Distr.
GENERAL

WP.6
12 April 2011

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 2011)
(Luxembourg, 23-25 May 2011)

Topic (i): Architectures, models and standards

Open GSBPM compliant data processing system in Statistics Estonia

Invited Paper

Prepared by Allan Randlepp and Maia Ennok, Statistics Estonia, Estonia

I. Introduction

1. Statistics Estonia is modernising statistical production phases according to the GSBPM and the ESS vision for the decade. Within the past several years Statistics Estonia has invested a lot to centralize and renew data collecting systems. Now is time for modernising data processing. 2010-2011 is in development new independent and metadata driven system of statistical activity data processing. Next is planned to renovate the phase of statistical analyse.

2. This paper describes the open-sourced freely available technological component-based architecture and working principles of this new system of unified statistical data processing.

II. Background

3. Development of unified data processing system in Statistics Estonia has been high level strategical goal since 2008. Strategy of Statistics Estonia 2008-2011 (http://www.stat.ee/strategy) includes three high level objectives and second of these is “High-quality information service” that includes goal “Standardise the process of data processing”. This goal has indicator: “Introduction of the unified data processing software”.

III. Introducing VAIS

A. Main objectives

4. VAIS (pronounced as wise) stands for Vaatluste Andmetöötluse InfoSüsteem -- system for statistical activity data processing. VAIS is a collection of tools and technologies aimed at automating data processing (Phase 5 in GSBPM) related to statistical activities to prepare data records for analysis.

5. In essence, the task of check, clean, and transforming statistical activity data can be identified as taking the raw data from one or more sources (sub-process 5.1 Integrate data) survey (administrative registry, sample questionnaire, census, etc) and transforming it to analytical system source data (data warehouse) structures (sub-process 5.8 Finalize data files).

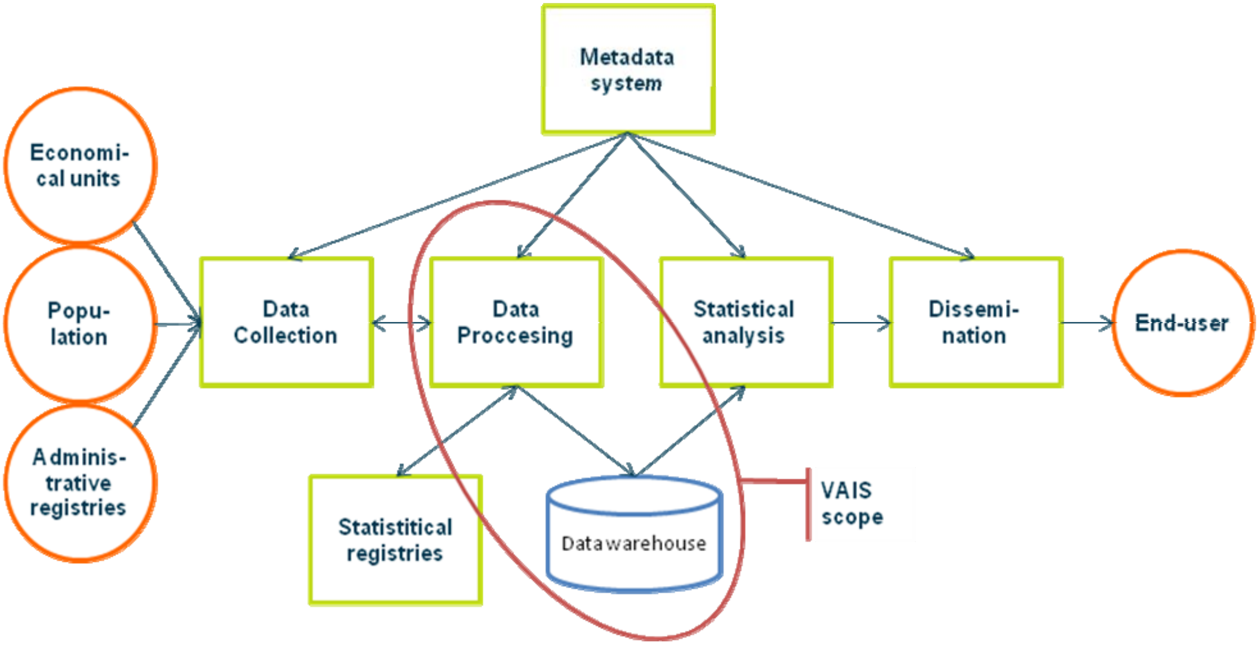

Fig 1: IT architecture in Statistics Estonia and VAIS scope

6. Besides pure data transformation from one structure to another, there are a number of operations that need to be carried out:

