Distr.
GENERAL

Working Paper No.13
20 April 2007

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

Meeting on the Management of Statistical Information Systems (MSIS 2007)
(Geneva, 8-10 May 2007)

Topic (ii): Statistical information systems architecture

DEVELOPMENT OF IST SYSTEM FOR STATISTICAL DATA PROCCESSING
Supporting Paper

Prepared by Branko Jirecek, Republican Statistical Office of Serbia, Serbia

I. INTRODUCTION

1. This paper was supposed to be presented last year, but due to changes resulting from the transformation of my country, it was impossible to attend and present it. With apologies for the fact that some things may be repeated, this is the state in which this project is at this moment.

2. For a long time data proccessing in statistical systems was (and still is) organized on a stovepipe principle, which results in a variety of often incompatible solutions dealing with similar problems. A more general approach was designed and implemented in the Statistical Office of Serbia and Montenegro, which deals with proccessing of all the statistical data in a completely different manner. As the above-mentioned Office ceased to exist, this approach, together with all the program-related and IT solutions, was transferred to the Republican Office of Serbia, where it is used and continues to develop.

II. THE "USUAL" WAY OF DATA PROCESSING

3. The most usual way of processing data of a statistical survey is a, so called, stovepipe principle, where the complete process is performed – from data entry to publishing the final results separately for each survey. Processing of survey data consists mainly of the following three phases:

(a) Data entry;
(b) Logical control (data "cleaning");
(c) Tabulating (dissemination, publishing).

4. The development of IT resulted in various surveys having over time various popular implementations. Depending on many factors, such as the importance of a survey, (un)willingness for changes, etc., we had (and still have) the situation where various surveys are handled by IT in a huge variety of ways – from surveys

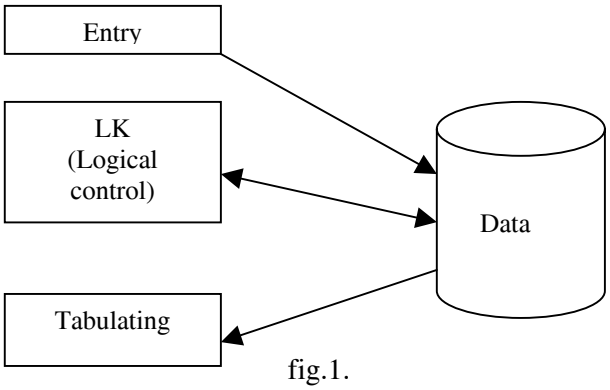

fig.1.

where data is stored in "plain" sequential files to those where they are in (unfortunately) databases of various kinds.

5. For a long time, IT was exclusively done on the IBM mainframe platform(s). Naturally, the development of PC-s caused some surveys to be migrated to the PC platform, and, of course, we have mixed solutions, just to make the situation even more complicated.

6. Therefore, we have a situation where the processing of statistical data is done on various platforms, implemented on a variety of software tools, with data stored in various ways, separately for every survey, and even worse, for every phase, as shown in fig. 1. This is a very simplified picture, which is done deliberately, because this was the biggest issue that we found all the surveys had in common and from where we had to start.

III. "AND NOW FOR SOMETHING COMPLETELY DIFFERENT"

A. Basic IST concepts

7. It appears to be very interesting that, although we contacted relevant people from several statistical offices to solve the above-mentioned problems, no one seemed to be interested. Hence, we had to start from scratch on our own. The only exception we found was EUROSTAT's Euro Trace, but, besides the fact that it is not able to deal with most of our surveys, we also heard about it too late, since by that time we had already developed the main foundations of our idea. Nevertheless, it was very helpful to exchange experiences with the people working with Euro Trace.

8. The worst of all is the fact that during all these years we have been developing, more or less, just three applications related to data entry, logical control and tabulating. We felt that it was high time for these three to be finally completed so that we would not ever (at least in the way we did until now) have to worry about them again. The aim is to develop a process of handling statistical surveys with the least amount of programming effort possible, or even better, with no real programming at all.

9. Although there is strong opposition to universal applications, we had some rather successful experiences in the past. We would like to mention TABS, a universal tabulating program developed for the IBM mainframe in the Federal Statistical Office and GoDar, a data entry and logical control system for the same platform, developed in the Statistical Office of the Republic of Serbia. Although both of these solutions were outdated a long time ago, we were positively encouraged with the existence of this kind of tradition.

