

Architecture and Design of a Flexible Integrated Information System for Official Statistics Surveys, based on Structural Survey Metadata

Gregory E. FARMAKIS¹, Yorgos KAPETANAKIS², George A. PETRAKOS¹,
Michalis A. PETRAKOS¹

¹ *Liaison Systems S.A.*

Akadimias 77 - Athens - 106 77 GR

e-mail: {FirstName.LastName}@liaison.gr

² *Unisoft S.A.*

Posidonos 18 - Kallithea - 176 74 GR

e-mail: captain@statistics.gr

Abstract: This paper presents the architectural concepts and the design methodology developed by the authors and used for the actual implementation of the new integrated information system of the National Statistical Service of Greece. Since this architecture is largely based on the usage of structural survey metadata, as well as on further design principles including a layered approach following the information life cycle, the derived architecture is generally applicable to different statistical organisations and surveys. The concept of structural survey metadata (i.e. data on variables and dimensions) implies that all the logical entities required for a given survey (and therefore the associated physical database objects) can be defined at any time upon the underlying generic structures of the data model. Apart from the obvious benefits of cost-effective extensibility and maintainability, this architecture ensures in an inherent way information quality, homogenisation of data and integration of statistical processes through different surveys, cross survey analysis on common as well as seamless treatment of time series.

Keywords: Statistical Databases, Statistical Information Systems, Metadata, Statistical Data Warehouses

1. Introduction

From the Information Technology point of view, statistical information systems have always presented a series of challenges and corresponding research and technological advances in many distinct areas, ranging from advanced data collection techniques and data validation to data visualisation, from metadata repositories and multidimensional data warehouses to data mining etc. Nevertheless, despite the abundance of data and tools, information still remains a scarce resource. Regardless of the merits of each individual solution, their potential can only be fully exploited through an holistic

approach focusing on the integration of statistical data and the synergy of statistical processes. Therefore, the concept of a truly integrated statistical information system is required in order to ensure the value and quality of official statistical information.

The concepts presented in this paper are largely based on the experience acquired through the design and development of the new integrated information system of the National Statistical Service of Greece. While this large scale project included work in such diverse areas as statistical metadata, dimensional data warehouses or data validation for many different statistical surveys, the urgent need for a unifying concept has early emerged. It has been clear, that this concept can only be materialised in the form of an abstract information model, encompassing both generic information structure and the information life-cycle. This paper presents the overall architectural concepts and the design methodology developed by the authors.

2. System Requirements and Methodology

2.1 Statistical Information Systems

While the primary aim of most information systems is the automation of structured business processes, an integrated information system for a statistics organisation such as the NSSG is fundamentally different. While the former have mainly to support data transactions and well-defined user requirements, statistical information systems have to produce and disseminate information out of data in a flexible way.

This information life-cycle aims to the transformation of elementary data, collected from a multitude of different sources, into valid and valuable information, suitable for statistical analysis and delivered to information consumers, with different and often ad-hoc requirements. This is somehow similar to the nature of MIS and Data Warehouses or OLAP systems of large organisations, albeit far more demanding, since in the case of the later the data to information life cycle is mostly confined within the limits of the organisation, which, in the case of Statistical organisations, is by definition not true. Moreover, since this information delivery process has to support effective decision making, the lead-time of the whole process has to be minimised.

Furthermore, statistical database schemas are conceptually different than transaction databases or even typical OLAP databases and corporate data warehouses. Statistical databases are the synthetic outcome of multiple data transfer from different sources, where data coming from small or large surveys, census or administrative records. In order to inherent the individualities of a statistical data into the architecture of an integrated statistical data schema, we consider the database as a set of random variables. To be more precise a statistical database is a set of measurable transformations from an abstract sample space Ω into an Euclidean space with known norm. This set of variables is accompanied by its underline structure (variance-covariance matrix, hierarchies) and its metadata specifications. These specifications refer both to the data set as a whole and to the individual variables separately, describe and where possible quantify the individualities of the statistical data set like the sampling scheme, non-response, editing, etc.

