

Distr.
GENERAL

ECE/CES/SEM.54/30
8 June 2006

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 Seminar on the Management of Statistical Information Systems (MSIS)
Sofia, Bulgaria, 21-23 June 2006

Topic (i): Changes in statistical processes

USING ADMINISTRATIVE SOURCES - GROUNDS FOR IMPROVEMENT OF THE STATISTICAL PROCESSES

**Supporting Paper prepared by Matjaz Jug,
Statistical Office of the Republic of Slovenia**

I. INTRODUCTION

1. Slovenia is a register-oriented country. We've been using data from the administrative sources for statistical purposes for a long time. However in the last period there has been even stronger orientation towards replacing primary data collection with using data from secondary sources in order to reduce the administrative burden on reporting units and cut the expenses. For us it is a great challenge how to reorganize the statistical process and IT architecture.
2. The paper will focus on the impacts of using administrative data in different parts of the statistical process: from data acquisition on input to the analysis of final statistical results.
3. Firstly, the current methods and solutions used in the Statistical Office of the Republic of Slovenia to effectively collect and analyze data from administrative sources will be revealed. We are collecting data in different ways and in different formats. In case of some data sources with frequent data exchange, the direct database replication method is used, whereas other data come to the office on media (CD, DVD) or by using e-mail. Data, when the volumes are big, are usually loaded into input databases, where statisticians have the possibility to analyze the data. Because of the increasing number of data sources, we plan to establish a general multi-modal input solution for data from secondary sources. The solution should also enable introduction of more flexible and powerful controls on raw data, providing the opportunity to reveal potential problems and inconsistencies as early as possible, and have the possibility to fix them in one

place in order to avoid inefficient and sometimes inconsistent corrections on many places further in the statistical process.

4. Extended usage of the administrative data has a great impact on statistical processing, too. Because of data from both administrative and primary sources and advanced statistical methods different types of metadata (conceptual, process etc.) are becoming more and more important. We have found the way to standardize them in order to be used in every survey where the need to use it could potentially arise. Namely, in databases can be stored both primary and secondary data directly accessible for analysis and production of statistical results, together with some important metadata. In Structural Business Statistics (SBS) there is the need to combine data from different sources to the same variable, depending on the part of population observed. In the new SBS production system it is now possible to change the data source or add the new variable in a more flexible way.

5. The data from administrative sources often integrated with primary or other secondary data are very important for external researchers. After the introduction of the possibility on secure remote access for researchers which enables them to do their analyses in a more comfortable way (data for their projects are stored on a therefore dedicated server, separated from the internal network) there is the need to improve the process of data preparation. The Statistical Office of the Republic of Slovenia (SORS) is often in the position of being the only institution capable to match data from different administrative sources for researchers. However, the process of data integration and anonymization could be extremely resource demanding, so the introduction of a system for data integration (including statistical processing needed for it) is planned.

6. The effective statistical process must be supported with appropriate IT system(s). In further modernization of the collection, processing and dissemination systems in SORS great importance will be put on the extensive use of administrative data.

II. CURRENT SITUATION

7. The Slovenian Statistical System is register-oriented for over 25 years, so the usage of data from the administrative sources for statistical purposes has a very long tradition. The main administrative registers, i.e. the Central Population Register (CRP), the Business Register of Slovenia (PRS) and the Register of Territorial Units (RTE) have been established in SORS and then transferred to other institutions¹. There are additional statistical registers maintained in SORS, such as the Statistical Register of Farms (SRKG) and the Statistical Register of Employment (SRDAP). Besides the main registers there are around 140 administrative sources used for regular statistical production.

8. CRP and PRS are used as a basis for the frames for persons and enterprises, respectively. Data from CRP are collected as quarterly snapshots of personal data accompanied with the lists of transactions for the main demographic events: births, marriages, etc. Access to PRS is realized

¹ The Central Population Register is now kept by the Ministry of Internal Affairs, The Business Register of Slovenia by the Agency of the Republic of Slovenia for Public and Legal Records and Services (AJPES), the Register of Territorial Units has been renamed to the Register of Spatial Units and has been kept by the Surveying and Mapping Authority of the Republic of Slovenia (GURS)

by using a database replication of the current version and the whole history of transactions in Business Register. A similar method is used for RTE which is used for populating the territorial aggregation lists and dimensional tables in the data warehouse.

