Distr.
GENERAL

WP.8
29 March 2010

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (i): Developing common high-level architectures

Toward Generic Systems at the Israeli Central Bureau of Statistics

Prepared by Shifra Har, Central Bureau of Statistics, Israel

I. Introduction

1. The IT department of the Israeli Central Bureau of Statistics supports more than one hundred statistical information systems, covering a wide range of subjects. Each year we are required to develop additional new systems together with the need to maintain the old ones. These requirements necessitate growing manpower and increasing costs.

2. Implementing a GIS unit within the Information Technology (IT) department enables us to combine the character data as well as the spatial data, but the integrated systems have become substantially more complex and time consuming, due to the need to integrate different technologies in various development tools. These changes in circumstances caused pressure on IT to conceive a new strategy for developing Information Systems.

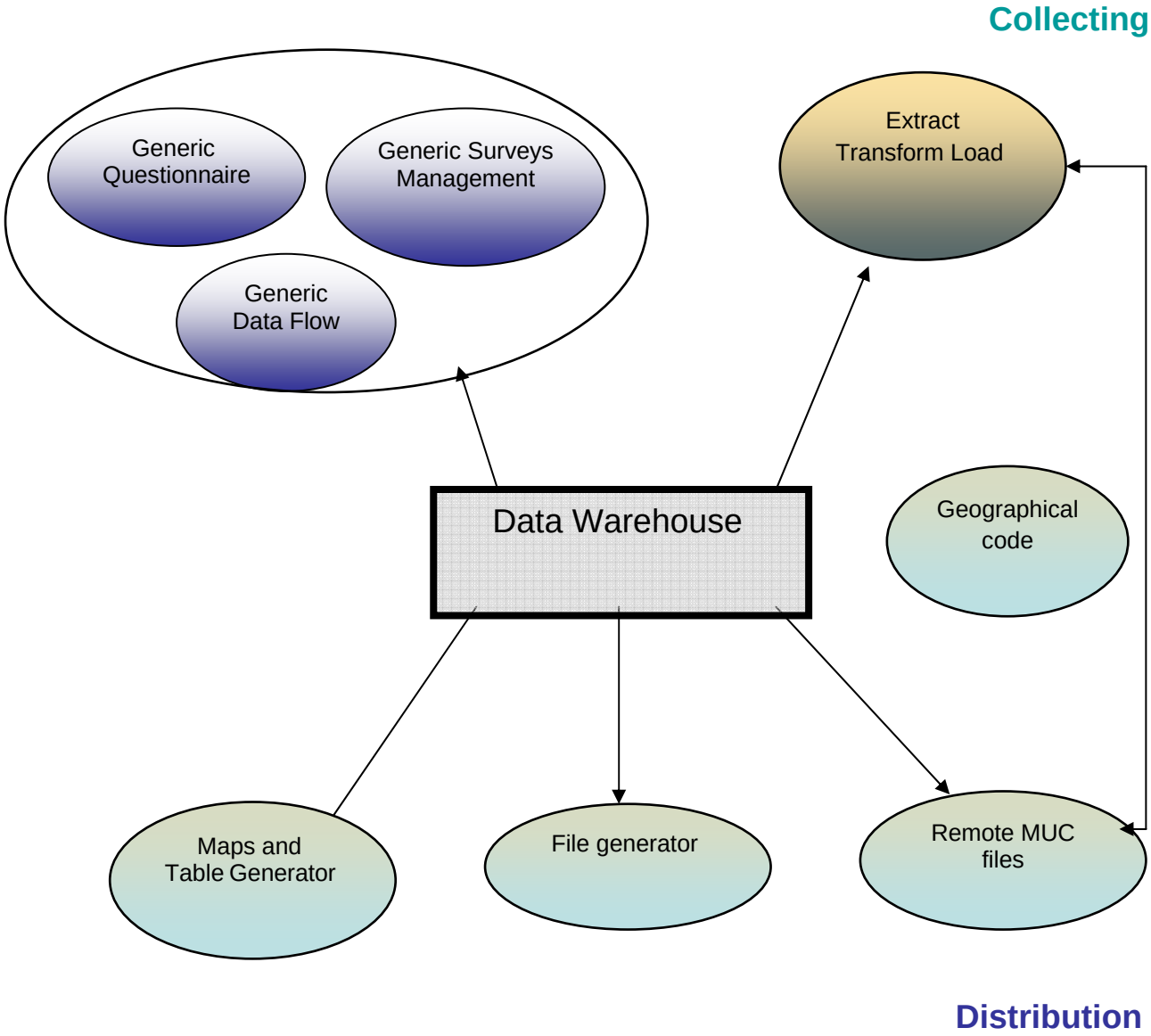
3. A partial solution was defined for more general systems. Our main goal was to shorten the creation time of a new information system by producing standard processes, standard metadata and standard documentation.

