Distr.
GENERAL

Working Paper No.20
24 April 2009

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 2009)
(Oslo, Norway, 18-20 May 2009)

Topic (iii): Architecture

CBS ISIS: ARCHITECTURE FOR SURVEY PROCESSING

Invited Paper

Prepared by the Central Bureau of Statistics, Croatia[1]

I. INTRODUCTION

1. Development of an Integrated Statistical Information System (ISIS) at Croatia's Central Bureau of Statistics (CBS) started in September 2006 and by now is ready for user-testing / implementation. The project was preceded by a metadata development project which resulted in the central metadata repository named CROMETA. CROMETA was necessary as the fundamental platform for the ISIS which in turn provides CBS with an automated metadata-driven statistical processing system.

2. The aim of the ISIS is to enhance all aspects of statistical production i.e. data capture, data validation and cleaning processes, data warehousing and presentation of statistics to the public. The metadata repository, containing descriptions of all data, activities and processes in CBS provides the fundaments of such a system.

3. In practice one general SW solution was developed consisting of modules for particular statistical business processes.

4. The basis and precondition for the automation is well structured metadata, entered and maintained by the owners of studies/statistics. Furthermore, the processes work in the opposite direction as well i.e. processes triggered by metadata produce other metadata to reflect the current processing status for monitoring.

II. STATISTICAL BUSINESS PROCESS

A. Survey Processing in CBS

5. The majority of CBS´s statistical surveys are processed in following steps:

[1] Prepared by Maja Ledic Blazevic (majalb@dzs.hr) and Branka Cimermanovic (BrankaC@dzs.hr).

(a) Planning all activities
(b) Survey design and description
(c) Data capture and file transfer
(d) Validity checking against preset rules and producing error-list to be presented to the statistician in charge
(e) On-line correction
(f) Tabulation, i.e. producing statistics for statisticians supervision
(g) Publishing, i.e. producing first releases and other statistics
(h) Archiving
(i) Monitoring

6. It should be noted that the majority of processes mentioned above refer to the *data processing* phase of any particular statistical survey or, in other words, to the statistical survey as seen from the perspective of the IT sector which is responsible for providing adequate system solutions for all statistical activities within the organization. Mapped to the Generic Statistical Business Process Model (Level 1)[2], the steps mentioned above fit the model as follows:

<table>
<tr><th>Generic Statistical Business Process Model</th><th colspan="2">CBS's Survey Business Process</th></tr>
<tr><td>Need</td><td></td><td rowspan="8">(i) Monitoring</td></tr>
<tr><td>Design
Build</td><td>(a) Planning all activities
(b) Survey design & description</td></tr>
<tr><td>Collect</td><td>(c) Data capture & file transfer</td></tr>
<tr><td>Process</td><td>(d) Validity checking
(e) On-line correction
(f) Tabulation</td></tr>
<tr><td>Analyze</td><td></td></tr>
<tr><td>Disseminate</td><td>(g) Publishing</td></tr>
<tr><td>Archive</td><td>(h) Archive</td></tr>
<tr><td>Evaluate</td><td></td></tr>
</table>

B. Survey Management in CBS ISIS

7. The mentioned responsibility of the central IT department to provide and process all solutions for all statistical activities has two main drawbacks: the production depends heavily on the IT staff, as they are the ones that create surveys, enter parameters, produce tables etc. and this in turn causes considerable time delays since the surveys are processed centrally, basically according to the first in – first out principle.

8. Therefore, implementing the new ISIS meant a significant shift in the approach to the survey processing in CBS. The idea is that statisticians should be able to produce and change the survey processing 'jobs' themselves, without assistance from the IT staff, something that should speed up the production of statistics significantly and also increase the quality of statistics in general. Underneath is one general solution that is expected to manage 80% of the entire CBS survey stock. In relation to that it is essential that all new applications are equipped with intuitive and user-friendly interfaces.

