Distr.
GENERAL

CES/AC.71/2005/16
2 March 2005

ENGLISH ONLY

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

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD) STATISTICS DIRECTORATE

Joint ECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems (MSIS)
(Bratislava, Slovakia, 18-20 April 2005)

Topic (ii): Development strategies for statistical information systems

STATISTICAL METAINFORMATION SYSTEM AND DATA WAREHOUSE AS MAIN SUBSYSTEMS OF SIS - CURRENT SITUATION AND PLANS

Supporting Paper

Submitted by the Central Statistical Office of Poland[1]

I. CONCEPT OF STATISTICAL INFORMATION SYSTEM (SIS)

1. The main goal of the integration of statistical databases was to assure, that services offered by official statistics will be continuously available by improving access to the information already gathered in many distributed databases. The history of SIS goes back to 1996, when a team from different units of Central Statistical Office (CSO) was established by the President of the CSO. The task given to the team was to prepare a concept of an integrated system of information processing for economic statistics as well as the creation of a database covering that range of data. In October 1999, the results of the work were presented to the President of the CSO and after broad discussions the decision was taken to implement the concept. Work undertaken in parallel included a few subgroups and covered: glossary of terms, classifications and other metadata systems, description of data entering processes, data dissemination and administration of the system. The creation of a Statistical Metadata System (SMS) was necessary to be able to create Statistical Data Warehouse (SDW), which, from the very beginning, had to be metadata driven.

2. In addition to the creation of SMS and SDW for the purpose of internal access from the whole statistics, the following decisions were undertaken:

- existing data entry systems (for over 200 questionnaires) were left temporarily outside of SDW – although some of them are SDW source systems;
- harmonizing above-mentioned data entry systems with SMS was postponed to a later date;
- aggregated data existing in SDW are transferred directly to other databases such as: Regional Data Bank, Eurostat, some government agencies, and will be used for dissemination;

[1] Prepared by Krzysztof Kurkowski (k.kurkowski@stat.gov.pl) and Stanislaw Sieluzycki (s.sieluzycki@stat.gov.pl).

- creation of a Public Data Warehouse (PDW) will be accessible for external users and also available from internet; implementation of this point requires additional hardware and software resources and will be done in 2005-2006.

3. At the concept level the following main activities were assumed:
- subject matter integration,
- physical integration,
- automatization of processes following the surveys.

4. By subject matter integration is meant composing terms, classifications etc. in one SMS. Names and definition of terms must be unique, sources and authors strictly defined. Some of the terms and classifications are treated as dimensions in SDW. The integrity of subjects used and stored in the survey frame is also important. Classifications (or small groupings) consist of elements, which must also be unique. It means that if one such element is included in a few classifications, its name is the same and there is a standard classification that includes the above-mentioned element.

5. For subject matter integration the following sequence of activities was assumed:
- within object integration:
  - building a glossary of terms with unique names and definitions,
  - classifications to ensure coherence and groupings according to standards,
- within subject integration:
  - building a statistical business register BJS with unique subject description in all surveys for a given period of time,
  - building a statistical agricultural register SRGR.

6. Physical integration was achieved by putting a maximum set of results into one integrated system. Data and metadata are available to any statistician with proper access. For physical integration the following sequence of activities were assumed:
- building of the SDW, which includes Census, economic and agricultural data,
- modernization of the Regional Data Bank (BDR).

7. The automatization of processes following the surveys was developed gradually, in an optimal way for a given time.

II. ORGANIZATIONAL CHANGES

8. One of the crucial points in integration and automatization of processes following statistical production is how business works are organized. Building of an SIS required some organizational changes: an additional section in the Survey Coordination Division (OB), which consists of 5 persons, was established, as well as a coordinator at Central Statistical Computing Centre (CSCC). It was important to have statisticians and IT people working together. When building the SDW, an outsource company was also involved as the technology used was very advanced.

9. The division of tasks was quite natural. All organizational, methodological, evolutional, informational activities were done by the Survey Coordination Division. The tasks were undertaken in cooperation with subject matter divisions and the CSCC. It covered in particular: concept preparation, designing, organization of work, and testing. CSCC was responsible for the technology used. It cooperated with OB during the design stage and evolutional work, maintaining programs, testing them, undertaking all administrative tasks, and coordinating hardware and software requirements. The wide range of the project meant that its implementation concerned all people engaged in the production process of statistics – designing, implementation, elaboration or dissemination of results.

III. STATISTICAL METAINFORMATION SYSTEM (SMS)

10. It was assumed that the most effective method is to build an SMS based on survey documentation. Such documentation concentrated on the Statistical Survey Schema (POS). Other important elements of the system were: the Glossary of Terms, Classifications and Administrative Data Sources System.

