STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

COMMISSION OF THE EUROPEAN COMMUNITIES

CONFERENCE OF EUROPEAN STATISTICIANS

EUROSTAT

Joint UNECE/Eurostat Work Session on Statistical Metadata
(6 - 8 March 2002, Luxembourg)

Working Paper No. 26
English only

Topic (iii): Metadata and quality

AUTOMATIC METADATA CREATION AT THE STATISTICAL OFFICE OF THE REPUBLIC OF SLOVENIA

Submitted by the Statistical Office of the Republic of Slovenia [1]

Contributed paper

ABSTRACT

The main goal of the Statistical Office of the Republic of Slovenia (SORS) in the field of metadata is to develop an efficient and effective, standardized and integrated system for collecting and editing metadata as an important part of the statistical information system.

We will test the hypothesis that effective metadata management can overcome problems of time-consuming and extensive manual feeding of the corporate metadata repository. We will discuss whether SORS Corporate Metadata Repository can be automatically fed by the following repositories:

?? Oracle Discoverer repository
?? Blaise repository
?? Klasje - Classification server.

I. A BRIEF OUTLINE OF THE SORS CORPORATE METADATA REPOSITORY

1. Within the StatCop98 project, the component 4.1: Development of conceptual, technical and software solutions of common (infrastructure) importance had the following goals:

?? creating the concept of a statistical warehouse with special emphasis on common functions and metadata as well as its testing on a pilot project;
?? specification, development and introduction of EDI tools and procedures;
?? classification database – upgrading the existing functionality, developing software for managing the concordances;
?? developing software for browsing Classification via internet.

2. Another component - 4.3: Development of databases and software solutions - aims at an integrated process of aggregation and dissemination of data from the Census of agriculture, horticulture and viticulture 2000 (AC2000) and other agricultural statistics (AGRISTAT). Within these two components, the basic common functions in the context of statistical data warehouse were defined according to Sundgren (Sundgren 1997): "Statistical metadata are descriptive information or documentation about statistical data, i.e. microdata, macrodata, or other metadata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data".

---

[1] Prepared by Julija Kutin and Jozica Klep.

3. In a broad sense, "production" covers the whole life cycle of a statistical survey or a statistical information system, including design, implementation, operation, monitoring, maintenance and evaluation. Producers of statistical data therefore include: designers, input data providers and statisticians. All these categories of producers of statistical data have their typical metadata needs. Therefore, the SORS Corporate Metadata Repository and the necessary user interfaces were built with an aim to facilitate knowledge sharing between the different producer groups in developing the computer environment for the selected statistical surveys.

