and are used within all the statistical activities. The model provides both a representation framework and relationships to classifications from which the nomenclatures may be derived. Nomenclatures with varying levels of decomposition, inheritance between nomenclatures, correspondence tables and periods of validity are all fully supported

Distributed operation is achieved by connecting independent systems based on this model together in the form of a directed acyclical graph, where each node represents a logical server. Each node in the graph has a unique identifier. Once introduced into or produced by the system, metadata flows down the arcs of the graph (e.g. from international to national or from national to regional institutions)

Data is captured according to the metadata, and flows up the arcs with its associated metadata. When data or metadata are introduced into the system, they are assigned locally unique identifiers. When they leave the local system they acquire globally unique identifiers by combining their local identifier with the node's identifier.

An initial version of the model is to be submitted to the Object Management Group (OMG), for evaluation as a possible Internet standard for the exchange of statistical data and metadata. It is intended that the model supports distributed operation through a CORBA interface, XML messaging and relational snapshots. It is intended to ensure the widespread adoption of this standard by distributing C++ libraries, Java Beans, Delphi & Visual Basic components.

Full details of the model can be accessed on http://www.anite-systems.lu/dmwce/ESIS. Progress on the OMG standardisation process can be monitored at http://www.omg.org/library/schedule/Statistics_RFI.htm

Implementation

A prototype relational database implementation is currently under development, and will be demonstrated A statistical metadata repository connects all stages of the life cycle and maintains the relationship between data and metadata. Thus metadata established at the nomenclature management stage may be retained through survey management, validation, aggregation, reference, analysis and onto dissemination.

The system is generic in that it may be applied to all statistical domains The metadata repository supports domain frames – complete metadata data-sets specific to a particular statistical domain. It acts as a metadata reference server. A CORBA interface allows external systems to query metadata constructs (e.g. Nomenclatures, Definitions, methodologies etc) across the Internet. An XML interface serves a similar purpose.

The system is based on a client/server architecture. A number of systems may be connected together to support the interchange of data between administrations in line with the IDA programme Other data-transfer mechanisms include proprietary file formats including Excel spreadsheets, 'Stub' systems act as collection points for statistical data from external systems.

The system is built from open and fully documented sub-components. In this way interfaces to existing legacy systems may be quickly and easily developed. The sub-components may also be used to enhance or extend such legacy systems.

As a related initiative an XML browser and parser are also under development, with extensions for the representation of statistical metadata objects structured according to the model.

Conclusion

The above model promotes the integration of metadata across the statistical process chain and distributed workflow across regional, national and international boundaries. It supports both CORBA interfaces and XML messages allowing both push and pull dissemination paradigms. It offers significant benefits to statistical institutes:

- Greater sales opportunities
- Improved timeliness
- Improved quality
- Greater transparency & traceability

Further information may be had from the authors or from http://www.anite-systems.lu/dmwce/ESIS