Distr.
GENERAL

ECE/CES/SEM.54/14 (Summary)
6 April 2006

Original: ENGLISH
ENGLISH AND RUSSIAN ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANIZATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD Seminar on the Management of Statistical Information Systems (MSIS)
Sofia, Bulgaria, 21-23 June 2006

Topic (ii): Dissemination and client relations

MULTIDIMENSIONAL STATISTICAL DATA DISSEMINATION ON THE WEB

Supporting Paper prepared by Stefano De Francisci, Giuseppe Sindoni, Istat, Italy and Leonardo Tininini, IASI-CNR, Italy

Summary

1. A statistical web warehouse is a system specifically designed for Web-based dissemination of statistical data. Data are stored in a data warehouse (Kimball, 1996) and are accessible from the Web through hypermedia navigation functions, enabling the user to select and dynamically visualize data in various formats.

2. There is a close correspondence between statistical databases and data warehouses (Shoshani, 1997): (a) microdata correspond to fact tables; (b) macrodata correspond to data cubes and macrodata summary attributes correspond to cube measures; (c) category attributes and their classification hierarchies correspond to dimensions and dimension levels; (d) roll-up operations provide summarisation, i.e. operations to shift from a more to a less detailed aggregation level; (e) conversely, drill-down operations provide shifting from a less to a more detailed aggregation level. Unfortunately, some peculiarities of statistical data with respect to business data call for extensions to data warehouse techniques with specific models and structures. One peculiarity is about surveys based on samples. Aggregates coming from sample microdata normally require significance checks that are not needed in traditional data warehouses. Another peculiarity concerns privacy and *secondary disclosure*. Due to these requirements, traditional data warehouse systems cannot be used for Web-based statistical dissemination without any customisation, because they would allow users to arbitrarily navigate all dimensions available for a given fact, without providing functions to check for conformance to the above principles.

3. This paper presents a generalized technique for Web-based dissemination of statistical data, the underlying conceptual model for spatio-temporal multidimensional data and the provided data warehousing functions. This technique has been used to design and implement a generalised system, called DaWinci/MD (Sindoni G., Tininini L., 2006), currently used for data dissemination at Istat.

4. The DaWinci/MD system is based on a model for statistical tables, designed for Web-based data access. The model is based on a decomposition of the information space into *basic multidimensional tables*, represented by a pair of components: the **object of interest**, i.e. the table measure (e.g. a summary attribute), and the set of **classifications**, i.e. the dimensions used to classify the table measure. Each basic table can have several spatio-temporal instances (spatio-temporal multidmensional tables) by adding a data **time stamp** and **spatial context**. Hence, when choosing a spatio-temporal multidimensional table to be visualized on the Web, each selection step defines incrementally the four table components. Then, a typical query to the system specifies all or part of the components of the <*t*, *s*, *o*, *c*> quadruple, where: *t* is the time stamp; *s* is the spatial context, i.e. the combination of a territorial detail *d* and geographical area *a*; *o* is the object; and *c* is a set, possibly empty, of classifications.

5. The above table decomposition is also the conceptual basis of the information storage model that represents the spatio-temporal table availablity in the system metadatabase. In this way, the storage of a <*t*, *d*, *o*, *c*> quadruple means that the basic table defined by the <*o*, *c*> pair is available for the *t* time stamp and for all territorial details up to *d*.

6. Statistical objects and classifications are organized in hierarchies, where a specialization relationship holds between parent and child objects (or classifications). In order to facilitate navigation, “virtual” objects (or classifications) can be part of the hierarchy. They group conceptually more specific objects, and they do not correspond to any data warehouse measure. In general, the more generic an object, the higher the number of basic multidimensional tables referring to it.

7. An object, a set of classifications and a spatio-temporal context identify a set of available multidimensional tables, i.e. a set comprising all tables with the specified, or a more specific, object, the chosen, or more specific, classifications, and such that they can be instantiated up to the required territorial detail.

8. Each table is compatible with one or more spatial contexts, depending on the maximum territorial detail, as specified in the system metadata. A geographical area and territorial detail are chosen at the same time. They are not independent and the choices available depend on the specific territorial hierarchy. The choice of a higher or lower territorial detail respectively limits or extends the number of tables corresponding to a chosen object and set of classifications. Each successive choice of a parameter value decreases the number of further available parameter values and compatible multidimensional tables. In this way users are prevented from requesting tables corresponding to sensitive data or dimension combinations that are meaningless or not planned for dissemination. Conversely, the removal of a chosen parameter usually decreases the number of constraints and therefore increases the number of further compatible choices.

9. The table chosen among all those compatible with the selection criteria is visualized in one or more web pages, depending on the number of involved classifications. Starting from the visualized table it is possible to remove or add a classification; increase or decrease the territorial or classification detail; change the geographical area. The removal of a classification results in the visualization of a less detailed statistical table. From the data warehouse point of view, this corresponds to a roll up operation on the currently visualised cube (table), i.e. to define an *(n-1)*-dimensional sub-cube starting from an *n*-dimensional one. Conversely, adding a classification corresponds to a drill down operation on the visualised cube, i.e. to define an *(n+1)*-dimensional super-cube starting from an *n*-dimensional one that is a more detailed statistical table.

References

Sindoni G., Tininini L. (2006) Statistical warehousing on the Web: navigating troubled waters. *Proceedings of the International Conference on Internet and Web Applications and Services. IEEE Computer Society Press.*

Shoshani A. (1997). OLAP and Statistical Databases: Similarities and Differences. *Proceedings of the PODS 1997 Conference*.

Kimball R. (1996). *The data warehouse toolkit*. John Wiley & Sons.
