Distr. GENERAL
06 May 2013
WP.13
ENGLISH ONLY

UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE
EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Work Session on Statistical Metadata
(Geneva, Switzerland 8-10 May 2013)
Topic (ii): Case studies and tools

METADATA MANAGEMENT AND STATISTICAL BUSINESS PROCESS AT STATISTICS ESTONIA

Working Paper

Prepared by Kaja Sõstra, Eda Froš, Statistics Estonia

I. Introduction

1. Statistics Estonia has been developing a metadata driven statistical information system for some years. The aim is to develop standardised systems for the statistical business process to improve the efficiency of statistical production.

2. Using the UNECE generic statistical business process model, we determined the activities and processes in need of standardisation to make them as transparent and compatible as possible. Standardisation enabled the development of common software systems for these processes. In order to manage these systems as an operational whole, a linking metadata system is needed. In 2011/2012 the old metadata management system was replaced with a new one.

3. In parallel with the metadata management system, a statistical data processing system and a system for the acquisition and management of administrative data were developed. These systems are linked to the metadata management system. The electronic data collection systems were developed earlier, but will be matched with the new metadata system. Integrated metadata system - iMeta

A. Purposes of integrated metadata system

4. The main purposes of the integrated metadata system are:

    a. to support the whole statistical business process;

    b. to enable the storage of metadata once for all usages and at the moment when they are made available;

    c. to act as a central repository for various outputs regardless of purpose, subsystem of SIS, media or time;

d. to act as an instrument for harmonizing and standardizing metadata;

e. to improve coordination of the statistical business process.

5. Additional benefits of integrated metadata system are more logical connection between the stages of the statistical process and smoother communication between those working in different stages of the statistical process and in different subject matter areas.

B. Components of integrated metadata system

6. The first phase of project on integrated metadata management system covered only the most important metadata objects. The central part of the metadata management system is based on the Neuchâtel Terminology Model for Variables (TMV), adapted to our needs. TMV complies with the standard ISO/IEC 11179, but has been adapted to statistical needs. TMV covers also statistical activities and other objects related to variables. The description of administrative data collection should be added to this Model.

7. Integrated metadata system consists of following components:

    a. Metadata navigator

    b. Description of statistical activities and instances of statistical activities

    c. Classifications, classification versions and variants, correspondence tables, code lists

    d. Concepts

    e. Statistical units types and statistical characteristics

    f. Measurement units

    g. Information about questionnaires, questionnaire versions

    h. Legal acts

    i. Databases

8. Metadata repository contains metadata managed by iMETA application and other applications. Metadata navigator gives an overview of all metadata stored in metadata repository (terminology objects, SQL objects, etc.).

9. Classifications constitute the main instrument in statistics. Classifications and code lists determine the value domains of a variable. Neuchâtel Terminology Model for Classifications is used for classifications. Classification is an umbrella that comprises one or more classification versions. Classification version is structured list of mutually exclusive categories. Classification variant is subset of elements of classification version. The original categories of classification version are split or regrouped to provide context-specific additions to the standard structure. Correspondence tables

10. Statistical activity is usually defined as the collection, storage, transformation and distribution of statistical information. In Statistics Estonia the concept of statistical activity is wider including not only the conducted statistical surveys, but also management of statistical registers, compilation of yearbooks and analytical publications, methodological developments as well as other works related to the production of statistics. Every year a new version (instance) of statistical activity is being described.

11. The description of statistical activity is based on ESMS concepts, supplemented by special attributes needed for Statistics Estonia. Because the large number of attributes describing statistical activity it was decided to group attributes into logical groups as follows:

a. General information – type of statistical activity, objective, content, cost, contact information, legal acts related to activity, time of approval by Government;

b. Methodology – attributes concerning different aspects of survey methodology: planned changes in methodology, source data, description of statistical population and frame, data collection, data processing, confidentiality policy and practice, revision policy and practice etc.

c. Quality management – attributes describing quality of survey: sampling and nonsampling errors, comparability, coherence etc.

d. Dissemination – all attributes related to dissemination: release calendar, dissemination frequency, news releases, publications etc.

e. VVIS and E-respondent – special attributes for CAWI mode surveys including explanations and instructions for respondents.