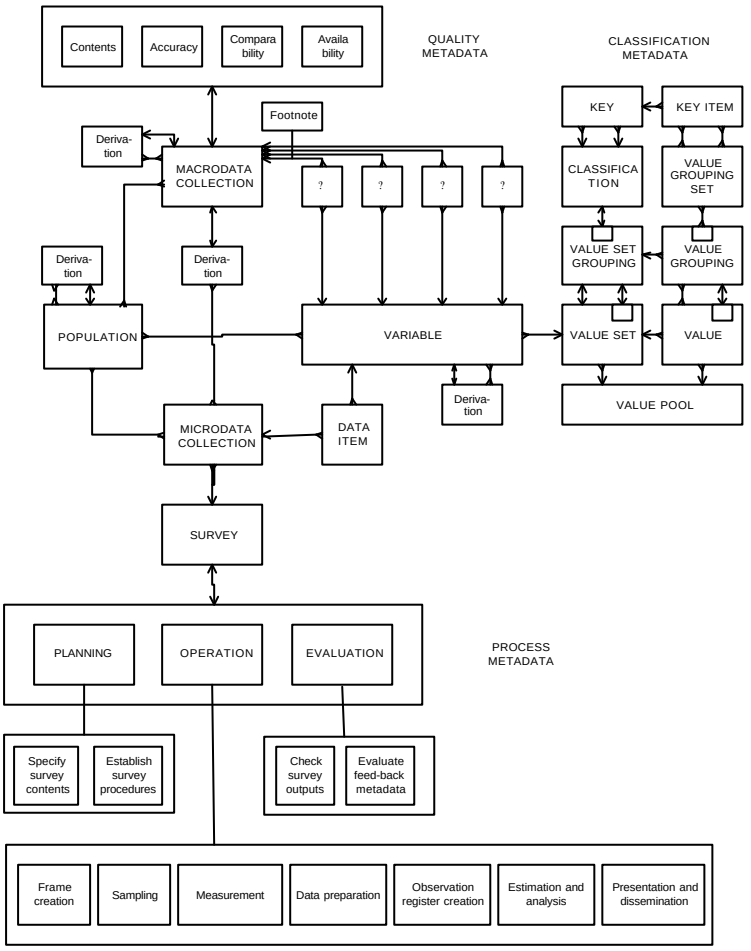

Figure 1: Metadata model for a statistical data warehouse (Sundgren 1997, p. 32)

II. THE MAIN MODULES OF THE METIS APPLICATION

4. The modules developed within COP98 cover the statistical metadata for the following: questionnaires, statistical variables and characteristics, planning and monitoring of the progress of statistical surveys (part of process metadata), selected data for the National Program of Statistical Surveys (NPSR) and SORS publications as well as the module with information on the physical location of data in the data warehouse (Oracle) and a link to unstructured metadata. Modules for metadata required for dissemination, quality metadata, sampling, etc., will be covered in the framework of the COP 2000 project.

5. Developing the solutions that will provide establishing and maintaining of the catalogue of statistical variables was the focal point of attention. The application enables definition of the variables with mapping and monitoring of the links between:

?? statistical variable and question asked (and the link to the questionnaire where the question was asked; explanatory notes for the question, if they existed, must be recorded as well);
?? object (population);
?? corresponding permissible value set of the variable.

6. SORS's application - METIS - for managing statistical metadata was built with Oracle Designer. The database is in Oracle8i and is comprised of 104 entities. METIS allows feeding of statistical metadata either manually or automatically - with only one click on the "Import" button on the form. Metadata can be imported from:

?? Excel files - data from database repository or from repository of a query tool;
?? Classification database;
?? Blaise repository.

II.1 Importing metadata from Excel files

7. Importing metadata from Oracle Discoverer: Excel is used for editing metadata. We have to convert Excel files into CSV files, from where we can import metadata into corporate metadata repository. However, where there is only little metadata to insert, we might wish to fill it interactively into the fields on the form - this might be the case for metadata for the questionnaire (Figure 2.).

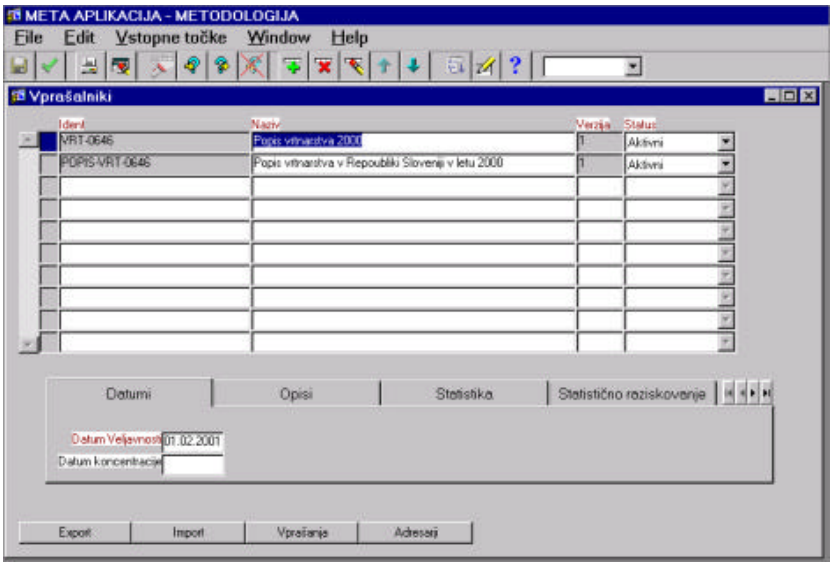

Figure 2: Form with the data for the questionnaire

8. But we think that it is easier to feed the corporate metadata repository automatically, especially when we have a set of metadata to insert. Metadata have to be prepared according to an exactly defined (Excel) template.