9. While the CBS' survey business processes remained the same, a new architecture, new technical platform and a new appearance has been implemented with a number of new functionalities and nice-to-have features. The environment is customized for multi-user access where all users are entitled only to specific, previously defined activities. It should be noted that a significant extra effort will be required from statisticians to get used to the new methodology of survey maintenance and especially to provide all

[2] UNECE Secretariat: *Generic Statistical Business Process Model*, Version 3.1; December 2008; Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS)

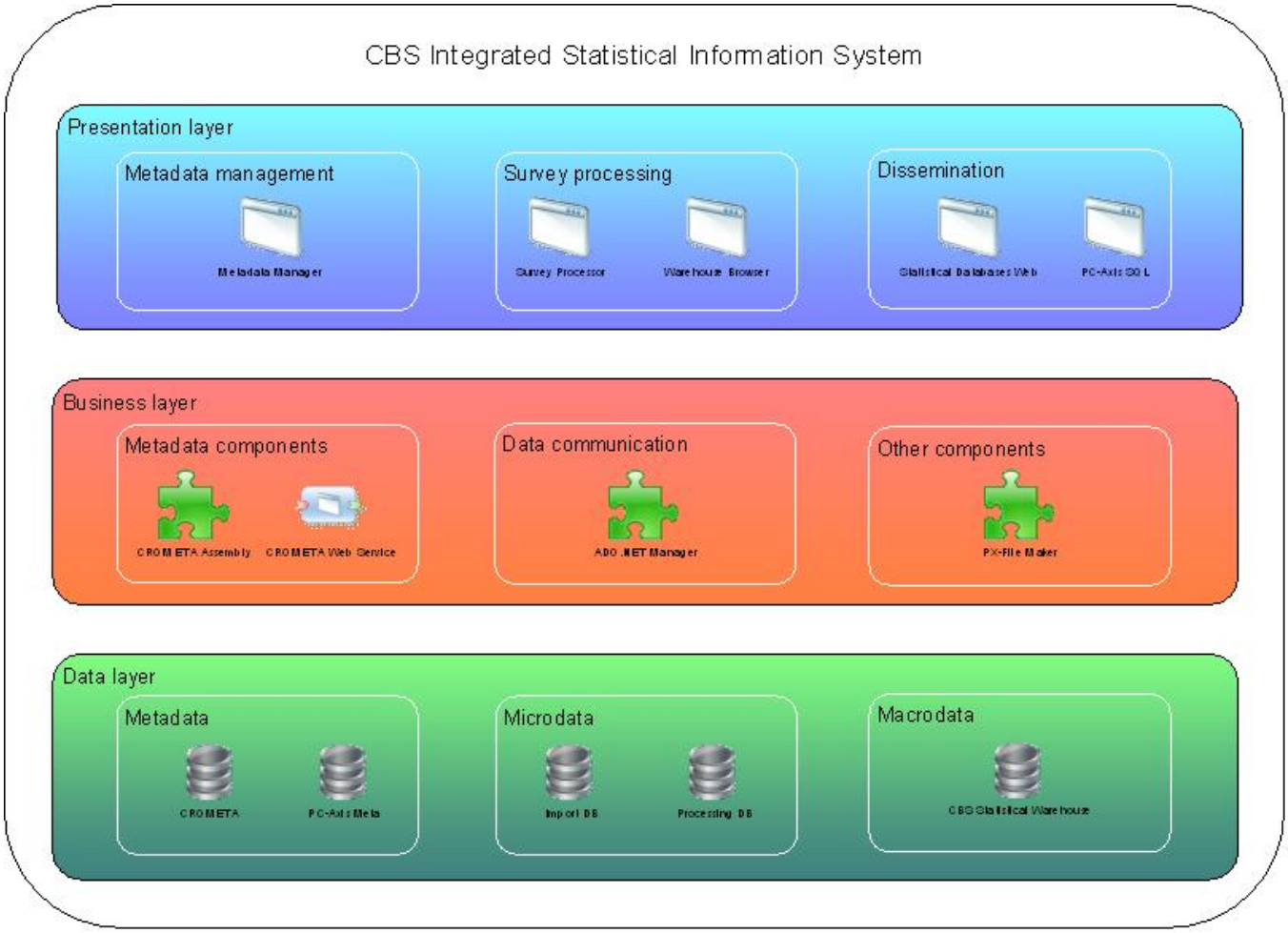
necessary metadata. It is expected that some statisticians will regard the new system simply as extra work, while some will, hopefully, gladly meet the challenges and benefit from that.

10. It is worth mentioning that the whole system is multilingual; the interface itself could be customized for any language and languages desired for entering and maintaining data and metadata could be defined through the interface. This is a very important functionality for a possible solution sharing.