12. For statistical surveys all attributes are compulsory to fill in while for other statistical activities only general information is needed. Metadata system and user interface enables to add new attributes and groups without the need for additional programming. In addition to descriptive attributes every statistical survey has the list of register and cube variables, statistical indicators, classifications and questionnaires.

13. Description of statistical activities enables to present descriptions of surveys according to Euro-SDMX structure (ESMS) on the web, to create a document of Statistical Programme for 5 years, to present list of conducted statistical activities with short description by years on the web and to create XML file according to Euro-SDMX MSD.

14. Process metadata is mainly be described by using the ETL (Extract, Transform and Load) procedures both for the system meant for the acquisition of administrative data and for the survey data processing system. Various attributes needed for the survey data processing system are added to a respective contextual variable.

15. Metadata system consists of three main parts: MMX metadata repository, iMeta web based user interface and XML services which is part of data processing system (see figure 1).

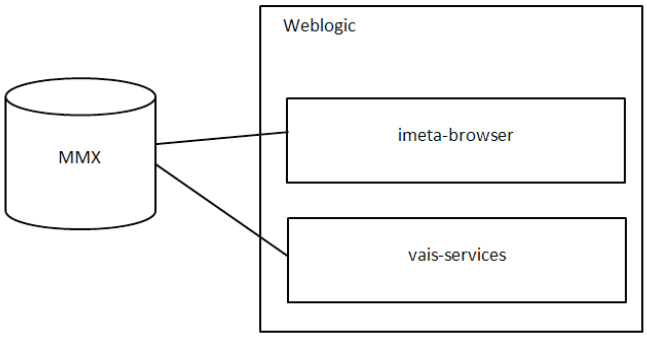

Figure 1. General architecture of metadata system

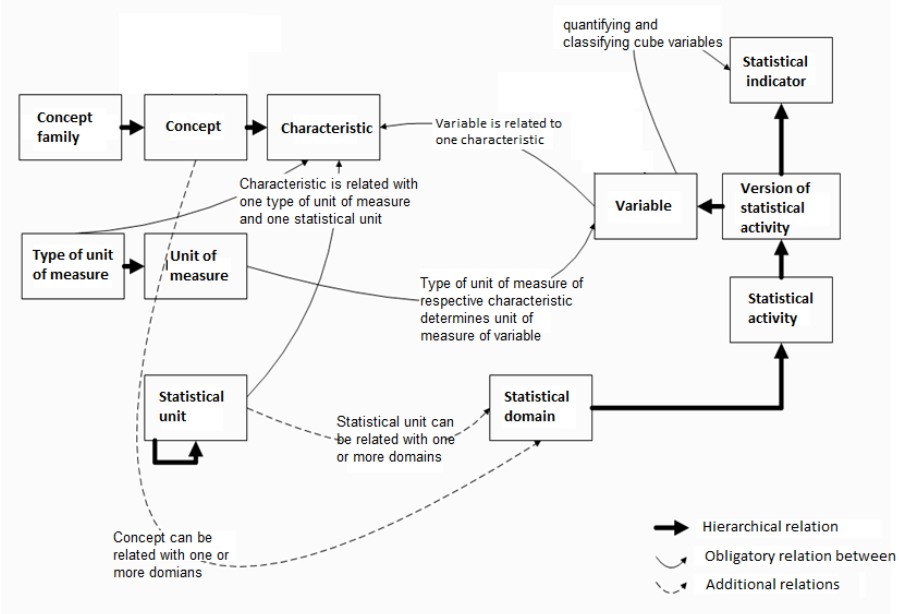

Figure 2. Relations between main objects of metadata system

16. A statistical unit type describes a class of statistical objects. Statistical characteristics (Object variables) define variable characteristics in connection with a defined statistical unit type regardless of their role. Conceptual variable is a concept from which statistical characteristic is derived by combining it with a statistical unit type. Specifying variables (classifying cube variables) are a subset of object variables, which are used for further specializations of the statistical unit type into sub-types. Measure is a quantitative object variable (quantifying or classifying cube variable). A contextual variable defines a variable in the context of statistical activity. It refers to the statistical characteristic (object variable) which provides a standard definition for the variable. The value domain of the contextual variable must be consistent with the conceptual domain of the object variable.