VPR_VPRNIK				
VPR_ID_OZN	VPR_IME	VPR_ID_VERZIJA	VPR_STATUS	VPR_DTM_VLJ
Ident	Naziv	Verzija	Status	Datum Veljavnosti
POPIS-VRT-0646	Popis vrtnarstva v RS v letu 2000	1	1	23.02.2001

Figure 3: Template for the data for the questionnaire (questionaire_id, name, version, status of validity, date of validity)

1								
2	OBO_ID_OBJ_OPZ	STS_ID_STA_SPR	STS_IME_STA_SPR	STS_DEF_STA_SPI	STS_DTM_VLJ	STS_IND	VLP_ID_I	VLP_ID_V	VLP_ID_F
3	Objekt opazovanja	Spremenljivka	Naziv	Definicija	Datum veljavnos	Ključ	Oznaka k	Verzija	Raven
4	POP_PODJETJE	MAT_ST	MATICNA_STEVILKA	Matična številka poc	12.11.2001	0		1	
5	POP_PODJETJE	IME_SUB	IME_SUBJEKTA	ime subjekta	12.11.2001	1		1	
6	POP_PODJETJE	SKD5	SKD5	SKD5	12.11.2001	1	SKD	1	5
7	POP_PODJETJE	SKD4	SKD4	SKD4	12.11.2001	1	SKD	1	4
8	POP_PODJETJE	SKD3	SKD3	SKD3	12.11.2001	1	SKD	1	3
9	POP_PODJETJE	SKD2	SKD2	SKD2	12.11.2001	1	SKD	1	2
10	POP_PODJETJE	SIF_HS	SIF_HISNE_STEVILK	šifra hišne številke	12.11.2001	1		1	
11	POP_PODJETJE	NASELJE	NASELJE	naselje SKTE5	12.11.2001	1	SKTE	1	5
12	POP_PODJETJE	OBCINA	OBCINA	občina SKTE4	12.11.2001	1	SKTE	1	4
13	POP_PODJETJE	REGIJA	REGIJA	regija SKTE3	12.11.2001	1	SKTE	1	3
14	POP_PODJETJE	POR_KAP	POREKLO_KAPITALA	poreklo kapitala	12.11.2001	1	Poreklo k	1	1
15	POP_PODJETJE	VRSTA_LAST	VRSTA_LASTNINE	vrsta lastnine	12.11.2001	1	Vrsta last	1	1
16	POP_ZAPOSLENI	MAT_ST	MATICNA_STEVILKA	matična številka	12.11.2001	0		1	
17	POP_ZAPOSLENI	LETO	LETO	leto	12.11.2001	1		1	
18	POP_ZAPOSLENI	VRSTA_ZAP	VRSTA_ZAPOSLITVE	vrsta zaposlitve	12.11.2001	1	Vrsta zap	1	1
19	POP_ZAPOSLENI	ST_ZAP	STEVILO_ZAPOSLEN	št. zaposlenih	12.11.2001	1		1	

Figure 4: Template with variables (object, variable_id, variable_name, variable_definition, date, key, classification_ scheme_name, classification level)

9. We have made Excel templates to import data to each table (Questionnaires, Questions, Variables, Columns of tables where data are stored). The files with the templates have to be saved on the exactly defined location on the personal computer, from where the procedure can load them directly.

10. As far as preparing the data in Excel is concerned, there are three methods which can be used:
?? write them manually;
?? import them from the query tool - at SORS we are using the Oracle Discoverer query tool;
?? import them from repository of RDBMS- at SORS we use Oracle 8i.

11. We can write metadata interactively - questions for a new questionnaire of a new survey (we can use questions from other surveys - if they are already stored in the corporate metadata repository).