2.2 System Requirements

Apart from implementing a complex and usually ad-hoc process, this "information production and delivery" nature implies a series of important requirements, not usual (or at least not dominant) in typical systems:

- i) data have to be collected from different sources, using different formats, classifications and standards;
- ii) data have to be controlled, validated and edited in order to ensure data quality;
- iii) data have to be homogenised and accompanied by metadata in order to allow statistical interpretation and ensure data value;
- iv) the information production process itself includes large scale "batch" operations, affecting entire data sets; there are no record oriented processes as in transaction databases; this implies that the concept of "atomicity" of a transaction is upscaled to an entire data set, a fact which available RDBMS's do not handle well;
- v) statistical information is by nature dimensional, which may complicate a relational implementation;
- vi) last but not least, there is a dominant temporal dimension, which most often than not proves problematic in case of even slow changing dimensions (such as changes in classifications and nomenclatures).

While the above would be true even for a single "isolated" statistical survey, these main characteristics have to be augmented by further requirements, in the case of an integrated system, namely the need to allow the constant re-design of surveys, the incorporation of new ones and the need to allow the integration of data from different surveys.

2.3 The survey life-cycle

For any given statistical survey, this life-cycle includes the following stages, each setting specific requirements for the design of the system:

- i) sampling (if required) and data collection from different internal or external sources;
- ii) data entry in different formats, according to the data source, conducted in staged batches and primary control of the data;
- iii) temporary storage allowing the subsequent manipulation of these primary data, and archiving of the data sets as collected,
- iv) data validation, editing and imputation in order to ensure data quality;
- v) data transformation (where needed) and integration with other data (such as those of previous surveys);
- vi) aggregation into multi-dimensional variables, including the temporal dimension, and computation of derived data;
- vii) compilation of methodological metadata;
- viii) flexible, interactive data analysis and storage of the analysis results;
- ix) confidentiality treatment to ensure the non-disclosure of sensitive data;
- x) information dissemination, in multiple formats and with different media and channels, also allowing ad hoc information requirements.

2.4 Methodology and Approach

One of the major risks associated with the design of statistical information systems is to follow a "vertical" approach, almost exclusively focusing on each individual survey. This approach, although seemingly safer and implicitly favoured by the organisational structure of statistical organisations (where a distinct and often autonomous organisational unit is totally in charge of a group of surveys, and thus owner of the corresponding data schema), would lead to a fragmented "data mart" -like architecture, i.e. a loose set of autonomous and often heterogeneous data schemas, each tightly coupled with survey-specific reporting applications. As a matter of fact, this data-mart approach has been extensively followed in the past, at least in the case of the NSSG, and is often promoted even for large corporate data warehouses, mainly due to rapid development life-cycles and the organisational benefits of delegating data ownership directly to users.

While it may be argued that this approach better accommodates the specific requirements of each individual survey, and is better suited for an incremental development of the information system, it has severe drawbacks. Apart from the obvious disadvantages concerning long-term system maintainability and extensibility, the most severe drawback concerns the fundamental requirement of data homogenisation, which can only be ensured by a suitable data design. Therefore, the authors have opted to exploit a "horizontal" approach, following the statistical data-to-information life-cycle and enforcing a uniform treatment of statistical data. This approach consisted mainly in the construction of (a) an abstract data model, generic enough to accommodate any statistical survey; and (b) a generalised process model implementing the information life-cycle. To achieve this aim:

- 1) Data structure and processing requirements have been registered and analysed for a representative set of existing statistical surveys, using the ESA PSS software engineering standards;
- 2) Typical individual data and process models (using the IDEF1X and IDEF0 methodologies respectively) have been designed; these models were structured along the actual entities examined by each survey (for example, household, person, employer etc. in the case of the employment survey), as would be the case for a data-mart approach;
- 3) Based on these models, an abstract, generalised conceptual data model has been prepared; rather than just unifying the models, the task was based in abstracting the actual entities into generalised metadata constructs, such as statistical population, variable, dimension etc.;
- 4) The abstract model has been validated using another set of actual surveys;
- 5) A layered horizontal architecture has been designed around the abstract model.