III. ARCHITECTURE OF CBS ISIS

A. System Layers

11. The architecture of the CBS ISIS solutions follows common techniques for modern system development, fronting a multi-tiered, scalable solution, customized for multi-user environment. Generally it consists of the data layer in the bottom, keeping a business layer on top managing the business logic of the system. On top of the business layer is the presentation layer, dealing with user interaction, providing several application interfaces for metadata maintenance, survey processing and macro data dissemination.



12. Data layer is basically divided into two tiers. The physical storage tier consists of several databases where three types of data are stored. Metadata is stored in central metadata repository CROMETA and partially in specific PC-Axis Meta database. Microdata (individual data, registers) is stored in databases used for data import and survey processing. Finally, the public macrodata (aggregated data, cubes) is stored in CBS statistical warehouse.

13. The second tier in data layer is the access tier that consists of numerous stored procedures on top of the databases. The complexity of metadata model consequently ended with a large number of stored procedures that had to be defined in order to access and retrieve metadata correctly.

14. Business layer could be divided in three tiers: data access, business objects (metadata) and other components. For data access between data layer and the application the standard component ADO.NETManager[3] has been used. In the case of metadata access there is a difference and an assembly is used in every application that needs metadata from repository. The assembly contains all the metadata objects as described in the CROMETA model and information about the physical implementation of the repository. Additionally, the data access to metadata repository is directed to a web service that manages all execution and results compilation. Every application that has to access metadata repository can use only the web service for that purpose and thus the security of the metadata system is increased.

15. The other components in the business tier are used for some specific purposes, i.e. PX-File Maker assembly that can be used for creating PX files in the Microsoft .NET environment. This component was developed in 2005 for the needs of the future data dissemination through CBS web site.

16. Presentation layer has only an interface tier consisting of customized applications/tools developed in CBS and PC-Axis, as the main dissemination tool chosen in CBS. Customized tools are thin client solutions, i.e. the application is installed on the client computers, while the processing and execution are carried out on application and database servers.

17. The main tool for metadata administration and maintenance is Metadata Manager. Simple interface but very complex background of Metadata Manager is intended to be used by metadata experts. Survey processing is supported by two applications developed by CBS, Survey Processor and Warehouse Browser. These tools are aimed at IT and subject matter users who do not need to have the expert level of metadata knowledge.

18. Beside the PC-Axis as CBS's dissemination tool for aggregated data a new prototype of the CBS web site is developed to access it from the intranet/internet. This web application should allow users to create their own tables and search the existing publications. However, this prototype uses metadata from PC-Axis Meta database for table creation and has to be integrated with the CROMETA repository.

B. Technical Implementation

19. As regards technical platform and development environment, the conceptual modeling of metadata model and processing database was done by using Sybase Power Designer 9.5. The same software has been used to generating the physical data model that has been implemented on the Microsoft SQL Server 2000 RDBMS[4] and later transferred to Microsoft SQL Server 2005.

20. All database development has been carried out on MS SQL Server platform and applications have been developed in MS .NET environment in Visual Basic .NET (final versions in MS Visual Studio 2005).

IV. APPLICATION

A. ISIS Tools

21. As mentioned before, there are several customized tools developed in CBS that support different aspects of ISIS: metadata maintenance tool Metadata Manager, survey processing tool Survey Processor and a tool for data tabulation/aggregation Warehouse Browser. These tools enable users to process a survey in an

[3] Developed by Conscious KGCS, www.conscious.se
[4] Relational Database Management System

automated fashion so they are guided in a very easy and natural way from survey design to output results in form of aggregated data and ready-to-use tables.

22. The following picture shows user entry points to the different ISIS components, presented in the form of a flow diagram of the applications and their sequence in the survey processing.

CROMETA Metadata Manager:
Define your survey using the metadata maintenance tool. Copy metadata from an existing version or start from scratch.

Survey Processor:
Process the study in an automated fashion – taking advantage of modern techniques and computer-aided functions.

Warehouse Browser:
Create cubes or ready-to-use tables and customize reports in various formats for printing or on-line publishing.

23. The core of the automated metadata-driven solution of the ISIS is the central metadata repository CROMETA. Metadata Manager provides standard methods for adding, editing or deleting any kind of metadata in the system, covering eight of nine sections of metadata defined in CROMETA model (classifications are still excluded). Surveys are described by the concept of study which includes them and other statistical activities as well. The study is an umbrella for different versions of a study defined with corresponding reference periods, while there may be one currently valid version at one point in time.