9. Data from other administrative sources are coming in the office on both individual and aggregated levels. For most of administrative data the process is similar: after the formal control of the data, important variables from CRP (location of the settlement, sex, age, etc.) or PRS (main activity code, main location of the business subject, etc.) are added, depending on the type of statistical unit (persons, enterprises). For some commonly used data sources (for example tax data) input databases with the possibility to analyze data and perform common validation rules have been established so there is the possibility to discover potential inconsistencies early in the process and to perform appropriate activities at a single point (for all surveys using that data source). There is not a single input solution covering all the administrative sources. For every source which is subject of input control, the IT solution for loading, storing, analyzing, validating and editing the data is developed as custom-made and usually from the scratch.

III. CHALLENGES IN EXTENSIVE USE OF ADMINISTRATIVE SOURCES

10. In the last period there has been even stronger orientation towards replacing the primary data collection with using data from secondary sources in order to reduce the administrative burden on the reporting units and to cut expenses. This process usually results in combined use of both primary and administrative data. There are the new challenges, connected with this approach:

- Harmonization is needed. For effective replacement of primary data sources with data from administrative sources there is a need to harmonize metadata (particularly variable definitions).
- The new challenges in the field of using administrative sources are drivers for new centrally managed statistical registers for persons and enterprises, covering all the populations we are observing.
- Multi-modal input solution for secondary data sources with the possibility to analytically access and control large volumes of data (data profiling).
- The need that metadata (describing data sources, methods used, etc.) be placed in a common production database together with data.

A. Harmonization of the variables (ISO/IEC 11179)

11. In SORS we have started to implement the international standard ISO/IEC 11179 to describe variables (data elements). The usage of this standard offers many advantages:

- Formalized metadata collection in a structured and exact way.
- The possibility to standardize and harmonize metadata by using the top-down approach.
- International comparability.

12. In the current statistical process variables are defined in inconsistent way. There are different reasons for that:

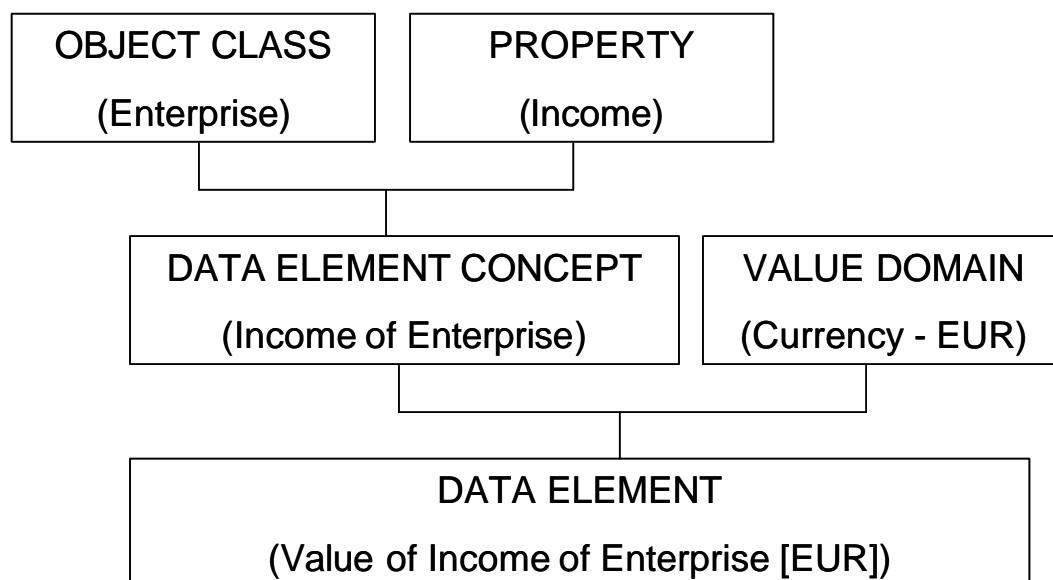
- The data structure on the questionnaire can differ from the data structure in the administrative source so the same variable can be defined differently.
- Variables definitions are often based on the IT tools used in the production process, so the same variable can have a different name in Blaise, SAS and the Oracle database.
- Variable names are defined locally for a particular survey, so there can be the same variable name used in different surveys (for different variables).

13. In order to establish a new standardized list of variables, we have started with three pilot surveys covering different statistical units and trying to cover all the lifecycle of the variable. The top-down approach is enabling us to define the basic entities (parts of variable definition) centrally (see figure 1):

- Object class describing type of statistical unit (Enterprise, Person, Household, etc.).
- “Global” properties used by different surveys (Income, Salary, Sex, Activity, etc.).
- Representation used to derive data elements from conceptual variables (ValueOf, NumberOf, TypeOf, etc.).
- Value domains used by different surveys (NACE, NUTS, Currency, etc.).