It must be noted that the use of abstract data structures in place of actual entities, can only be achieved if the system is aware of the inherent structure of the information, or in other words, of structural metadata.

3. Structural Survey Metadata

By the concept of structural survey metadata, we mean data defining both the

conceptual information model of any given survey and the physical implementation of this model, on generic data structures, common for all surveys. In contrast to methodological metadata, these metadata are generally not intended for the interpretation of statistical information by end-users. They are always "machine-understandable" metadata, necessary for the operation of the system. This kind of metadata are usually referred to in the literature as "active" metadata (while the former are referred to as "passive").

The logical schema of structural metadata includes entities representing surveys and survey instances, variables, their data types and their usage by surveys, dimensions for the variables etc. In this aspect this schema is overlapping (at the entity level only!) to the schema of methodological metadata, and as a matter of fact, these two schemas have been unified during the final implementation.

The primary aim of the structural metadata schema is to store and manage the definitions of the logical links among the consecutive architecture layers (see following sections). In other words, structural metadata allow the bidirectional mapping between the schema of the transient data layer (which stores questionnaires) to that of the stable data layer (which stores values for combinations of variables and dimensions), as well as between the later and the data warehouse layer (which stores derived, aggregated multidimensional variables). In this aspect, the structural metadata schema is the backbone of the whole data model and actually enables the integration of the different views of statistical data during the information life cycle. On the other hand, these active metadata provide a layer of transparency between the user and the actual RDBMS physical implementation. These characteristics offer many advantages, not feasible otherwise, more specifically:

- i) The data transformation processes between the consecutive layers can be based on the metadata declarations and thus be abstract and generally applicable, eliminating the need for programming and maintenance of survey-specific application packages;
- ii) All the logical entities required for a given survey (and therefore the associated physical database objects) can be defined at the logical level (i.e. in terms of statistical entities) and created at any time upon the underlying generic structures of the data model. Thus, support for new surveys or modifications to existing ones can be accommodated for with no need for complex database maintenance.
- iii) User navigation across layers, from aggregated multidimensional data to fine granularity variable values to "raw" collected data, becomes feasible.

4. The Architectural Framework

As already stated, the architecture is structured upon horizontal layers corresponding to the different phases of the data-to-information life cycle, i.e.:

- 1) Data Collection Layer;
- 2) Transient Data Layer;
- 3) Stable Data Layer;
- 4) Data Warehouse Layer;
- 5) Data Extraction and Delivery Layer.

Obviously, for each layer, metadata, data, and application modules are required. Thus, an alternative, orthogonal model of the system would be structured along three typical vertical layers or views, namely:

- a) the metadata view, including structural and methodological metadata, as well as nomenclatures and registries;
- b) the data view, including raw data, variables and multidimensional aggregates; and
- c) the application view.

This leads to an integrated matrix-like framework of the architecture, as depicted in Fig. 1, in which individual system components (i.e. cells) and the interfaces among them can be further analysed.

4.1 Data Collection Layer

The aim of the data collection layer is (a) the collection of data from heterogeneous sources as well as (b) their primary control. Data sources may range from paper questionnaires that must be manually typed-in, to files generated by OMR, separate form-based applications or even files of differing formats (including, potentially GESMES/XML) received by third-party administrative sources. These data need to be transformed (in cases where classifications, definitions or data types differ), controlled (for simple cases of record-level validation) and finally mapped and transferred to the transient data layer.

The data view at this layer consists of the different flat data files generated or received by data sources (including "internal" forms clients for data entry of paper questionnaires). The metadata view includes file structure definitions, mappings to the transient level data structures, as well as primary validation controls and links to lookup tables (nomenclatures and classifications). The application view consists of a generic metadata-based application module which acts as the unique entry point to the system and is capable of data file parsing and performing transformation, control and mapping tasks.

