WP.17
ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 9-11 February 2004)

Topic (i): Functions of metadata in statistical production

METADATA DRIVEN STATISTICAL DATA WAREHOUSE SYSTEM AT THE HUNGARIAN CENTRAL STATISTICAL OFFICE

Contributed Paper

Submitted by the Hungarian Central Statistical Office [1]

SUMMARY

In the Hungarian Central Statistical Office there is long tradition of the metadata driven applications.

The first application was introduced in 1984. It was an on-line query system for economic statistical database (SOLAR), which was built on the metadata description of statistical indicators (their content, population, aggregation level, etc.) and classifications (nomenclatures). Unfortunately this system had relatively few users because of its obsolete software solution and the lack of terminals in the Office but the application lived for eight years.

The second system was made for the generalised solution of the survey control task. The GESA system is responsible for assigning the population of data collections, personalising, mailing questionnaires, supervising responses and non-responses, quality control, estimating costs from 1995. This system is based on the description of questionnaires, the characteristics of data collections in the metadata base. Nowadays the GESA manages 80 percents of data collections (180 data collections).

In the latest years more general applications were developed: a frame application for data editing (ADEL), another application for electronic data collection (KSHXML) and the most generally used application for data warehouse (STATINFO). All of them are built on the metadata-base.

The paper describes the metadata driven Statistical Data Warehouse System at HCSO, the structure, the applications and the metadata which are necessary for their operation.

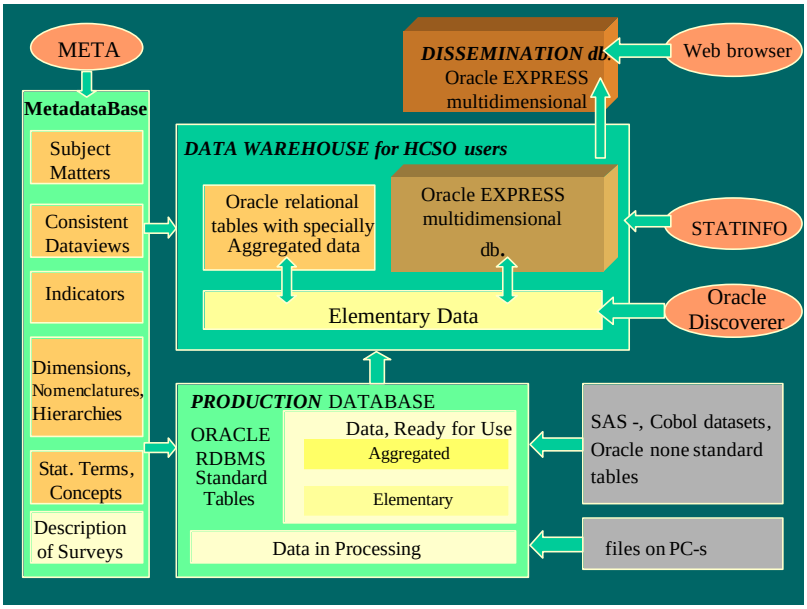
[1] Prepared by Ildikó Györki (ildiko.gyorki@office.ksh.hu) and Imre Pap (imre.pap@office.ksh.hu).

INTRODUCTION

The new idea, the data warehousing turned the attention to the OLAP (On-line Analytical Processing) tools at HCSO. Having purchased Oracle Express Software Package in 1998, HCSO started developing a statistical data warehouse system (referential database) and a dissemination database. The HCSO statistical data warehouse system serves statisticians in the central and regional statistical offices in Hungary. The dissemination database is a Web enabled system for users outside the Office.

The structure

The structure of the HCSO approach is the following:



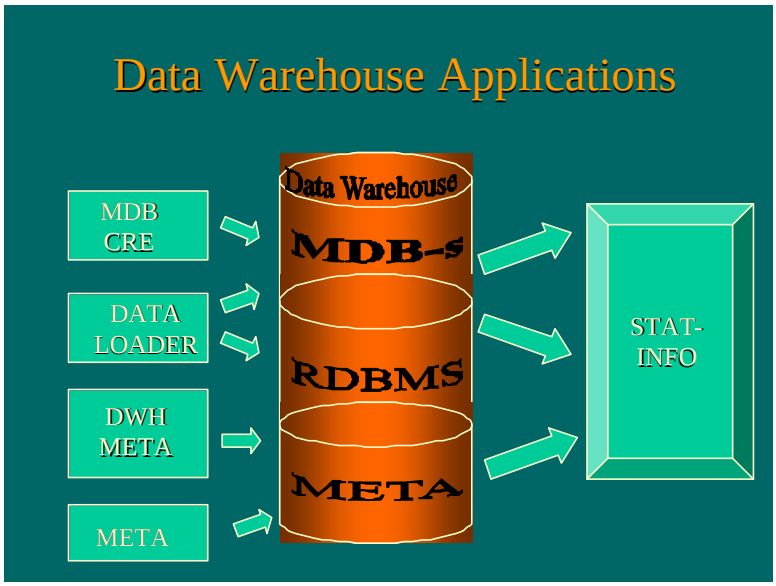
where „META" and „STATINFO" are home made applications. Oracle Express Analyzer called from the STATINFO application and Oracle Discoverer are OLAP tools to visualize data in the forms of tables and graphs.