14. With the proper application, registration of new variables can be viewed more as a compounding of new combinations (data elements) from predefined parts (and adding missing parts) enabling a more harmonized approach.

Figure 1: Basic entities covering ISO/IEC 11179



B. Multi-modal input solution for secondary sources

15. The extensive use of administrative data requires a new, more centralized data management approach on the input side of the statistical process. The future target architecture for administrative sources will comprise main event-based statistical registers (starting with the

business register and persons register), connected with the common input database for secondary sources. The system will enable the multi-modal data acquisition (covering different electronic channels usually used for the exchange of administrative data and metadata between institutions), formal controls on data (data profiling and automatic or manual editing), linking with data elements from main registers, querying/analyzing data and extracting data elements for further usage in statistical surveys.

16. Metadata, collected in a standard way, will be used for driving the process. The ISO/IEC 11179 based metadata structure can be a basis for centralized metadata - driven data management:

- The conceptual variable (combination of the object class and property) defines to which basic register (frame) the input variable should be linked (business, person, etc.) and what is the grain of data (level of detail),
- The data element (derived from the conceptual variable with the value domain specified) can be a basis for formal control on the input variable. For each value domain, formal control can be specified and used on any input variable connected to this value domain.

17. For now we have started with the prototypes needed to discover the feasibility of such a solution on more detailed level (definition of the business rules for collecting metadata based on ISO/IEC 11179, testing new IT tools for data profiling, ETL and analysis, testing generic data models etc.). The development of the solution is planned for 2007.

C. Surveys combining administrative and primary collected data in a common, metadata rich production database

18. In recent years different databases have been built to support effective data processing and analysis of data for surveys based on data from both primary and secondary sources. In 2001 Structural Business Statistics database was built combining 13 different sources (the Business Register of Slovenia, the Statistical Register of Employment, SBS surveys, annual account reports of enterprises, tax declarations, etc.). The solution was designed as a dimensionally modelled data warehouse. There were 13 fact tables (one for each data source) with up to 1 million individual records. The aggregates for about 50 variables defined according to the EU Commission Regulation No. 2701/98 were stored in the SBS macro database.

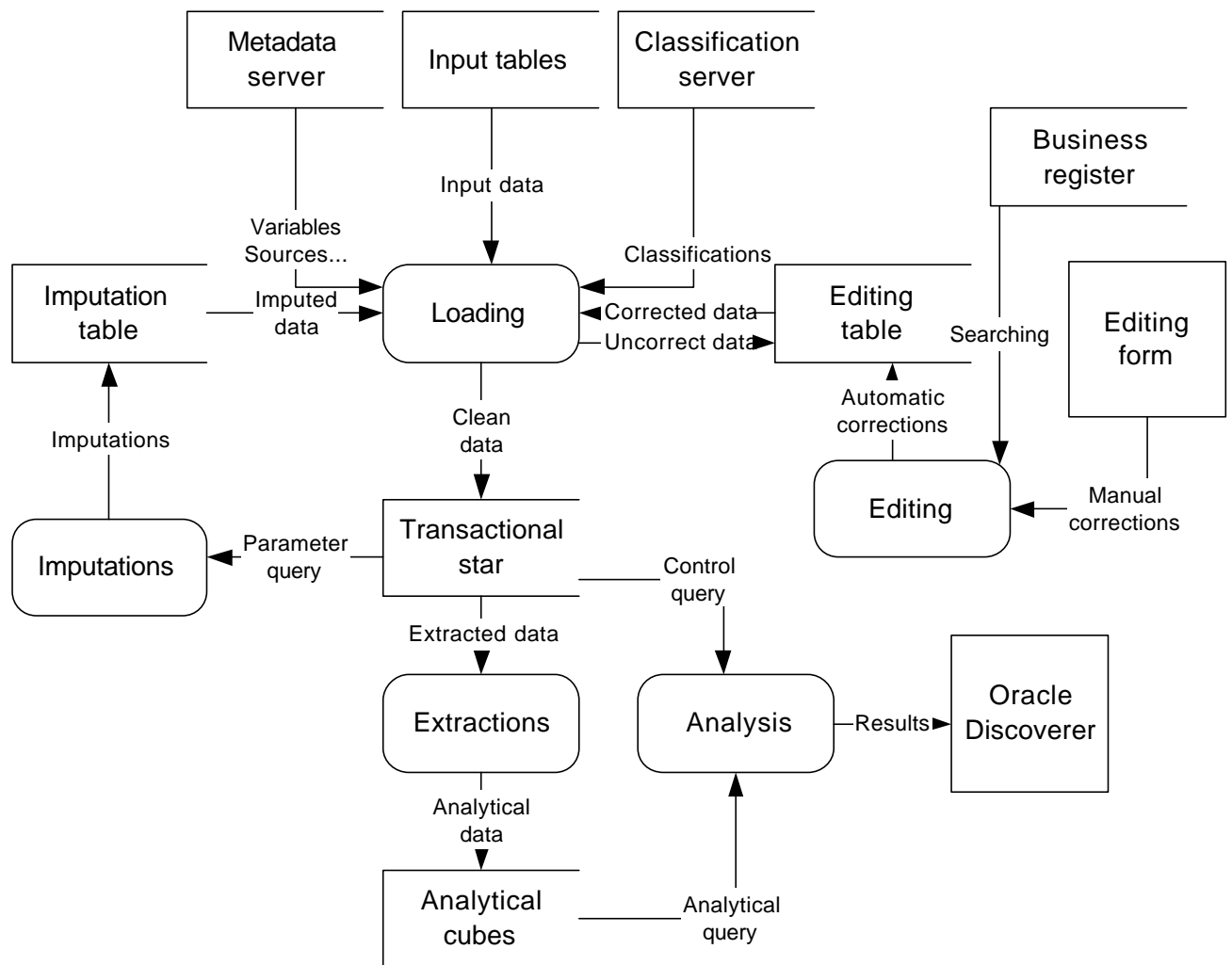
19. In 2003 the database for Structural Earning Statistics (SES) combining variables on both the enterprise and person grain was developed. The SES micro database contains integrated data, which come from many different sources: the Business Register of Slovenia, the Statistical Register of Employment, education, tax, the SES survey, etc. SES was designed with the proposition to be a production system enabling joining the sources, imputations and analysis, so tables for different processing stages were created in the database, resulting in one big fact table for administrative sources with 1.5 million records (data for 2000 and 2001) and one small one for joined results of the survey.