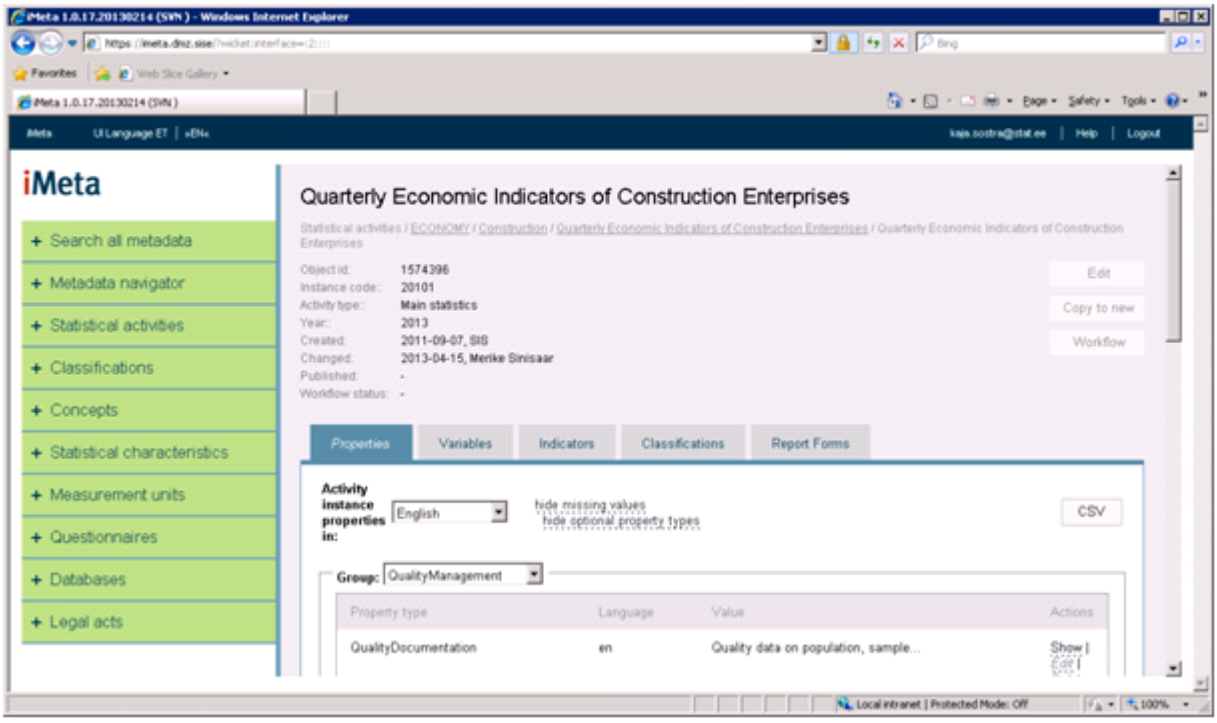

Figure 3. User interface of iMeta

II. Statistical information system

C. Architecture

17. Architecture of statistical information system and the relationships between its components and relations with GSBPM is presented in the figure 4.