4.2 Transient Data Layer

The aim of the transient data layer is the temporary storage of received data sets in order to allow for operations requiring manipulation of the entire set (i.e. which are not confined at record level) before it is transferred to the stable data layer and integrated with other data. These operations include data validation, editing and imputation tasks, complex set transformations (such as complex mappings to different classifications) and may be of a highly interactive or ad hoc nature.

Another important characteristic of the transient data layer is that it acts as a soft-locking check-in check-out mechanism for entire data sets, i.e. entire sets may at any time and for any reason be checked-out from the stable layer back to the transient one, reprocessed and checked-in again, while allowing the simultaneous existence of different data set versions and ensuring that a consistent version of the set is always available at the stable layer. Note that locking and journaling mechanisms available in commercial RDBMS's are transaction oriented and not capable of dealing with entire

row sets at an adequate performance level.

The data view includes a working database whose generalised structure allows the storage and manipulation of the data sets in a generic way, while metadata on data sets structure, surveys, samples and strata as well as mappings to variables at the stable data layer are available at the metadata view.

The application view at this layer includes:

- a) Data set versioning and archiving mechanisms;
- b) Sampling and stratification applications;
- c) The facility to allow check-in, check-out to/from the stable data layer;
- d) A connectivity gateway to external tools for data set processing.

4.3 Stable Data Layer

The stable data layer is where valid statistical data are stored as values for combinations of variables and dimensions (i.e. fact tables), and integrated with older data of the same survey or even data from different surveys. In this aspect, and mainly due to its dimensional nature, it would be better to think of the stable data layer as a data warehouse. Nevertheless, since we are using this term for the next architecture layer, we will avoid the confusion of terms. As a matter of fact one could note that, in standard industry terms, both layers are data warehouses, with the difference that while the next is a typical aggregated star-schema based first-generation data warehouse, the stable data layer is a fine granularity second-generation one.

The data view includes a homogeneous schema consisting a series of distinct but similar fact tables, one for each measured statistical variable, while the metadata view includes structural survey metadata as well as repositories for registries and nomenclatures.

In a uniform way, each fact table is logically connected, on the one hand to an abstract entity expressing a combination of a specific survey instance and a specific statistical population unit, and on the other hand, through a possibly multiple relationship, to the abstract dimension entity. It must be noted that, using extended E-R concepts, the statistical unit entity is the abstract top-most super-type of the hierarchy of registries, while the dimension entity is the logical link to a system of temporally aware classifications and nomenclatures. These concepts are depicted in Fig. 2.

Therefore, one can define the data structures (i.e. the fact tables) of any given survey, simply by declaring the variables and the dimensions for each fact, and storing these structural survey metadata in the appropriate table of the structural metadata schema.

This concept has a series of interesting implications:

- The appropriate physical database tables can be generated or altered directly from these structural metadata, using a dynamic SQL application; thus the task of implementing a new survey or modifying an existing one does not require complex database maintenance tasks.
- All surveys are using a uniform landscape of fact tables sharing common dimensions, which facilitates user navigation and enables the integration of statistical data.

- Through structural metadata the data transformation processes among the consecutive layers can be automated by generic applications.

It is worth mentioning here that the temporal nature of this layer, as well as of the next, (i.e. the requirement to integrate data across the time dimension) is particularly challenging, especially due to slow changing dimensions, such as classifications or registries. The problem stems from the fact that the relational model is inherently designed to provide consistent snapshots of the reality (as required in typical transaction databases). Therefore, until temporally aware databases become available, we are limited to relational constructs with some drawbacks.

The metadata view also includes a repository of methodological passive metadata. Although they serve different purposes since the two metadata schemas overlap at the entity level, they have been integrated.

4.4 Data Warehouse Layer

The aim of the data warehouse layer is twofold, namely to facilitate subsequent statistical analysis and integration of information by providing multidimensional variables sharing common dimensions, as well as to ensure query performance by providing aggregates of these variables.