20. In 2005, the SBS database was upgraded in the production data warehouse. Experiences from SES revealed that the fixed data model with different tables for each processing stage doesn't adequately suit frequent changes in the data structure and methodology (for example

changes in the questionnaire). Moreover, the requirements for a SBS production database comprised a metadata rich model with classifications (including drill up/down hierarchies), process indicators (describing whether a continuous variable was reported, edited or imputed), variable definitions (with variable code, short name, description, English name, etc.), data source indicator (with the possibility that some particular continuous variable can be loaded from more than one source, depending on the period or part of population). The process required a robust loading mechanism with the possibility to change data sources (for example from primary to administrative) for existing variables and to add new variables, perform general imputation and editing procedures, standard extraction procedures and analytical possibilities.

21. Data for the SBS database are loaded from different data sources into one single transactional fact table, structured in a way that changes in the data source or data structure don't influence the loading procedure. All data transformations are based on records in two dimensional tables: the data source and data elements (continuous variables) dimension. Since data are coming from both the primary and secondary sources, the mechanism for loading weighted and unweighted data is used, filling the corresponding two columns in the central SBS fact table. The third metadimension in the database contains the indicators for status of data, describing whether data are from the primary/secondary source and whether they are reported/edited or imputed. This information is submitted as a parameter in the loading procedures. A procedure for refreshing classification-related metadata directly from classification server has been developed; a similar approach is planned to be used for metadata describing data elements, too. During the initial SBS fact table data load, all incorrect data (violating constraints) from any data source are transferred to a special staging table, where data can be cleansed and loaded again. A similar approach with a special table is used for the imputation procedure. Data, needed for imputation programs, can be browsed using a query to the central fact table. For data export purposes standard procedures have been developed for transforming any data selection to formats suitable for further processing done by specialized tools for dissemination, confidentialisation, etc. Advanced analyses and calculations can be performed by using data cubes regularly extracted from the central fact table. The whole data flow is shown on figure 2.

Figure 2: Data flow scheme for SBS architecture



IV. CONCLUSION

22. SORS has a long tradition on using data from administrative sources. Step by step we are building new production systems for effective data and metadata management. Based on our current experiences, the key success factors comprise the extensive usage of well-defined metadata, complete definition of the process (requirements not known in the design phase can slowdown the implementation), usage of generic data models insensible to changes in methodology and data structure and good cooperation with methodologists and subject-matter statisticians.