12. We can create an Excel file directly from Oracle Discoverer repository of the query tool. Oracle Discoverer repository is the interface between relational tables and specific presentation of data structure. This means that in Discoverer repository one can map the relations between columns and tables in data schemas and items and folders. So users can make their own reports and use user-friendly names of tables and columns in the database.

Database schema (Oracle)	Discoverer	Metadata repository
Column	Item	Variable
Table	Simple folder	Object
View	Complex folder	Object

Figure 5: Relationship between elements in database schema, Discoverer and Metada repository

13. However, Discoverer repository has the ability to create a lot more of the data that are actually needed for the metadata repository, and there is no need to replicate them again. Therefore, a view can be created f.i. on the tables to show the three elements needed for the tables. A lot more can be inserted as parameters (storage_date, user_id....) (Figure 7).

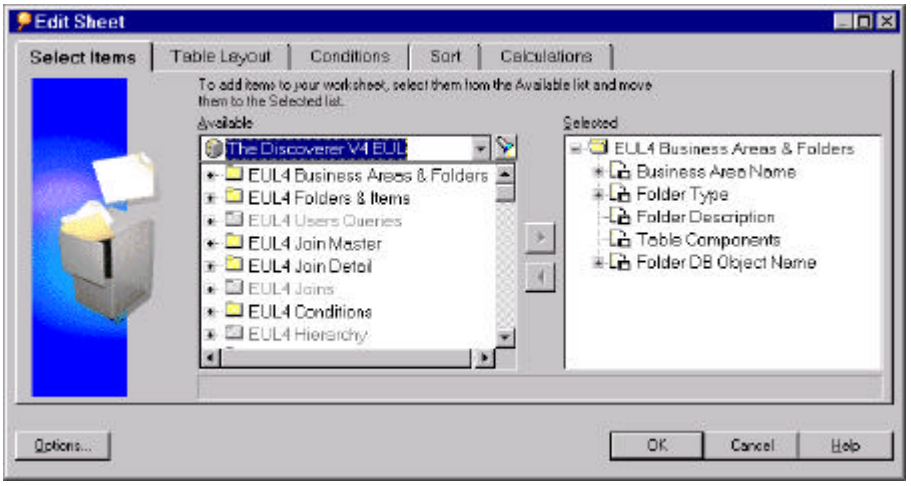

Figure 6: Discoverer view for tables

14. By simply pressing the OK-button (Figure 6) on Edit sheet in Discoverer and then on icon Microsoft Excel, we export the predefined view with the data on tables to Excel. After inserting other necessary parameters (version of table, status, time, user), we import them automatically to the metadata repository.

3	Business Area Name:VRT-POP	Folder Type:Simple		
4				
5	Table Components	Folder DB Object Name	Folder De	Folder Components
6	AC2000.D_CAS_RAZISKAVE	D_CAS_RAZISKAVE	Čas popisa	
7	AC2000.D_CENILNI_OK	D_CENILNI_OK	Geografska delitev na cenilne okoliše in katastrske občine	
8	AC2000.D_ENERGENTI	D_ENERGENTI	Šifrant energenti	
9	AC2000.F_ENER_OGR	F_ENER_OGR	Poraba energije za ogrevanje	
10	AC2000.D_INSTRUKTORJI	D_INSTRUKTORJI	Šifrant inštruktorjev	
11	AC2000.D_IZOB_VRT	D_IZOB_VRT	Šifrant vrtnarska izobrazba	
12	AC2000.F_IZO_VRT	F_IZO_VRT	Strokovna izobrazba za vrtnarstvo	
13	AC2000.D_MERE	D_MERE	Šifrant merskih enot	
14	AC2000.D_NACIN_PORABE	D_NACIN_PORABE	Šifrant način porabe pridelkov	
15	AC2000.D_NAC_PRIDEL	D_NAC_PRIDEL	Šifrant način pridelave	
16	AC2000.D_NAC_PRID_ZPRO	D_NAC_PRID_ZPRO	Šifrant način pridelovanja v zaščitenem prostoru	
17	AC2000.F_NAC_PROD_NPOR	F_NAC_PROD_NPOR	Namen porabe izdelkov glede na način prodaje	
18	AC2000.D_NAC_TRZEN	D_NAC_TRZEN	Šifrant način trženja za kmetijske pridelke	
19	AC2000.D_NAC_ZALIV	D_NAC_ZALIV	Šifrant način zalivanja	
20	AC2000.F_NACIN_PROD	F_NACIN_PROD	Namen porabe pridelkov	
21	AC2000.D_OGREVANJE	D_OGREVANJE	Šifrant ogrevanje	
22	AC2000.F_OKR_RAST	F_OKR_RAST	Pridelava okrasnih rastlin	