24. The basic metadata information about a survey version is all what is needed for its further processing, but the design and description of a study/survey can also be described with adequate metadata (questionnaires, registers, context variables etc.). Thus, not only the metadata about the survey itself can be added by Metadata Manager, but all survey design metadata by adding new or copying existing metadata for new versions of surveys from previous ones.

25. When essential metadata about a study version is added, users can process it by using Survey Processor tool. The access to a specific study and its versions is also described with metadata. Users can access their surveys at the level of their access rights, which are handled by the Metadata Manager tool. These rights can be limited not only to a specific study version, but to a specific process, i.e. population definition, data import, data cleaning, etc. Survey Processor consists of several modules corresponding to the main statistical processes identified in a standard survey processing.

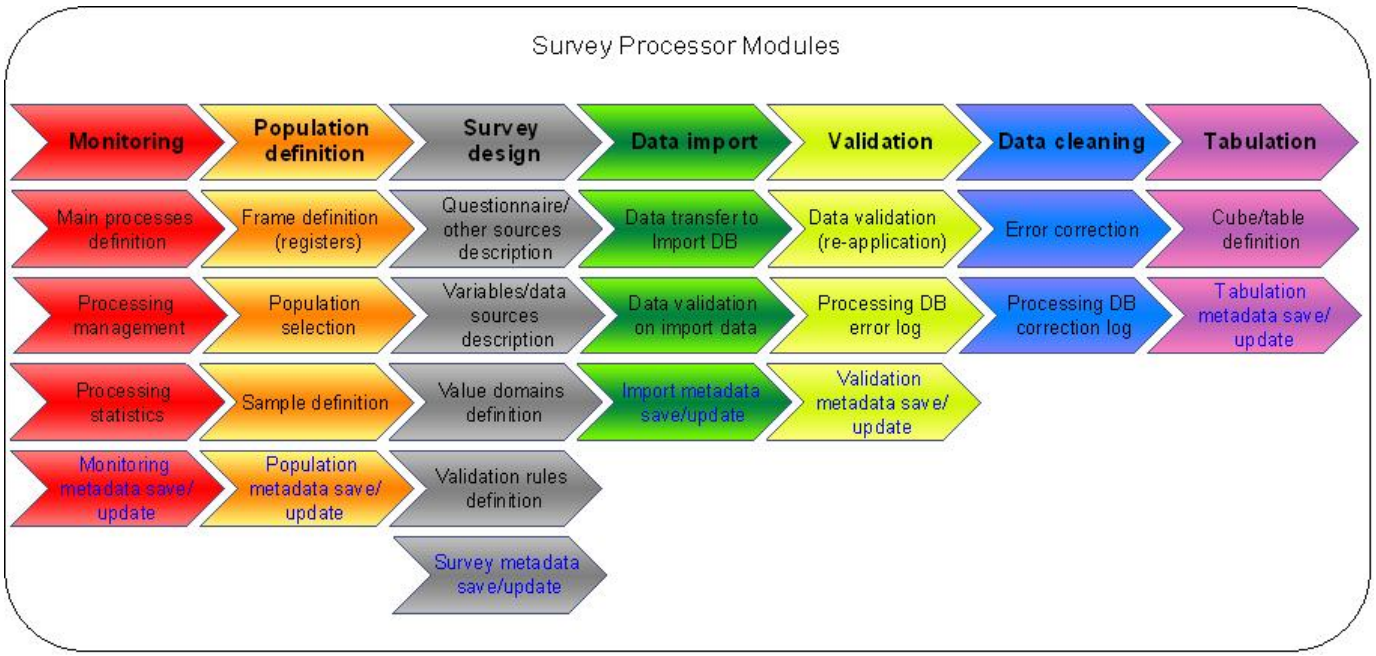
26. The modules of the Survey Processor will be described later, but one of them is actually developed as a stand-alone application. The module for tabulation is presented in the form of a Warehouse Browser tool which enables users to produce result tables from processed data. The results can be saved as cubes of aggregated data and/or ready-to-use tables, both stored in CBS Statistical Warehouse. Metadata about aggregated data is stored from Warehouse Browser to the central metadata repository, while initial metadata about surveys, registers and rights for tabulation is retrieved from there as well.

B. Survey Processor Modules

27. As mentioned before, each module was developed to cover a particular phase in the statistical survey processing. So far, the developed modules of the Survey Processor are:

- Monitoring
- Population definition
- Survey design
- Data import
- Validation
- Data cleaning
- Tabulation

28. The following picture shows modules of the Survey Processor, along with the main functions that each of them provides. The function marked with blue letters presents some sort of a metadata transaction in that specific module.