The data warehouse has been designed using a typical star schema dimensional model methodology, in order to implement a set of multi-dimensional hyper-cubes for each survey. The data view consists of relational implementations of these hyper-cubes, while the metadata view holds the definitions of the later in terms of dependent variables corresponding to fact tables of the stable data layer and dimensions.

In contrast to typical data warehouses, due to the uniform structure of the stable data layer, the transformation of data from the later to multidimensional aggregated variables is straightforward and might also be automated by a metadata aware generic application.

4.5 Data Extraction and Delivery Layer

The aim of this layer is the extraction of data (normally from the data warehouse but in some cases directly from the stable layer), the extraction of the corresponding methodological metadata and the creation of suitable products, ranging from pre-defined reports and statistical publications to ad-hoc queries, possibly via a dissemination web site.

The metadata view in this layer includes report and query definitions in terms of variables and dimensions. Since the data warehouse structure is also known to the system in the same variable and dimension terms, data extraction and delivery applications can be designed to be aware of the structural metadata, and therefore flexible and generic enough, at the same time eliminating the need for the user to know the details of the physical database schema implementation.

Furthermore, the existence of dimension and variable metadata makes navigation

feasible. Two kinds of data navigation are possible, namely:

- Vertical (or drill-down) navigation, allowing the user to get data for specific dimensions at increasing levels of detail;
- Horizontal navigation, allowing the user to navigate across different variables through shared dimensions.

The application view includes mainly:

- a generic query and reporting facility which can undertake the job of formulating at run time suitable queries according to stored or ad-hoc query definitions; different front-end interfaces can use this same facility, ranging from reports to interactive query tools.
- confidentiality treatment tools;
- a gateway offering connectivity to external statistical analysis or data visualisation applications.