A. Statistical Survey Schema (POS) and Survey Documentation System

11. It was assumed that the list of elements, which constitute survey documentation, as well as their form will be standardized. Survey documentation is based on a description given of the Statistical Survey Programme for Official Statistics (PBSSP) and Statistical Survey Schema (POS).

12. By the gathering of survey documentation, we are referring to the topic described in PBSSP, as it is a legal description of the survey. The rules that apply to the creation of documentation are the following:

- The basic document for each survey is its description in PBSSP, which is a legal base for implementation of the survey,
- Each survey from PBSSP should have one or more positions in POS,
- The implementation of pilot surveys should be described in POS,
- POS is the general schedule of the surveys.

13. Survey documentation consists of:

- A survey description from PBSSP,
- The POS description of surveys, which are associated with a given survey,
- Annexes to some positions of POS:
    - The description of survey methodology, which combines some positions from POS into one survey topic. The POS position can include the whole topic from PBSSP or part of it. If the position from PBSSP is the same as in POS, the methodology and above-mentioned survey description are combined together,
    - Survey documentation is taken directly from PBSSP and its content is created every year. In the case, where some parts of the documentation were not changed, they are copied for the next year, however leaving the date of document creation unchanged. If modifications were done, the base information on such modifications are put in the survey description and details are put as directives for Regional Statistical Offices and interviewers,
    - Information about the survey is available for Regional Statistical Offices as sets of documents, eventually supplemented with necessary directives.

14. Within the framework of cooperation with Statistics Sweden (SIDA), the project on a documentation survey system was established. The project was implemented using the TQM method. The group working on the project prepared a concept schema, where survey documentation is part of SMS.

15. Preparation of the survey documentation process is described in the following sub-points:

- Director of author's division announces to administrator of SMS the name of the person (author), who will be responsible for a given survey and survey theme;
- Survey author prepares the project of survey methodology (survey description in SMS marked as temporary);
- Survey author enters changes in the Glossary of Terms and Classifications during work on the methodology;
- Organizational and subject matter supervision, which ensure coherence of elements in Glossary of Terms and Classifications, will be done by Survey Coordination Division (OB);
- Survey author prepares first version of forms – in forms database of the SMS;
- After acceptance by Programme and Methodology Commission, survey description is created by the system in PBSSP project;

- Elaboration(s) became visible in POS – author's division, together with OB supplement of general information (also with persons responsible for every task);
- OB prepares the last version of forms (in system database), which, after acceptance, will be copied to proper annexes of project Decree of the Cabinet,
- Schedule of work proposal is agreed with employees of OB and CSCC, entered to POS and accepted by authors,
- Schedule of work acceptance finishes work on POS and make its dissemination available (e.g. printing). The rules described above are used currently and will be transferred to the informatics' system directly,
- Persons, responsible for document preparation mentioned in POS, include the created files with following documents into system:
  - survey description;
  - assumptions for the creation of a list of entities for a given survey (frame BAKAR based on Statistical Business Register BJS);
  - assumptions for the informatics' system;
  - running the documentation of the informatics' system;
- Preparation of the above-mentioned documents may be done before formal acceptance and publishing of POS;
- Such documents may require acceptance, according to the rules, by many CSO units. All documents, accepted or not, will be visible for a given groups of users only.

B. Glossary of Terms

16. The goals for creating the Glossary of Terms were the following:
- To have a unique meaning of terms used in official statistics through the creation of a standard description of terms;
- Support for designing and implementation of surveys processes;
- information services for users of statistical surveys.

17. Types of subjects stored in the Glossary of Terms were:
- basic terms;
- secondary terms (calculated with a calculation algorithm);
- classifications' positions – chosen by survey author.

18. Currently thousands of terms are stored in the database and each of them is described through several object types. Structure of Glossary of Terms considers internal hierarchy between existing terms.

19. For the purpose of management of terms the following types of users were defined:
- term protector, which is responsible for definition and coherency of a given terms,
- subject matter administrator, who is responsible for coordination of works and content of database, where Glossary of Terms is stored,
- internal and external users, which have read only access to the terms. Range of terms available for external users is narrower.

20. The database model for classification purposes differs from the typical structure, where classifications consist of many positions. It was assumed that positions of classifications could be treated independently. It means that a given position can be linked to a few classifications. It is important in the case where classifications are somehow combined. For instance, one classification may be a subset of another. In searching a given term (position of classification) the user can achieve many results from different classifications.