Figure 7: Data for tables from Discoverer view

15. In the same way (view on Discoverer repository) we prepare and then import the data for columns.

16. The Oracle repository of RDBMS - Data Dictionary permits the selection of select data and then the preparation of the Excel file with the data required.

17. Views on Oracle repository of RDBMS - Data Dictionary currently used in select statements
a) for tables:
?? user_tables or all_tables
?? user_objects
b) for columns
?? user_tab_columns
?? user_col_comments.

18. The idea is that the user executes the procedure by clicking on the button in corporate metadata repository application. The procedure is saved in METIS API (Application Programming Interface). The procedure is executed only if the database administrator grants the user all required privileges on the selected database schema.

II.2 Importing data from Classification database

19. With our new development environment (Developer6i, Designer6i Release 2, HeadStart 6, Oracle 8.1.7 repository), we found new opportunities to load data from one database to another. It is a challenge to automatically import classification - allowed value sets for variables from classification database.

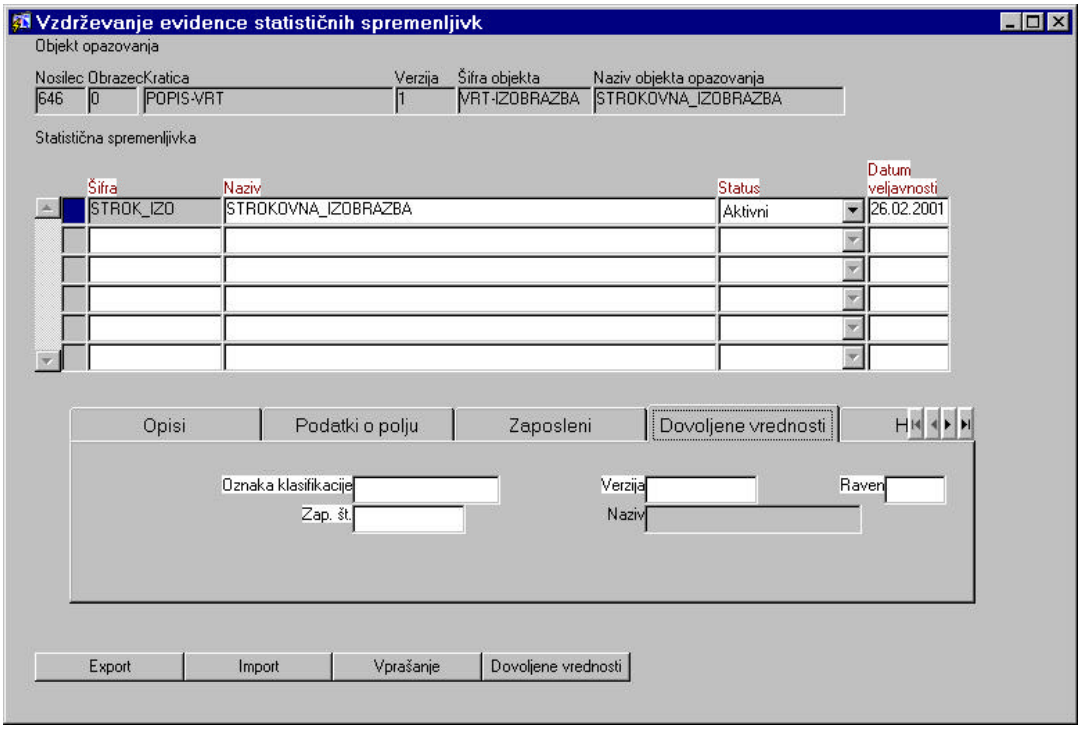

Figure 8: Statistical variable with its value set name

20. Where there was no corresponding value pool in the corporate metadata repository, we found it easier to import it from the classification database.