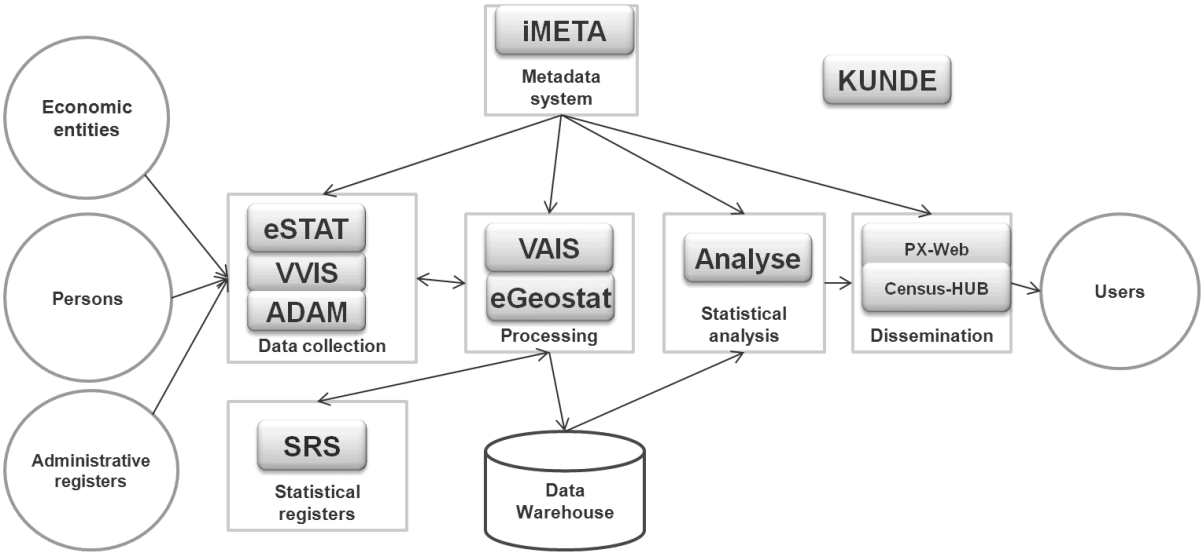

Figure 4: Architecture of statistical information system

18. Information system consists from following information systems:

    a. iMETA – integrated metadata system (2011);

    b. SRS – system of statistical registers (2012);

    c. Kunde – customer relationship management system (2006);

    d. eSTAT – data collection system for economic entities (2006);

    e. VVIS – data collection system for persons (2011);

    f. ADAM – data collection system for administrative data (2011);

    g. VAIS – template based data processing system (2012);

    h. Analyse – system for statistical analyses (2013);

    i. PX-Web – output database (2001) will be replaced by .Stat (2014);

    j. Census Hub – dissemination tool for European statistics (2013-2014).

D. Metadata flows within the information system

1. Statistical register system (SRS)

19. The aim of SRS is to create common system for managing statistical registers of population, businesses, agricultural holdings, buildings and dwellings; use more effectively data collected to the

state information system; to avoid duplication in data collection; increase quality of statistics produced based on statistical registers; start to develop register based population and housing census.

20. SRS functions are forming and managing statistical units, managing links between the units of different statistical registers, forming and managing sampling frames for statistical surveys, selecting samples according to sample design options, preparing lists of respondents with updated contact information and other data, providing register data for users of statistical registers.

21. SRS uses following metadata from metadata repository: classifications, statistical activities, questionnaires. SRS creates detailed metadata about frame and sample selection supporting subprocesses 2.4 Design frame and sample methodology and 4.1 Select sample.

2. Data collection system for economic entities (eSTAT)

22. eSTAT enables respondents to view the list of statistical reports, which a particular economic entity has to present to Statistics Estonia during the current year; to view deadlines for presenting these statistical reports; to order reminders, which notify by e-mail about upcoming deadlines; to compile statistical reports, i.e. to fulfil cells on screen or to upload csv-fails (particularly meant for large enterprises with a big number of records such as the Intrastat statistical report and some statistical reports on wages); to run controls, i.e. check whether they have compiled the statistical report as required; to correct statistical reports immediately upon compilation thereof; to submit statistical reports to Statistics Estonia; to look at all earlier statistical reports submitted to Statistics Estonia via eSTAT by a respondent concerned; to print out a paper copy of a compiled statistical report; to administer users, i.e. to create, change and cancel rights and access; to accept or correct one's contact information, etc.

23. eSTAT uses following metadata from metadata repository: variables, classifications, questionnaires. The subprocesses 2.3 Design data collection methodology and 3.1 Build data collection instrument is performed based on this metadata. Further metadata supports data collection and processing steps.