4. Brainstorming processes performed in the department divided the statistics activities into three main stages:
   (a) Data Collecting
   (b) Data Processing
   (c) Data Distribution

5. Each of these stages was analyzed for the potential of implementing a generic system. Six systems have been identified that could meet the goal:

- Two general systems were diagnosed in the data collection stage: Generic Survey and Extract, Transform and Load (ETL) for handling administration files.
- Three systems were diagnosed for the data distribution stage: Maps and tables generators file generator and remote work on micro data under control files.
- One system, the geographical coding and anchoring system, was diagnosed as an infrastructure system for the three stages.

Figure 1: The six systems that can bring us closer to the target will be described below:



II. Data collecting

A. Introduction

6. The statistical data is received from two main sources: administrative files and surveys. For each of them, we characterized the needs, and looked for "on the shelf" software that would enable us to use "as is", or alternatively, to be the base software adapted to our needs. Results of our searching indicated that there was no available software for the survey procedures but we found several tools that were capable on handling the administrative files.

B. Administrative files

7. Each year the CBS receives hundreds of files from many institutes and governmental offices. The files have different structures depending on which systems they were created, different descriptions and different definitions. To a single variable there are several categories that are not identical.

8. In the CBS, each department has its own autonomy to handle those files and document the processes. It was obvious that we had to find a way to standardize these procedures, including standardized metadata and document the processes.

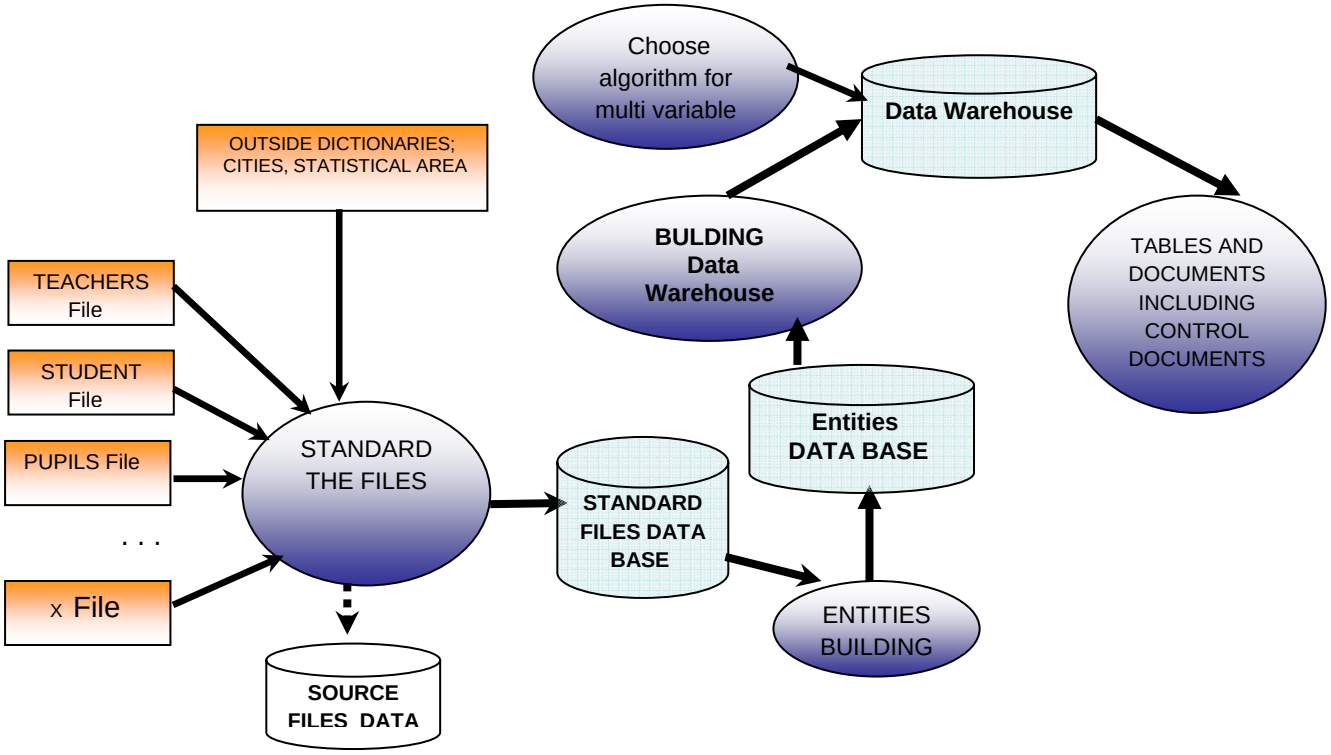
9. The Extract Transform Load (ETL) software would enable us to read data from several types of technologies (text files, xml files, excel files and databases) and convert the data from its previous form into the required structure, including databases. During these processes it is possible to check the data by using lookup tables, by combining the data with other files or cleanse the data according to our predefined rules. Finally, the data are loaded into databases at predefined entities. All the processes are "accompanied" by control tables and documentation.