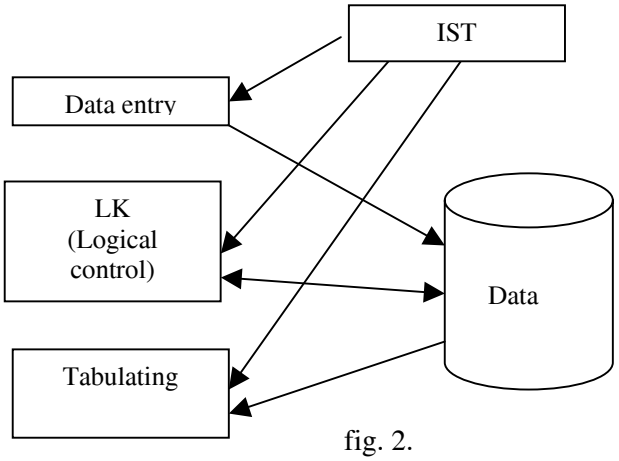

fig. 2.

10. It was obvious that data had to be stored in a relational database. The problem was which database (DBMS) to choose, because we already had two (IBM's DB2 and Microsoft's SQL-Server) in use, without any knowledge about whether this number would expand. Due to this reason we had to make our solution independent of the physical implementation of the database, which, on the other hand, limited us to the use of only the most basic set of SQL commands, with absolutely no enhancements or improvements of any DBMS, which, unfortunately, excluded stored procedures, triggers and all of this more useful stuff.

11. The main idea, as shown on fig. 2, was to try to organize the surveys to take care of their own data and to process it, instead of the current situation where every survey is a world onto itself. Technically speaking, IST as shown in figure 2, is a database where all the survey relevant data, both operational (such as data descriptions, conditions for logical control, definitions of the output tables) and descriptive (methodologies, field descriptions, etc.) is stored. The first "test rabbit" was the TU-11 survey in the field of tourism statistics, that once all expected and unexpected troubles were addressed, was implemented successfully and is presently running on the IST system.

12. However, there is still a lot of work to do on the whole idea, many things need to be improved, written and/or rewritten, but we are glad that it has been brought to an operational level, which means that the whole idea was correct.

B. "Inside" IST

13. The IST is mainly designed as a meta-data database "handler". The whole IST database is, as mentioned before, made in that way.

14. Although it may look complicated, the whole IST database is rather simple. Here are the main tables:

(a) **IST** Table that consists of global information about particular surveys. Here we have some global information, such as the full name of the survey, database(s) where data includes the periodicity of the survey, paths to the methodological texts, queries for output tables, etc.

(b) **ISTBaze** Table with the real connection strings, which the application uses to connect to the real databases and data within them.

(c) **ISTPolja** Definition of the data fields of surveys with accompanying properties, such as the type, length, eventual bounds with some consultation tables etc.

(d) **ISTLK** Logical Control definitions, or in other words, a list of possible errors and their definitions.

(e) **ISTTABS** List of output tables. As a matter of fact, this is a list of names of queries and eventual Excel, Word, SAS etc. files that the application uses to make an output table.

15. At the moment, logical control criteria and queries for output tables are more or less limited to pure SQL. Although this should not have to be a real problem, it may be difficult or inconvenient to manage some things with only pure SQL, so certain efforts are already being made to develop functions for some of the more usual situations, and even a bigger effort will be made in, probably, creating a little language that will ease this process.

16. As a complete solution for data entry already exists in the Republican Statistical Office of Serbia, the existing solution was included in IST for the time being. Work on a better way to handle data entry has started, but there are no results in this moment.

17. It seems reasonable that IST should consist of, at least, two components – one for the end-user (statistician), and the other for the IT specialist who would "install" a survey "into" IST. At the moment, only the first part exists.

IV. IMPLEMENTATION

18. IST, as already mentioned, is created with the intention to work with any relational database. This also stands for the IST database. Maybe it would be better to say that it works with any number of databases of different kinds, because there is no limitation that data must be in only one database, the data from one survey can be in different (even physically different) databases, which is not unreasonable to expect, because this way we can plan to have unique registers, classifications and other "consulting tables" for all surveys.

19. There was no doubt about the operating system on which IST would function. It had to be made as a Microsoft Windows application, because all the infrastructure of the Republican Statistical Office of Serbia is

running on this platform. There were certain doubts about how Windows would communicate with the IBM mainframe and, especially, the DB2 database, but, fortunately, this works very well.

A. Programming language

20. There was much more of a dilemma about the programming language in which this would be implemented. Initially, the decision was made to do it in Microsoft Access, or specifically, in its VBA language. The main reason for this decision was that Microsoft Office is already installed on every PC in the Statistical Office, and users (statisticians) are somewhat familiar with it, so they may be able to do some elementary operations in Access (certainly much easier than in, for instance, DB2).

21. The user (statistician) is supposed to choose the survey he wants (and has the right) to deal with, and either input data, logically test and clean it, or produce outputs. The logical control phase is the most interesting one, where the user was allowed either to test and change the data "locally" (on his PC) or "globally" (directly in the database where the data is stored, enabling more people to simultaneously perform this operation).