(a) Classify and code the raw data. Including automatic and clerical coding routines which assign numeric codes to text responses according to a pre-determined classification scheme (sub-process 5.2 Classify and code);
(b) Validation of raw data for technical and logical errors, correction of errors and amendment of raw data (sub-process 5.3 Review, validate and edit);
(c) Imputation of missing data fields and/or statistical units using a rule-based approach (sub-process 5.4 Impute);
(d) Derive variables and statistical units that are not explicitly provided in the collection, but are needed to deliver the required outputs (sub-process 5.5 Derive new variables and statistical units);
(e) Create weights for unit data records according pre-defined methodology (sub-process 5.6 Calculate weights);
(f) Create aggregate data and population totals from micro-data (sub-process 5.7 Calculate aggregates).

7. VAIS is being designed to provide tool framework for all these tasks.

B. Metadata driven template based tool

8. Essentially, VAIS is metadata driven process oriented ETL tool. This means that all data transformations are stored in the metadata repository. Storing data transformations in metadata repository is not a novel task, but VAIS approach is different. Instead of storing any kind of transformation e.g. by analyzing SQL statements, VAIS is template driven.

9. Each operation (e.g. remove duplicates) that can be performed in VAIS is implemented in a template. Templates can be in SQL, SAS or in other programming languages. Each template implements a specific data transformation task and is parameter driven. Parameters can be either specific values or pointers to other metadata stored in the metadata repository (e.g. database structures).

10. Template driven approach provides an universal solution for three main goals of the VAIS project:

(a) Create an easy to use statistical data processing tool requiring **minimal programming skills** for transformation package creation;
(b) Create **a metadata driven process-oriented and automated** statistical data processing tool;
(c) Create an **extendable** data transformation tool.

11. By encapsulating the code in templates, VAIS removes the requirement to program data transformation for each and every statistical activity separately. New templates can be added, if the existing ones do not provide a required functionality, without any changes to the tool itself. Usage of existing templates does not require any deep programming skills.

C. Data processing with VAIS

12. Data processing with VAIS start with the design of the data transformation package. The package includes definitions and mappings for:

(a) data extraction from raw data providers to data staging area (DSA);
(b) loaded data validation rules;
(c) data transformation such as imputations, removal of duplicates, coding, calculations for new variables etc;
(d) transformed data validation rules;
(e) data loading to data warehouse (Final Observation Register).

13. All these definitions and mappings are stored in the metadata repository. On package execution VAIS run-time engine (VAIS RT) accesses the stored metadata and using template engine builds the execution code of the data processing packing, different steps of which are executed in the corresponding environments – SQL, SAS/Base and others.