C. Administrative Data Sources System (SMA)

21. Within broad range of SIS problems metainformation system on administrative data sources (SMA) was also elaborated.

22. Work goals undertaken on SMA system are the following:
- building of information system and creation of databases:
  - to have tools for analyzing coherency of official statistical system with administrative sources,
  - to elaborate model for data exchange between official statistics and administrative data sources,
  - to be able use of administrative data in efficient way,
- formalization of administrative data problems (metamodel),
- gathering and actualization of knowledge about administrative data sources.

23. SMA functions were defined as follows:
- gathering and storing data on administrative information systems.
- actualization knowledge base on administrative sources consistent with modifications of those systems.
- reports generation for the purpose of detailed analysis and estimation of quality of administrative data sources, usability in official statistics, current and perspective usage of administrative data sources.
- gathering information about range of data possible to use by statistics and methodological, organizational and technical conditions associated with.
- dissemination of information gathered in the system.

24. The architecture of SMA consists of:
- Databases describing administrative data sources,
- Glossary of Term for administrative data sources,
- Classification dictionary used in administrative data sources.

D. Implementation of the SMS

25. Building such a metainformation system required the proper environment and tools. For that purpose a server with Oracle software was purchased. It was also decided to buy the DDS system – INSEE tool designed for complete survey documentation. SMS will be accessible by all users of corporate CSO network.

26. SMS will generate different types of messages:
- event messages – e.g. form acceptation,
- time messages – defined time is over.

27. Messages will be sent to users or to activate system processes:
- running system tasks (management, errors),
- information on coming time limits,
- information on data availability.

28. SMS users are grouped according to their roles:
- survey author,
- subject matter administrator (for each subsystem),
- coordinator of survey frame,
- organizer,
- application designer, programmer,
- analyst,
- task manager,
- manager of external users,
- other users.

29. Access to some areas of data and metainformation depends on a person's role in statistical productions' system. Security contexts are defined allowing access to reading information as well as its modification:

- information status (private, first version, last version),
- type of activity or domain,
- form,
- territorial code.

IV. STATISTICAL DATA WAREHOUSE (SDW)

30. The goals of SDW were:

- to facilitate the analysis and secondary surveys based on distributed data sources through dissemination of the results of different reports from one place,
- to create a foundation for commonly accessible databases as a natural extension of databases accessible for statisticians,
- to disseminate data efficiently to permanent users including EUROSTAT, OECD and other international organizations,
- to elaborate efficient methods to secure data – from damage as well as from non-authorized access.

31. SDW is a set of data, which assists in decision-making processes, with the following features:

- subject oriented
  SDW structures are based on fact and dimensional tables. Fact tables include data on subjects, which are described in SDW and will be analyzed. Each subject is described by many attributes. Dimensional tables consist of attributes' values and their descriptions. Dimensions allow for grouping and selection of data stored in fact tables.
- integrated
  Integration means unique description of all data stored at SDW - in accordance with the metadatabase and internal homogeneity of data achieved by structures and procedures of SDW.
- with time regarding
  One of the main dimensions is time. Time intervals are different, depending on SDW specifics, but in any case allow for analyzing of data in time series.
- unalterable
  As opposed to relational data base systems, where data modifications and transaction are allowed, at SDW data should stay unalterable.

32. It is assumed, that DW will be used by the following users:

- Authors' divisions
  CSO divisions, linked within one LAN, will have the widest access to stored data, with respection of data protection (proper access rights). Individual and aggregated data stored at relational tables will be accessible as well as agregated data stored at multidimensional cubes. The typical two-layer connection client-server will be used (Oracle Discoverer client and Oracle Reports client – RDBMS Oracle).
- Regional Statistical Offices and other CSO units
  Users working at Regional SOs, which are connected through WAN, will have access to the data using intranet. Range of available data will depend on user access rights. Within intranet the following programs will be available: intranet versions of Oracle Discoverer and Oracle Reports, browser applications proparred by CSCC programmers.
- External users will not have direct access to the SDW. It will be created special Public Data Warehouse (PDW), feeding from SDW and other sources, which will consists aggregated data on proper level to protect individual data. Access to that database will be done through internet.

33. At the current stage on SDW development the following systems were loading:

- National Census NSP 2002,

- Agricultural Census PSR 2002,
- Economic surveys: C-01, DG-1, F-01/I-01, Z-03,
- Agricultural surveys: R-05, R-08, R-08G, R-09A, R-09B, R-KSRA, R-KSRB, R-CzSR,
- Architecture and tourism surveys: B07, KT-1, KT-1a

34. SDW is based on Oracle solutions and consists with described below subsystems.

Metabase

Metabase is essential part of SDW. All other SDW subsystems are tightly integrated with metabase. Metabase consists of two parts. System part includes description of system structures, format of data, inexes, views definitions, user access rights, users contexts, partitions etc. Statistical part of metabase consists of: Statistical Business Register BJS, Glossary of Terms, Classifications and MDane (description of data with metadata connections).