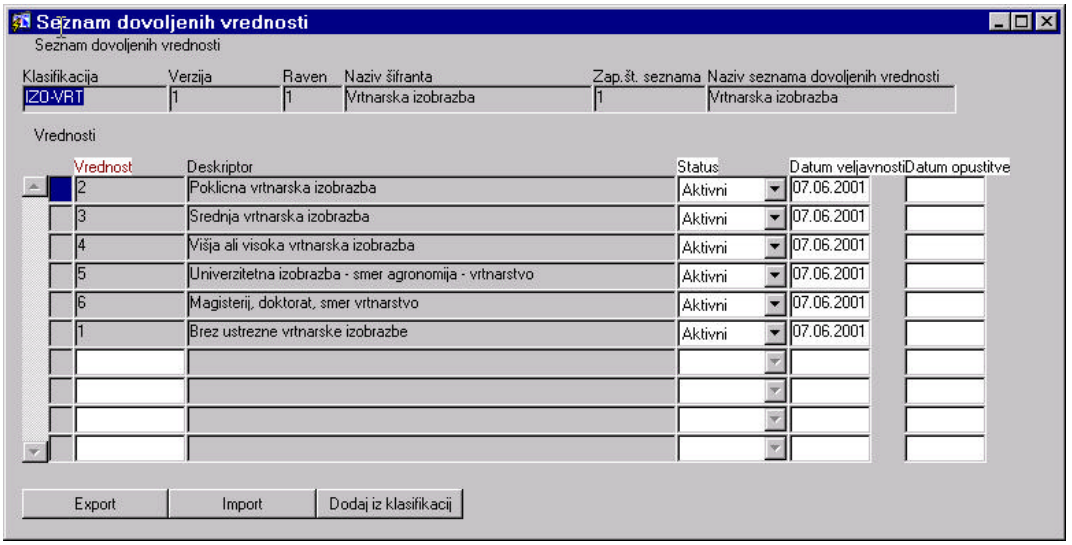

Figure 9: Form for importing the variable value set from the classification server

21. Into field Classification we type the name of the classification, and then press button "Dodaj iz KLASJA" - meaning "import from classification database". This application offers all versions of the required classification, so we select the right one and import it to the corporate metadata repository (Figure 10).

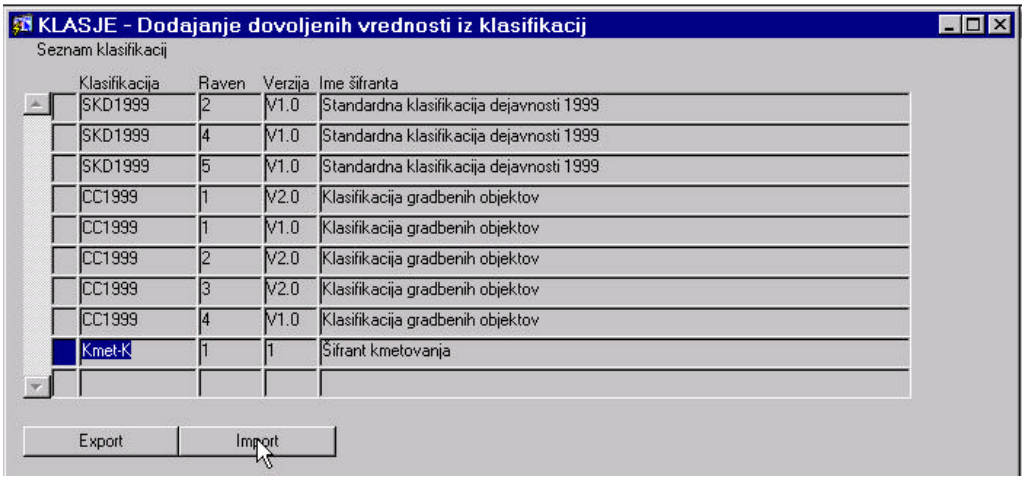

Figure 10:Selected version of the classification will be imported into metadata repository