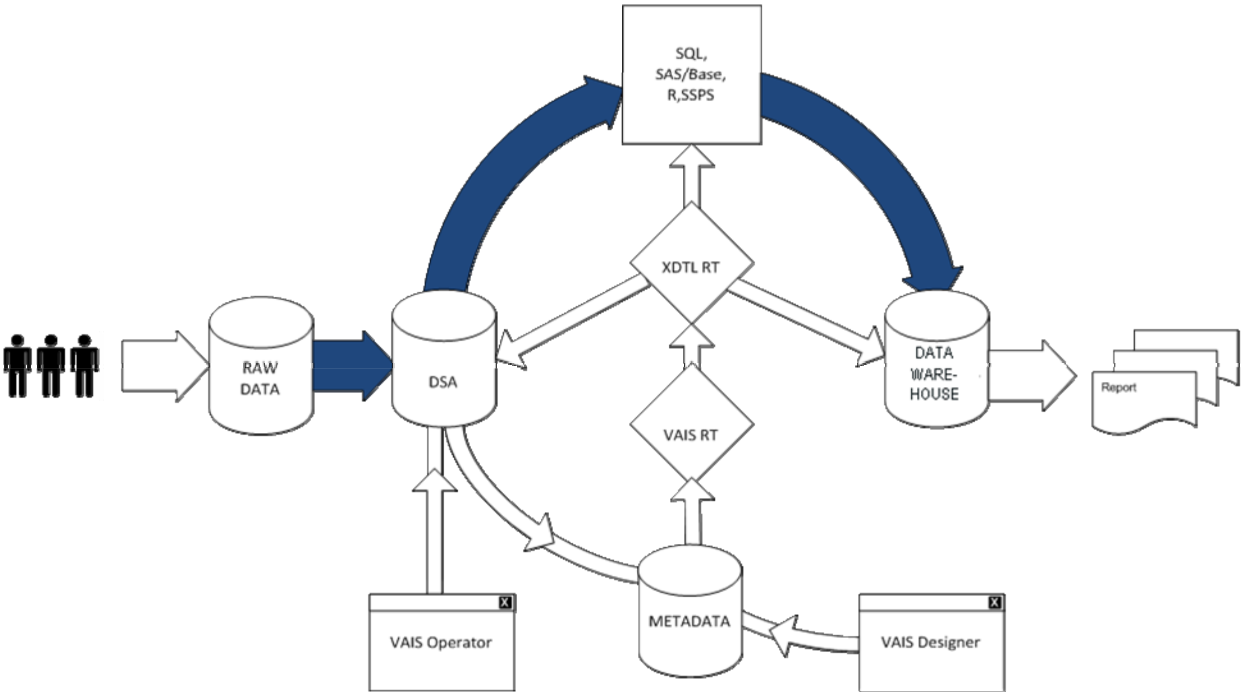

Fig 2: Data processing with VAIS

14. Transformed data is then stored in the data warehouse according to the mappings defined in the metadata and is available for analysis using analysis tools.

D. Automating and speeding up data processing

15. One of VAIS project goals is to provide means for speeding up the data transformation processing: not so much in means of processing time, but being able to start with data transformations as soon as some raw data is available.

16. Starting to transform data before all the raw data has become available allows to:

(a) test incoming raw data for overall quality and possible make necessary adjustments to the acquisition process;
(b) test data transformation process and make necessary changes;
(c) provide some incomplete data for analysis to test the survey data for logical problems.

17. VAIS design includes versioning and time-stamping of the cleansed data. This allows the analyst to refer to specific version of the cleansed data.

E. Raw data, transformation metadata and cleansed data audit trails

18. One of the topics that being given special attention in the VAIS design is the traceability of the statistical data processing process. The legislative requirements state that cleansed data, transformation process and raw data should be auditable for all produced statistical reports.

19. As not all of the raw data providers have built-in versioning of the raw data, VAIS is designed to provide for these needs as well. Data time-stamping routines include creating a restorable copy of the VAIS staging area used for data processing, including all raw data, corrections made to raw data either automatically or manually, the amended data and all transformation procedures.

F. Balancing automation and manual intervention

20. Data transformation in VAIS begins with defining the transformation package (sub-process 2.6. Design production systems and workflow). This is done in VAIS Designer application first by product manager, who outlines the needed transformations and then by data manager who specifies all necessary technical details. If a need to develop a template for some operation arises in the design of the transformation package this can be handled either by data manager or programmer from IT department (sub-processes 3.2 Build or enhance process components and 3.3. Configure workflows).

21. After data transformation package has been designed and tested, it is scheduled for execution. Scheduled packages run regularly and transform any new or corrected raw data extracted from the raw data provider. Packages can also be executed on external trigger events, VAIS provides a special web-service for those purposes.

22. In the package execution, data validation rules constantly check the incoming and transformed data. Data validation rules are amended with instructions how to handle errors - either automatically correct mistakes or raise an issue for the data operator. All automatically and manually done corrections are logged. Errorous data is excluded from further transformations and from loading the data warehouse.