Operational base

The role of the operational base is to import data from external source systems. The source data (individual or/and aggregated) are eventually filtered, verified and transformed. Additional indicators, control tables and internal SDW keys are created at that stage also. Structure of operational base are very simple, some possible relations and indexes were excluded also.

Central base

The central SDW base has relational, simple structure (star schema with fact and dimensions tables). It was decided to exclude relations between fact tables. There were two main reasons of such decision: simplicity from user perspective and efficiency. Dimensions tables can be linked to several fact tables however. SDW has many stars; for instance for economic data: individual data and aggregated data fact tables. In National Census case structure is more complicated - stars were created for: buildings, flats, households, families and persons. Connections between those stars are done through specially created dimensions (junks – realization of ideas of Ralph Kimball).

Data marts

Data marts are thematic databases, stored within SDW, created for fast and frequent data searching for a given users (eg. for one division). Data are agregated according to the user needs. For the time being two data marts were created: for financial survey
F-01 and labour surveys.

Dissemination subsystems

As described above, for the purpose of dissemination of data client programms (Oracle Discoverer, Oracle Reports so-called fat clients) are used within LAN for access to RDBMS Oracle data. Within WAN only intranet solutions are available. It is possible to use thin Oracle Discoverer client, thin Oracle Reports client and intranet applications (special kind of browser). From any of mentioned above subsystems Excel or CSV text files can be created, which allow user for further data analysis.

V. FUTURE PLANS

35. As stated at the beginning, some activities concerning SIS were postponed to a later date. There were three main areas:

- data entry systems based on internet/intranet forms,
- data dissemination system through CSO portal,
- system for interviewers.

36. In June 2004 a feasibility study was done to elaborate a general concept for the creation of a CSO portal. The portal, which is to be built in 2006, consists of two parts:

- portal for reporting units (PS),
- portal for dissemination of information (PI).

A. Portal for reporting units (PS)

37. This part of the CSO portal allows reporting units to enter the data on electronic forms and send them through internet to the CSO. It is assumed, that about 10% of units, which are obligated to send data, would like to use such a method. Other units can use more traditional ways such as paper form or e-mail. Data coming in such a way will be entered by Regional Statistical Offices' operators, but application will use intranet as a transfer media and will be similar to that used by reporting units.

38. During the design of internet forms, all necessary information from SMS will be used. In particular only terms, which are included in Glossary of Terms, could be used and control expressions, which will be stored in SMS, will be used.

39. The heavy burden of data entering and checking will be undertaken by the reporting units, although additional verification of data will be done by the CSO server.

40. As on-line mode will be generally used for the transfer of data forms, off-line mode will be possible also. It is especially important for units, which would like to prepare data for statistics based on their internal systems. In such a case file will be created by reporting unit and transferred in off-line mode.

41. One of important problems, which must be resolved, is to assure that data will be transferred in a secure mode. Our system will use SSL (https protocol) and authentication done by login and password given to each reporting unit (or even more – reporting unit could have more than one login for different forms). Use of electronic signatures was discussed and generally postponed to the future, as it will require serious financial support.

B. Portal for dissemination of information (PI)

42. For the past 10 years, dissemination of statistical information is done through www pages. The creation of the portal allows for dissemination in clear and unique form according to the generated schema and style sheets (CSS). Publishing of information on the portal will be done not by the Dissemination Department only, but by many statisticians, which will have proper rights to publish the information. The portal will be managed through Content Management System (CMS).

43. On the portal data will be published also. Sources of data can be divided into two categories:

- data prepared and published directly by statisticians – mainly in Excel,
- data available in on-line or off-line mode and taken directly from Public Data Warehouse (PDW).

44. Data published on the portal as well as data taken from PDW will be connected almost in all cases with metadata information. Portal will have English version also.

C. System for interviewers

45. There are some surveys like the price survey or social surveys that require interviewers' activities. These systems differ from the above-mentioned portal for reporting units. Hardware is based on laptops and PDA, software and applications must be specially designed. There are two possibilities considered by the CSO (evaluation works started in 2004):

- a system based on MS SQL, which are installed on different platforms (server and mentioned above laptops and PDA) with replication of data,
- our own application based on exchanging files through internet.