10. The organization published a public tender that required the functionality we needed with an understanding that: there is no tool that has all the needed functions and that the provider will have to develop the missing features. Benchmarking the results, we measured the gap between our needs and the software capability. As we could not anticipate the future needs, one of the test results criteria included the possibility of add functions by utilizing standard development tools.

11. Several ETL tools competed in the tender: SAS, Data stage, I-Bolt and "Informatica". The latter won, as most of the required features were pre-embedded.

12. The tender was published in the end of 2009 and included two steps: the first was the installation of the tool and development of all the missing functions; the second step was the pilot to implement the tool on our education subject, including handling of 350 files obtained from educational institutions in one database arranged by the pre-defined entities i.e. student, teacher, institute, class, etc.

Figure 2: The first ETL project



C. Generic Surveys

13. "Generic Survey" is the project's name but it includes three separate systems: Generic Survey Managing System, Generic Questionnaire System and Generic Data Flow (transfer the data from the field to headquarters). The integration between them provides the global solution. It was decided to develop each of them separately in order to enable independent use in the old non generic projects.

D. Generic Survey Management System

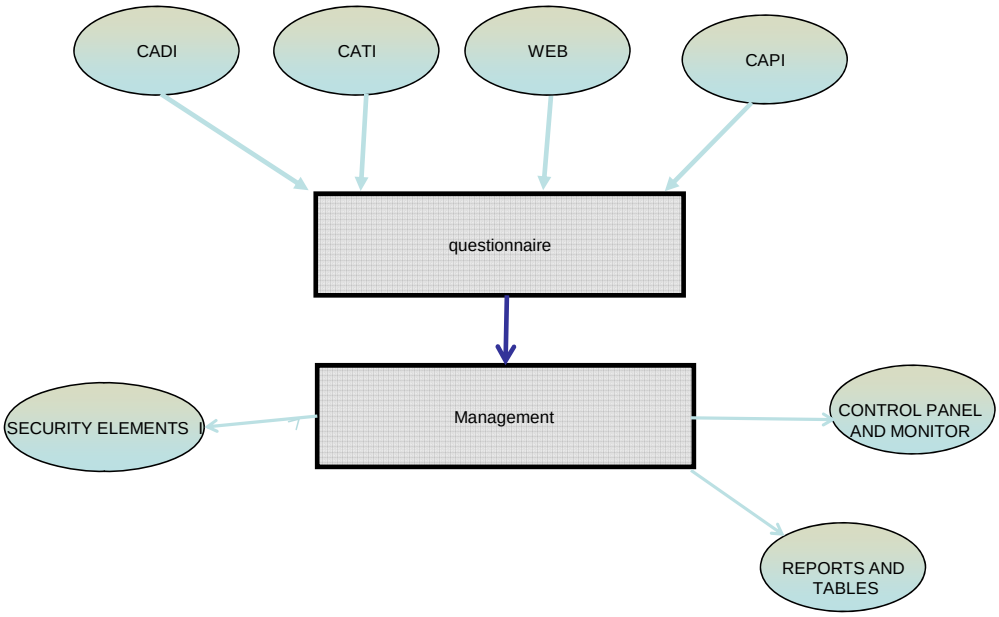
14. The CBS is developing computerized tools for a large number of surveys throughout many years. In most cases the surveys management system have the same basic functions. The differences between the surveys depend on the project manager and his personal point of view, according to the subject, technological changes and the capabilities and availability of human resources.