In the future there should be more modules included, i.e. automatic coding, manual coding, publishing, etc.

29. The **Monitoring module** is intended for creating main processes for the survey and their management. This module allows starting, stopping and ending each of the phases, in case the data is ready for that particular phase, e. g. data cannot be tabulated before the tabulation process is started. Some basic statistics about processing is also presented in this module, i.e. number of respondents, rate of response, error statistics, data correction statistics, etc. It should be stressed that relevant process management information is stored to central metadata repository.

30. The **Population definition module** is intended to create the survey population from a frame that can be selected from statistical and/or administrative registers and populations from previous study versions. The selection is supported with features like filtering, that allows subject matter people to search their specific respondents from large data sets such as statistical business registers, or adding specific information to population units, such as contact information, etc. All definitions are stored to the metadata repository as well.

31. The **Survey design module** contains many features; from questionnaire and other sources description, variable and data sources description up to validation rules and value domains definition. User-friendly interface helps users to easily create and describe structures corresponding to their questionnaires and data sources, if they are not already copied by the Metadata Manager tool. Every part of survey design is stored then as adequate metadata information in the central metadata repository.

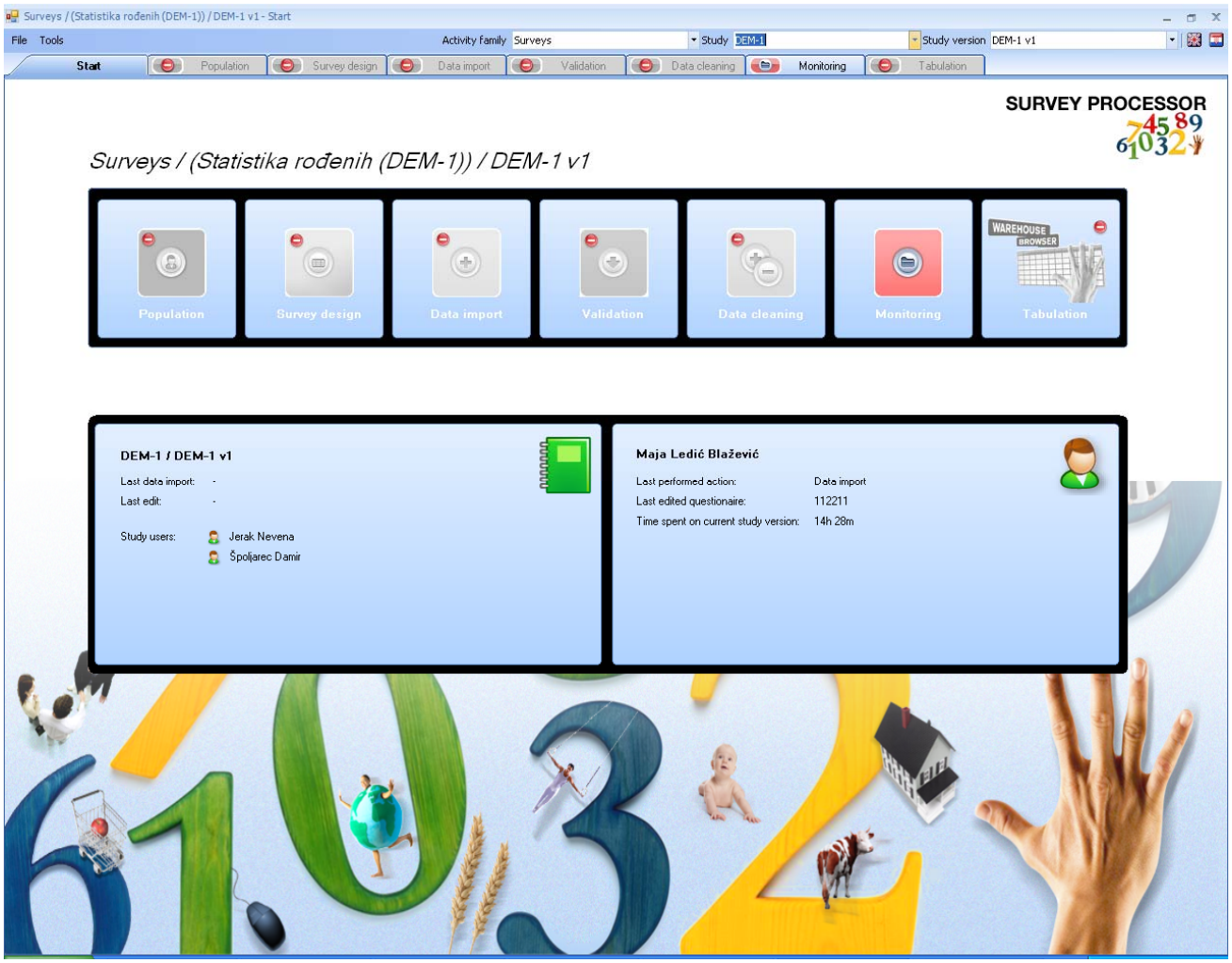
32. The **Data import module** is used for importing data in several data formats (delimited, fixed-width, etc.). During the import data is checked and cleaned in several phases. All mirror duplicates are deleted and non-validated data is stored in the Import database. Then the data is checked by standard rules (key duplicates, mandatory information and value domains) and by specific validation rules if they exist. All metadata about the validation is stored to the metadata repository, while errors found in the data are logged in the Processing database.