3. Data collection system for persons (VVIS)

24. VVIS is complex system for collecting information from individuals and monitoring whole field work. System enables describe questionnaires in three languages, collect data by CAWI and CAPI mode, update contact information, perform data controls, code data, remove duplicates.

25. VVIS uses following metadata from metadata repository: variables, classifications, questionnaires. The subprocesses 2.3 Design data collection methodology and 3.1 Build data collection instrument is performed based on this metadata. Further metadata supports data collection and processing steps.

4. Data collection system for administrative data (ADAM)

26. Functionalities of ADAM are automatic extraction detailed personalized data from administrative sources using X-road (data exchange layer) or ftp, storing data in raw data databases, data processing, coding, duplicate removal, making data available for in-house applications.

27. Special statistical activities are described in metadata system for administrative sources including description of variables. Metadata about variables are used in data collection and processing.

5. Template based data processing system (VAIS)

28. VAIS is a collection of tools and technologies aimed at automating data processing (Phase 5 in GSBPM). 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.

29. 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).

III. Current metadata projects and future plans

E. Grant agreement "Implementation of ESS metadata standards on describing statistical activities and disseminating data"

30. The main objective of the project is describing reference metadata (in the metadata system) for all statistical activities. The following actions are planned and partly made: review of existing metadata; collection of recommendations for harmonization and improvements; compilation of guidelines for describing content of metadata concepts, based on requirements for ESS metadata structures; trainings of survey managers, to give overview of new recommendations, guidelines and processes; update of metadata based on recommendations (by survey managers). Planned result of the actions is complete and high quality reference metadata about all statistical activities.

31. Second objective of the project is modernisation of reference metadata describing and update processes. Detailed plan and process description for updating metadata enables to keep metadata updated after the end of present project.

32. Third objective is dissemination of reference metadata on the website. For this objective the vision document is compiled and accepted by the IT Project Committee of SE). Based on this vision processes will be developed for getting metadata automatically and directly from the metadata system to the website of SE. Final phase is implementation of the processes and programming of systems/links between metadata system and website (by website administrators and developers).

F. Technical development of iMeta

33. During development of system for electronic data collection eStat and system for analysing data the new needs for metadata system were specified. It was decided to change the data model moving variables from statistical activity instance under statistical activity. Old model generated every year new set of variables which courses technical difficulties to other systems.

G. Future plans

34. Dissemination of ESMS based reference metadata on the web is planned on the 1st of July this year. Currently descriptions of statistical activities is disseminated partly only in Estonian in Word document. After disseminating ESMS metadata on the web users can easily access systematic and updated information.

35. Release of concepts and definitions on the web of Statistics Estonia. Replacement of current HTML version of concepts and methodology in output database.

36. Further development of iMeta in line with developments of other components of statistical information system.

IV. Lessons learned

37. Several parallel developments course problems of specification common requirements for all systems. Unfortunately, all new developments bring along some changes in the metadata system, which needs continuous development and improvement in order to meet the emerging requirements of developing systems.

38. Responsible unit and persons should be appointed for management of metadata. The metadata group of Methodology Department is responsible for development, implementation and management of metadata repository. A separate manager has been appointed for each set of metadata, descriptions of statistical activities, classifications and code lists, contextual variables, conceptual variables and statistical unit types, etc. Metadata group is also responsible for the harmonization of metadata.

39. Creation of new metadata, filling in the gaps and harmonisation is very labour-intensive. Support from management and other people in the office is essential for success. Special guidelines and rules should be created. All potential internal users of metadata system should be informed about the development process and involved in it. A training plan for implementation has to be created. At the very beginning, terminology has to be introduced to make the whole staff use the same terminology and understand each other.

V. Documents and links

40. Neuchâtel Terminology Model – classification database object types and their attributes http://www3.ssb.no/stabas/DOCS/Neuchatelversion2.1.pdf

41. Neuchâtel Terminology Model – variables and related concepts http://www.ssb.no/english/metadata/metadatadocuments/varneuchatelnodel.pdf

42. MMX Metadata Framework – implementation of MOF (built on relational database) technology http://www.mmxframework.org/