15. The entire information about the needed functionality was gathered from the field managers and IT developers that have vast experience in surveys. Our intention was to rely on past experience and reused procedures.

16. A generic management system requires unity in both working procedures and technologies. Our aim was to create a collection of procedures, so the project manager could choose and customize the survey from several given possibilities.

17. The generic system creates integration for all data collection types (CADI, CATI, CAWI, and CAPI) each system has an independent management with overall integrated mechanism

Figure 3 – the architecture of the Generic survey system

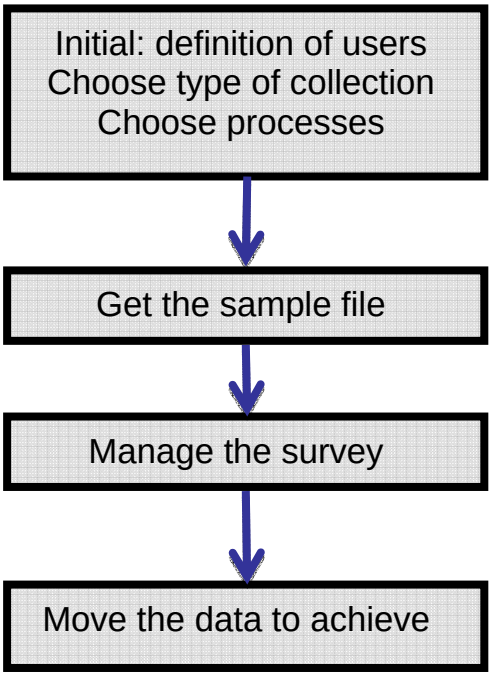


18. The components of the system are:

(a) Initialization of each type of collecting data
(b) Interfacing with the questionnaire
(c) Managing the survey
(d) Managing external and internal users
(e) Control panel – monitor
(f) Reports and tables
(g) Security elements

19. Two types of users were identified for each survey: external users i.e. sampled web questionnaires and CAPI enumerators, internal users i.e. headquarters' personnel, CATI and CADI enumerators.

20. The four stages required to generate a survey management system:



(a) Initiate the process, set parameters and define the internal and external users. Define the databases and update the dictionaries used in the survey.

(b) Obtain a standard structured sample file and check it.

(c) Customize the survey management parameter. The customization must be created for each of the collection methods and general administration. Customization takes place when the survey manager is independently able to choose the set of needed activities. Additionally, it allows the survey manager to initiate activities i.e. sending letters and messages to the surveyor by post, mail or responding to events that had been displayed through the monitor.

(d) The final step - check the questionnaires' integrity. The system enables an inspection of each sample in any stage, with the possibility to handle the multi questionnaires (WEB and CADI) problems. Move the data to archive.

E. Generic Questionnaire

21. Generic Survey Management significantly shortens the time needed to create a new survey, but every survey needs a specific questionnaire. Generating a specific questionnaire, especially in different languages is time consuming.

22. An idea to shorten the questionnaire development time lead us to create about 20 computerized designed templates of questions. The field unit will choose for every question in any survey, a designed and computerized template.

23. A system will be developed to allow the field unit to independently fill in the templates. The integration, the navigation rules and records' checking will be added in the second edition and in the meantime it will be programmed separately.

F. Generic Data Flow for the CAPI Surveys

24. The system is a multi protocol channel, allowing each of the CAPI surveys developed over the years to send the data to the center through a "Hi-way". This "Hi-way" allows several collection media i.e. laptops and hand held computers, to transmit the data via telephone or cellular lines to the CBS. The system includes components required for data security. In addition, we developed a unique monitor that gives the possibility of tracking each questionnaire from the laptop to the database inside the organization.

