

**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**

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Geneva, 3-5 April 2006)

Topic (iv): Implementation

STATLINE 4 METADATA IMPLEMENTATION

Invited Paper

Submitted by Statistics Netherlands¹

I. INTRODUCTION

1. StatLine 4 is the new dissemination system of Statistics Netherlands. A very important new feature is its metadata server that will be used to coordinate conceptual output metadata. In 2006 the current output database will be migrated to StatLine 4 including its metadata and will be put into production. After conversion all available conceptual output metadata of Statistics Netherlands will be moderated and coordinated.

2. The paper starts with a short description of the classification and variable properties that will be coordinated within the StatLine 4 model. The main focus of the paper is the transformation from uncoordinated metadata to StatLine 4. The strategy to convert the existing metadata includes a preparation, conversion and coordination phase. These phases and their rationales will be described as well as their organisational aspects. The paper concludes with the benefits of StatLine 4 and future improvements.

II. COORDINATED METADATA IN STATLINE 4

3. All output data of Statistics Netherlands is available online in the output database StatLine. Begin 2006 this database contained approximately 1350 multi dimensional cubes containing over 2 billion data cells. Most of the metadata of these statistical cubes is currently not coordinated. In 2006 Statistics Netherlands will take the StatLine 4.0 software in production. One of its major features is an integrated metadata server. This metadata server facilitates the coordination of conceptual metadata. In StatLine 4 conceptual metadata are statistical variables and classifications as modelled in the Cristal model [1]. Cristal models variables, categories, levels ("flat classifications") and hierarchies

¹ Prepared by Edwin de Jonge[ejne@cbs.nl].

(hierarchical classifications). Categories can be shared between levels and between hierarchies. For each of these items a global id (GUID) and a local key is stored. All items have multilingual labels and descriptions and for each language it is possible to have an expert and a popular version of the labels and description. This feature is used throughout the StatLine 4 web software: every table or page can be viewed in popular or expert mode. A data cube about consumer price index can be shown as a table on inflation or on CPI.

4. Furthermore Cristal models versions of categories, levels and hierarchies. Each item has a begin date and an end date, during which it is valid. The Cristal model guards and checks the constraints classifications and their versions have. For example a level can not contains overlapping categories and a hierarchy can only contain categories that are valid during its “life time”. This version information is used within the Web application of StatLine [2]

5. Storing variables and classifications in StatLine 4 can only be done by a organisational unit within Statistics Netherlands. This coordination unit acts as a gateway for storing and publishing conceptual metadata. This means that all data in StatLine will use centrally moderated and coordinated conceptual metadata which make up the borders of the tables retrieved from StatLine.

6. StatLine 4 stores also other types of metadata that are not modelled in Cristal. These metadata will be checked and edited by dissemination department but they will not be coordinated.

III. CONVERSION STRATEGY

7. The introduction of the StatLine 4 software will be in phases, mainly because the conversion of the current 1350 StatLine data cubes to StatLine 4 takes months: it includes a manual coordination step.

8. In StatLine 4 each classification (or variable) has a coordination status. This status is one of the following decreasing values of coordination:

- National standard: The classification is a (inter)national standard.
- Coordinated: The classification is coordinated within Statistics Netherlands.
- Shared: The classification is shared, but has not reached (yet) a sufficient level of coordination.
- Private: The classification can only used for a specific dataset. This status is mainly used during and shortly after the conversion of the StatLine database

9. The conversion uses the following coordination strategy. Only classifications that are equal to standard classification must be coordinated, where feasible other classifications may be coordinated. Classification that are not coordinated are marked as private. After conversion data cube owners have a period of two years to replace this private classification with a coordinated one.

10. The conversion phases are preparation, conversion, [launch StatLine Website], and a coordination phase. The conversion process ²is supported by a conversion tool that calculates similarity measures for classifications and aids in (auto)mapping a data cube classification onto a coordinated classification.

² The conversion process description in this paper is not complete. In this paper only the meta data aspects are mentioned

Preparation phase

11. In the beginning of the preparation phase the standard classifications are determined. After this each data cube is checked for the existence of standard classifications using the conversion tool. Classifications that are coordinated can be used. Other classifications are marked as private. After the preparation phase it is clear for each StatLine 3 data cube what the structure of it corresponding StatLine 4 data cube will be and what variables and classifications will be used.

Conversion phase

12. In the conversion phase each StatLine data cube will be converted into a StatLine 4 data cube.

13. Each variable and classification will be replaced by a coordinated or private variable or classification. Variants of coordinated classification are allowed and the variations will also be stored in the central Metadata server. Each converted data cube will be checked for content, metadata differences and presentation in the StatLine web application by the dissemination unit and the statistical subject matter specialists.

Coordination phase

14. After conversion of the StatLine database, the StatLine web application will be put into production.

15. The Metadata server will contain variables and classifications that are marked as private. These private metadata are the result of the conversion process and will be removed within two years in a coordination process. This process will be monitored by the coordination department.

IV. STATLINE 4 BENEFITS

16. One of the main benefits of StatLine 4 is the coordination of classification and variables as described in this paper. For a user of StatLine 4 this will be a great benefit. Coordinated metadata mean that figures from different data cubes can be compared more easily. It also means that all metadata have uniform names and descriptions in multiple languages and in expert or layman terms. Standard and coordinated classifications can be browsed and downloaded by internet users.

17. Not described in this paper are the many user interface improvements to the web application of StatLine. Also not discussed is that StatLine 4 changed the publication process of Statistics Netherlands. The publication process in StatLine 4 is split into three different activities. Managing metadata, supplying output data and design data cubes. Metadata and data can be shared between data cubes.

V. FUTURE IMPROVEMENTS

18. The metadata server of StatLine 4 does not cover all metadata coordination. Firstly it coordinates conceptual metadata and no process and quality metadata. Secondly is the coordination of conceptual metadata incomplete. Statistical populations and their relations are not coordinated. Therefore the Cristal model will be extended but a population model, which make coordination over populations, their variables and classifications possible.

19. In the longer term it is planned that process and quality metadata will be integrated in the metadata server.

VI. REFERENCES

- [1] E. van Bracht, Cristal a model for Data and Metadata, Working paper No 29 METIS, Geneva Februari 2004
- [2] J. Reedijk, E. de Jonge, O. ten Bosch, Handling Time dependence of metadata and data in StatLine 4, International Marketing and Output Database Conference, 2005 Oslo.