33. The **Validation module** enables re-applying of standard and specific validation rules in checking. This can be done at any moment after the first import of data or cleaning of erroneous data. All metadata about the validation is also stored to the metadata repository, while eventual errors are logged in the Processing database.

34. The **Data cleaning** module is intended primarily for correction of the erroneous data but it can be used for viewing data. All data errors are clearly presented with many graphic features and data can be searched by several criteria (the whole population, respondents only, non-respondents only) and filtered by population variables or errors by type. Every correction is logged in the Processing database with the information about the user, old and new values, etc. This module is the only that does not produce updates in metadata system.

35. The **Tabulation module** triggers the Warehouse Browser application. Again, users can access only those survey version registers and/or cubes that they are granted to. The variables are clearly presented and divided in groups (qualitative, quantitative, time and weight variables), can be easily filtered and presented in an appropriate way in result tables by using value domains. All information about them is retrieved from metadata system and new variables can be derived and stored to the repository. The results can be customized by users (different style templates, column widths, page settings, etc.) and saved in several data formats (MS Excel workbooks, PDF, HTML, px-files for PC-Axis, etc.).

36. Following is the start screen of the Survey Processor, where the modules for the selected study version are shown as tabs as well as buttons. The Monitoring module is started by default if the user has the access rights and an authorization level for survey processing. User can access only those modules which are granted to him/her in the metadata repository.



37. A module is enabled when its corresponding process is started, but there is a predefined scheme of enabling modules, depending on the survey version specifics. For example, if there are questionnaires used in the survey, then they must be first linked to the population by matching appropriate variables. Thus, the

Data import process cannot be started before the links are defined and corresponding module is not available until all conditions are met. For obvious reasons, process of Data Cleaning will not be started before data import was successfully completed.

V. CONCLUSION

38. The absolute target of CBS was to develop and manage a completely metadata-driven automated processing system. This objective was met by now and the system is ready for production which will by no means give further feedback and generate new ideas. By fulfilling this target CBS has opened the possibilities for a full transfer from mainframe to the client/server environment i.e. to move production work closer to the subject matter experts. This should lead to an absolute abandonment of the mainframe platform saving a large amount of maintenance money for the organization. No less an achievement should be the possibility to redirect a considerable amount of IT manpower into further development, taking into account that no software development will be needed for specific surveys, as regards general survey business processes. The solution developed upon the ISIS architecture should provide to that.

VI. SOURCES

(a) CBS/SCB – Sida Project, presentation Zagreb, Croatia, November 30th 2005, *Development of a Central Metadata Repository and a Public Macro Database at the Central Bureau of Statistics of Croatia*

(b) CBS/SCB – Sida Project, Presentation Zagreb, Croatia, June 17th 2008, *Presentation of Integrated Statistical Information System: Metadata Manager, Survey Processor and Warehouse Browser*

(c) Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS) – Luxembourg, April 2008, Metadata and the statistical cycle and Implementation, WP19, *Case Study: Central Bureau of Statistics of the Republic of Croatia*

(d) Andreas Goldman: *Development of an Integrated Statistical Information System (ISIS) at the Central Bureau of Statistics, Croatia*, Conscious™, October 2007