G. Generic coding and anchoring

25. Translating the addresses to geographical entities coordinates is one of the generic systems that are used as infrastructure for the three steps mentioned above. The system allows each of the subject units to independently use this system and to online view the coding result.

III. Distribution

26. The team presented several suggestions: Generic Maps and Table generators, a Generic File Generator from the data warehouse and a remote Micro Data under Control file for the researcher.

(a) Generic Map and Table generator

27. The ICBS developed and distributed two map and tables generators: Car Accident and Construction. These two generators were built on one template which can be the basis of a generic tool.

(b) Generic File generator

28. From the data warehouse, the last step of the generic survey system and the ETL system is to move the data to the Data Warehouse archive. Having those files with the standard metadata, enables us two important products: a dictionary and an index of all the variables existing in the Data Warehouse belonging to each file and subject. In addition, each employee that has authorization may choose any variable to create a new file to observe outside the system without a programmer's intervention.

(c) Working remotely on Micro Data under Control Files

29. One of the ICBS main goals is to allow the researchers' to utilize the CBS data while having reliable data security restrictions. At the present, the CBS has a research room located in the CBS main building, dedicated to external researchers who can use files authorized to them by the Data Security Committee. In the research room they can process and analyze the data on PCs with installed statistical software. Their output is under scrutiny of the CBS.

30. The CBS has already developed the first version of a virtual research room allowing researchers to securely connect from their own offices and homes and access the relevant data they were cleared to use. The researchers would be able to perform analytical procedures on the CBS servers. At the end of the research the results will be sent to them after appropriate inspection.

31. It is important to mention that the virtual research room is allocated on a separate network and should not be connected to the organizational network. Each research project will be isolated from the rest, with no connection between them. Even if a researcher has a

number of researches, he will not be able to view them simultaneously to avoid the possibility of matching between files.

32. In order to meet security requirements, the technology architecture is built on three tiers: The first refers to the user interface and its authorization and authentication. The second layer is the data processes and the third is the data itself. In this project we facilitated new technologies including "Virtual Machine" (VM) and Virtual Desktop Infrastructure (VDI).

IV. Summary

33. The ICBS started to develop the mentioned systems in parallel during 2009. At the present we are in various stages on each of the subjects listed above.

34. The Generic Managing Survey - The first survey has already been implemented in the new "Generic Survey Managing System" (without CAPI). Seven new surveys are planned until the end of Aug 2010. These eight survey when run in parallel (in the first year), will probably prove the tool benefits. The CAPI tool's method for collecting data in the system will take place at a later stage.

35. According to our declaration in the beginning of the development, there is a need for only one month to generate a new survey management system with one condition: The questionnaire has to be utilized with one of the following developing tools: Blaise or Dot.Net.

36. The Generic data flow - is already under production with the "wages in construction" Survey. Next month the social survey and labor force survey will run on the same platform.

37. The first version of the remote MUC files has already been checked and we are waiting for the first researcher.

38. At the end of February we shall release the first version of the table generator from the Data Warehouse.

39. Our timetable for a new definition of the geographical code and anchoring brings us to the end of 2010.

40. The ETL project has been postponed due to legal problems. The company that lost the tender submitted an appeal for a request to check the tender action, so hopefully by the end of February we should be able to start our project.

41. With reference to the generic questionnaire, we are in the first steps of the project. Our intention is to continue during 2010.

42. Regarding the Tables and Maps generator we decided to test the Australian "Superstar Systems" prior to making any decisions.

V. Conclusions

43. We are in the early stages of using a generic system and encountering the problematic aspects. "A **Single Point of Failure**, (SPOF), is part of a system which, if fails, will bring the entire system to a halt.

44. There is no doubt that the establishment of generic systems saves valuable time in developing information systems. However, we must take into consideration that consolidated information systems means consolidated hardware too, thus the mechanism becomes more complex. The system must run continuously as numerous surveys run in parallel. This reduces the opportunity of handling systems, upgrading tools, maintaining and repairing when necessary. Therefore the establishment of such a complex system requires high availability in the network, hardware and software application.

45. Understandably, such modifications have to affect the users' habits. Instead of dedicated programs developed specifically for a small group of users, they will have to adapt themselves to the general systems and for that concept we have very solid support from the Chief Government Statistician. The outcome of this inconvenience will result in improved and significant Return of Investment in cost and time.