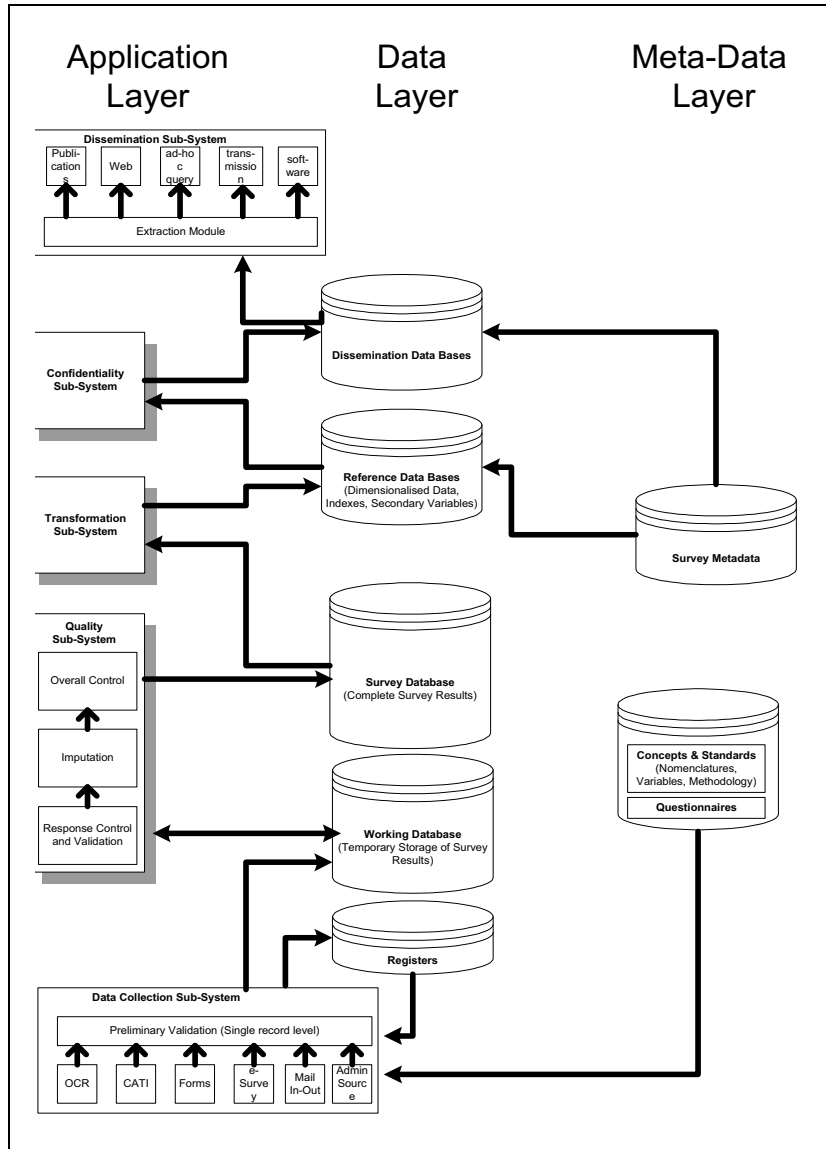


Figure 1. Architectural Framework.

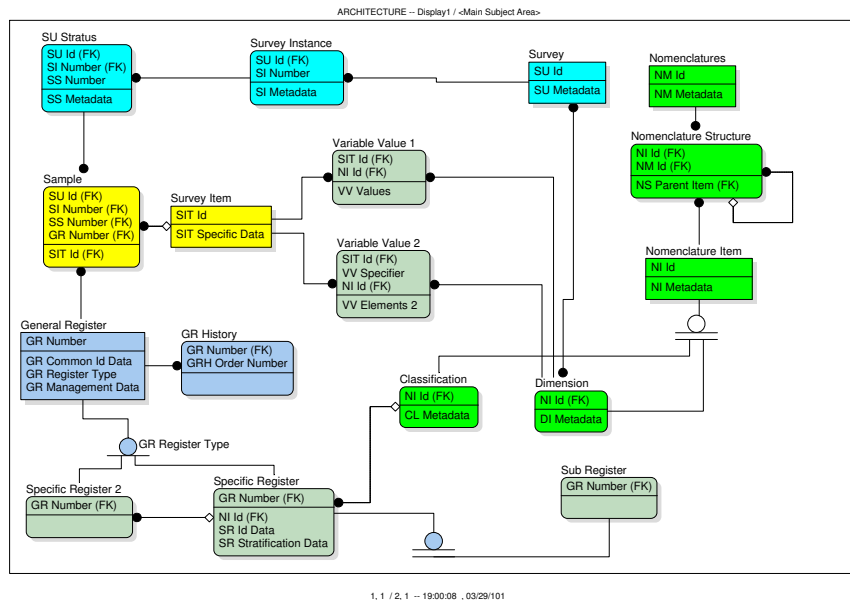


Figure 2. Stable data layer schema.

5. Conclusions

A methodological framework for the design of an integrated statistical information system has been presented. While this framework is layered according to the data-to-information production life-cycle, active metadata on the information structure at each level allow the effective integration of the different architectural layers and the design of generic applications.

Apart from the obvious benefits of cost-effective extensibility and maintainability, this architecture ensures in an inherent way information quality, homogenisation of data and integration of statistical processes through different surveys.

References

- [1] United Nations, 1995, Guidelines for the modelling of Statistical Data and Metadata, Conference of European Statisticians, Methodological Material, Geneva.
- [2] United Nations, 1999, Information Systems Architecture for national and international statistical offices, Guidelines and Recommendations, Conference of European Statisticians, Statistical Standards and Studies, No.51, Geneva
- [3] Rakesh Agrawal, Ashish Gupta, Sunita Sarawagi, Modeling Multidimensional Databases (1995), IBM Almaden Research Center 650, Proc. 13th Int. Conf. Data Engineering, ICDE
- [4] A. Shoshani, OLAP and statistical databases: Similarities and differences. In Sixteenth ACM SIGACT SIGMOD SIGART Symp. on Principles of Database Systems, pages 185#196, 1997.
- [5] De Giacomo, G. and Naggar, P. 1996, Conceptual data model with structured objects for statistical databases. In Proceedings of the Eighth International Conference on Statistical Database Management Systems (SSDBM'96). IEEE Computer Society Press. 168-