22. And yet, all of this was abandoned when it came to the implementation in the Republican Statistical Office of Serbia, where the whole thing was converted to Microsoft Visual Basic 6, which was their strategic language. Nevertheless, this conversion was a good chance for rationalizing the whole project, so it was all rather simplified from the users' point of view, several new possibilities were added, and everything was made more robust.

23. This (Visual Basic) version is currently running in the Republican Statistical Office. However, a process of a further conversion to Microsoft's .net environment is being done at the moment. The final goal is to make IST a completely web-oriented application.

B. DBMSs

24. IST is basically supposed to be independent of the database platform, so it was interesting to see how it would work in real life. After the first test with TU-11, which ran on Microsoft's SQL Server, three surveys were set up in the Republican Statistical Office of Serbia, using IBM's DB2 on a Z800 mainframe. Also the IST database was migrated to DB2.

25. As expected, different database products have incompatible SQL dialects. Except for some minor solutions, this problem stands. The most uncomfortable thing that had to be dealt with was the absolutely unexpected hostility of the DB2 product towards programmers – for instance, DB2 does not allow joins inside an update statement, so every relatively simple update had to be turned into a rather long and quite non-understandable query. The situation is even worse in creating queries for output tables - DB2 forces you to write rather long (and complicated☹) queries, so in the end the limit of 32K for a SQL query begins to be a problem. This certainly is not the subject of this topic, but I simply couldn't resist not to air a grudge and complain a little☺.

C. Samples

25. One of the surveys currently "running on" IST is PO-512, "Cattle sample", which is a survey based on a sample. Although discussions about how to generally deal with sample surveys will continue for a long time, the experience of how the IST apparatus dealt with this was more than positive, which also was a confirmation that the general idea is good.

V. THE USER'S VIEW

26. Beside the main IST goal of organizing and uniformly dealing with all statistical data, one of the things that had to be achieved was simplifying the job of the end-user (statistician). So, let's see how it looks from the user's point of view.

27. After identification, the user has to choose the survey and the time period he is intending to work with. Then the next step can be cleaning the data or making outputs files.

28. After pushing the button for logical control, the application identifies all the errors in the material. Cleaning the material is made through a grid (grids, if there is more than table with data), where the statistician sees the material and the list of logical errors in the row on which he is positioned, and has a chance of easily changing the data in the grid by simply typing over it. After changing the data, the logical control is done for that row, so he can interactively see the results of the change(s). There is also a possibility of invoking the data entry program and changing the data that way. The statistician, of course, can choose which columns and rows he wants to see in the grid(s). All that a programmer has to do is to make a list of logical errors and define them with a little "enriched" SQL language. No real programming to do.

29. Output tables can be created in two ways – the application enables the statistician either to create ready-made tables or to create his own custom tables.

30. The creation of output tables functions relatively easily – on the left is a list of all possible fields and the statistician's job is to select fields from the list for the resulting table's headings, fore-columns or as an expression that will be summed. The application generates a SQL query, and, after execution, the table with the results of the query can be exported to an Office application.

31. Some output tables are made very often, so it seems like a good idea to make them once and just call them when needed. The statistician just has to choose one of the offered ready-made tables and do with the resulting table whatever he wishes. Behind the curtains, the application executes a pre-set file with a SQL query (or several queries), exports this to Excel, eventually executes an Excel macro (if there is a need for it), exports all of this to Word and, by running a little Word macro, combines a pre-existing Word table heading with the data. The result is a printable Word document. The interesting thing is that the programmer has only to write the SQL query(ies), select the table headings and, eventually, create an Excel macro. No real programming needed☺.

VI. THINGS YET TO BE DONE

32. First of all, IST must evolve over time to a true web-oriented application. Steps are already being taken, but this will surely be a rather long process.

33. The data-entry part has to be done. Work has started, some ideas are present, but it's yet to be seen how this will be solved.

34. Solving the "time-dependence" problem – consulting tables alter through time (for instance, new counties are formed, all classifications get changed now and then, etc.), even surveys change – all of this works in the present, but won't work for the same survey in the time before the change occurred. A solution is within reach, and doesn't seem to be as complicated as it seems.

35. Definitely, a little "language-oid" will have to be made to upgrade and unify various SQLs and make the life of programmers much easier.

VII. RESUME

35. Although there is still a lot of work to be done, we are satisfied with the progress we have made so far. The whole system is functioning in the Republican Statistical Office of Serbia and was chosen to be the future strategic tool for processing statistical data.

36. We are open for all sorts of discussion and are willing to cooperate with other statistical offices in this matter, hoping to share experiences and gain further knowledge.