II.3 Importing data from Blaise repository

22. Blaise is the server processing system for collecting and editing statistical data. We plan to import metadata (data on questions) from the Blaise repository in the XML format. This means that the conversion tool must be capable of transforming the dedicated questionnaire definition format into an XML format. Microsoft has made available MSXML.DLL, and it can be used in Visual Basic, Java, or Java Script program.

23. There are a lot of questionnaires that are not created in Blaise, so it would be better to find a universal way of preparing metadata files and importing prepared metadata into the metabase - independently of the tool that is used for designing the questionnaires. There are two suggestions:
i) the pass procedure is executed from Word: subject-matter statisticians prepare questionnaires in Word documents and mark question text with special headings (METIS-heading); in Word there will be

a macro with WSA (Window's Script Application) procedure, which will parse the document and generate an Excel file with metadata about questions; subject matter statisticians will continue the importing process in METIS. There will be a button for the procedure for importing metadata into the corporate metadata repository.

ii) The parse procedure is executed from METIS application: subject-matter statisticians prepare document with metadata in the same way; work is continued in the METIS application, where WS (Window's Script) procedure is executed. It will read the file with a standard name and location, pass the file and load metadata into the corporate metadata repository.

III. CONCLUSIONS

24. We are convinced that the SORS Corporate Metadata Repository can be automatically fed and that the possibilities can be even further developed and expanded in the near future.

?? For the time being the same Excel file can be used for loading questions to the Blaise repository and to the Corporate Metadata Repository.

?? If the database administrator carefully designs and names the Discoverer Folders and items contained therein, objects can be mapped with the tables and variables with columns in the tables.

?? There is a direct connection applied and functioning between the classification server and the CMR. The permissible (if defined) value set of the variable can be loaded directly.

REFERENCES

Dippo C., Gillman D., The role of metadata in statistics, UN/ECE METIS, Geneva, 22-24 September 1999

English L., "e-QUALITY: Quality in Internet and e-Business Information", SRC.SI, Grimšce 19.-20.3.2001

English L., Improving Data Warehouse and Business Information Quality, Methods for Reducuing Costs and Increasing Profits, Wiley Computer Publishing, 1999

Gillman D., Appel M.V., Highsmith S.N., Building a Statistical Metadata Repository at the U.S. Bureau of the Census, UN/ECE Work Session on Statistical Metadata,WP No. 11, Geneva, 18-20 February 1998

Harold E. R., XML Bible, IDG Books Worldwide, Inc, An International Data Group Company, Foster City, CA, 1999

Klep J., Metadata Quality From A Business Perspective, UN/ECE Work Session on Statistical Metadata, 28-30 November 2000, Washington, D.C., United States

Marco D., Building and Managing the Metadata Repositiory, A Full Lifecycle Guide, Wiley Computer Publishing, 2000

Moeller R.A., Distributed Data Warehousing Using Web Technology, How to Build a More Cost-Effective and Flexible Warehouse, Amacom, 2001

SORS, Software Development for Management of Data of the Statistical Office of the Republic of Slovenia, Project No.: SL – 9803.02.0001.03, SUBPROJECT: 4.1.1 Statistical data warehousing, data concepts and general solutions; Working code: DWSURS - 4.1.1 Special annex: Metadata management

Sundgren B., An informations systems architecture for national and international statistical organisations, Methodological report, draft, April 1997

Sundgren B., Documentation and Quality in Official Statistics, Statistics Sweden, International Conference on Quality in Official Statistics,14-15 May 2001, Stockholm,

United Nations, Terminology on Statististical Metadata, Statistical Standards and Studies No. 53, Geneva, 2000

Andersson E., The present and future metadata system at Statistics Sweden, Statistics Sweden, 2001-04-12