The HCSO DWH system is based on the Production Database where data are stored in 3rd normalised form and their content (observation units, population, relational data tables, indicators, statistical terms and concepts, classifications) is described in the metadata base. Additional statistical and technical metadata are added to the metadata, which are used in the Production Database, and on that common metadata base special processes are used to create DWH elements and load data into them. The bilingual (English-Hungarian) Dissemination Database is part of the HCSO DWH system. Its content is checked in the DWH system, but the data belonging to it are copied to the HCSO Web site server.

The aim of the HCSO DWH system is to store the whole HCSO data "wealth" in a demand oriented form. The metadata base (structured and textual descriptions of statistical terms, methods, dimensions, variables, hierarchies, etc.) is an essential part of the statistical data warehousing system. It plays an important role in providing information on the content for end-users and in the explanation of the presented data. The combined OLAP tool gives end-users an interactive, flexible way to retrieve and analyze statistical data. An important goal is to automate the DWH builder processes to as great an extent as possible, in order to reduce the necessary workload and workflow. A metadata driven solution serves all these ideas.

The applications

The applications used in the HCSO DWH approach are as follows:



- Creation of the DWH elements.

  After the DWH Editorial Board has approved the application form for a new consistent data view (this is the basic unit in the HCSO DWH system) statisticians using the META application enter or finalise (if metadata are already available from the production processes) the metadata for which they are responsible. The IT DWH staff use the DWHMETA application to add the technical metadata to the system and to check the whole set of metadata for completeness and correctness. The IT DWH staff use the MDBCRE application to build the Oracle Express databases from which the new consistent data views will be retrieved.

- Loading data.

  Using the technical metadata the DWH DATALOADER loads the data belonging to one or more consistent data views to the newly created Oracle Express or Oracle RDBMS database. (Very large in size or very frequently changed data are loaded to Oracle RDBMS database.) The loader process can be executed interactively or in batch mode. The source data must be described and stored in the HCSO production database. If the data are originally stored in an other way (in MS Excel, SAS, COBOL files, or in non-standard Oracle tables) special E(xtract)T(ransfer)L(oad) processes are used to put the data into the HCSO Production database.

- Data retrieval.

  Statisticians in the central and in the regional statistical offices use the home made STATINFO client-server application to retrieve data from the HCSO DWH system. The visualisation of the data is made in tables or graphs by calling the Oracle Express Analyzer program (OLAP tool). Special functions were added to calculate percentages and statistical indexes. The retrieval system is also metadata driven and can show statistical metadata belonging to the data.

Main metadata in these applications

In the normal data warehouse concept we need the data warehouse database and we have to have a system that replicates data from production databases to the data warehouse database. In addition, we need some tools to report and visualise data. All of the tools have to use and display metadata.

HCSO metadata used in the STATINFO application:

The object **Subject Matter Areas** describes the topics (hierarchically, e.g. demography which is divided into the observation of births, marriages, depths, migrations, etc.) used in statistics to classify the statistical data groups.

The object **Consistent data views** describes a set of comparable (homogenous) indicators with the same dimensionality (e.g. time period, territory, sex, marital status, etc.) and same population for the observation units (e.g. enterprises in the industrial branches where the number of employees greater than four).

The object **Population** describes what is the population of the observation (e.g. enterprises in the industrial branches where the number of employees is greater than four).

The object **Indicators** describes the so-called statistical indicators or variables (like production value or number of employees, etc.). The indicator is a quantitative measure of the observed unit.

The object **Indicator versions** (or variants of the indicators) specifies the different (real) instances of the indicators available to the users. The instances differ according to the time period and the set of classifications (or nomenclatures) used to summarize (aggregate) the collected (from the observation unit) values.

The object **Statistical Concepts** describes the terms, which can help the users of the statistical system to interpret the content of the statistics and the statisticians to plan data collections. For supporting the users this part should be complete, to cover all indicators available for the users. There is possibility to specify the relations of the concepts like broader, narrower, related terms, synonyms, abbreviation, ascendant and descendant, etc.

The object **Definitions** helps to give different interpretations for the concepts e.g. for different time intervals if the meaning of the concept was changed.

The object **Dimensions** describes the final collection of the classification version elements which serve the user to analyse statistical indicators. The elements of the dimensions can be shown in different hierarchical (subordinated) orders.

The object **Hierarchies** specifies the rules, how the dimension elements are subordinated to each other, and which sort of elements the user can find in the given hierarchy (e.g. the territory dimension can contain a hierarchy with the settlement classification according to the number of inhabitants in the settlement, but it can also contain an other hierarchy, where the legal status (i.e. village, town, capital town) of the settlement is used).

The object **Hierarchy Levels** (i.e. aggregation level) describes the content of the hierarchy. Using the name attribute of the level the users are able to select the dimension elements belonging to the given hierarchy level (e.g. the territory dimension can contain country, region, county, settlement hierarchy levels).

The object **Time periods** is one of the dimensions (i.e. the most important dimension whose behaviour is very special and therefore it needs different treatment). The time period is so important during the specification of the analysis process, that it should be displayed as characteristics for the given consistent data view.

The object **Roll up needs** describes the hierarchy level at a given hierarchy of the dimension, where the system should start rolling up (aggregating) the value of the indicator (for the not specified levels the values should be loaded directly from the source table).

The object **Multidim. DB-s** describes the data storage for the given consistent data view. The content of the object is hidden for the users and is used only by the applications.

To build DWH elements we also need the following objects:

- The whole classification database as it was described in the Production database part,
- The data sources from which the DWH DATALOADER loads the data,

- the description of the data on elementary level if there is a need to show elementary data behind a cell of the displayed table.

The experience with the metadata driven systems and the conclusion will be presented on the METIS conference.