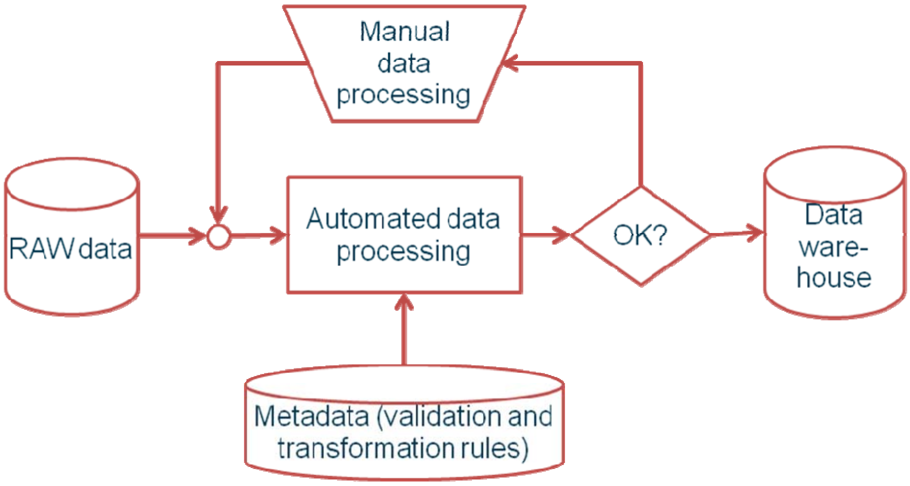
Manual data processing
RAW data
Automated data processing
OK?
Data ware-house
Metadata (validation and transformation rules)

Fig 3: Balancing automation and manual intervention

23. Data operator can look at the errorous data and decide further course of action: either dismiss the problem or correct the data manually. All corrections are allowed only to data processing phase (i.e. no corrections are done in data warehouse) and are logged. As the operator resolves the data conflict, the data object is included in the next data transformation package execution.

G. Technical platform

24. VAIS is built on open-sourced freely available technological components.

25. VAIS is based on XDTL (eXtensible Data Transformation Language – an XML based descriptional language designed for specifying data transformations, see http://xdtl.org) run-time engine (XDTL RT). This is an open standard, extremely lightweight and completely metadata driven ETL environment. The cornerstone of every ETL tool is efficient, automatic creation of transformation code (SQL, SAS etc.).

26. Instead of relying on clumsy, inefficient visual tools or extremely complex code generation XDTL RT takes advantage of an innovative concept of reusable and reconfigurable transformation templates that are merged with metadata into executable code to accomplish various ETL tasks. This merging process can be almost fully automatic (based on transformations and data mappings defined in metadata repository), leaving the transformation designer the task of specifying the correct template parameters.

27. VAIS is based upon MMX Metadata Repository, part of Metadata Framework (a MOF compliant metadata management environment designed with a wide variety of metadata-driven applications in mind, see http://mmframework.org).

28. XDTL RT is built in Java, taking advantage of several best of breed, open-source Java components. Apache Foundation's Velocity template engine (http://velocity.apache.org) is used as the template engine combining excellent template rendering functionality with very easy to use template language.

29. The user applications (VAIS Designer for data transformation package design and VAIS Operator for handling data problems) are programmed in Java, based on Wicket MVC framework (http://wicket.apache.org)

30. Quartz scheduling framework (http://www.quartz-scheduler.org) is used for execution scheduling.

IV. Implementation

31. Project of developing VAIS started 2010 and will be finished on October 2011. At first VAIS will be implemented on statistical data processing of Population and Housing Census 2011 that will start at the end of this year. Secondly VAIS will be implemented on reusing administrative data. 2012 Statistics Estonia is obligated by law not to recollect data that is already in administrative registries, this means that Statistics Estonia will prefill questionnaires in web collection system eSTAT with annual bookkeeping report data. VAIS is used for converting administrative data into the statistical data format. Data processing of other statistical activities will be implemented in VAIS from 2013.