WP.10
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

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

Topic (ii): Metadata Concepts, Standards, Models and Registries

SORS THESAURUS OF STATISTICAL TERMS

Supporting Paper

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

I. INTRODUCTION

1. The main roles of thesaurus in the office are assisting in the search for information in metadata repositories and providing definitions for terms found in the various metadata repositories. These two functions are not yet developed to the full degree.

2. However, another very important functionality is being developed, namely preservation of bilingual (Slovene – English) professional terminology development.

3. Statistical terminology is organised as a "macro" thesaurus, where 14 (field) thesauruses are organized, maintained and managed. We are now using it as a terminology tool - fort the purpose of standardisation of bilingual terminology and as a bilingual (Slovene - English) dictionary - (as part of the bookshelf with electronic dictionaries). Software support was developed step by step during last years.

4. From March 2006 it is available for the internet users from SORS web site.

II. SHORT BACKGROUND

II.A. Hierarchies as definition systems

5. To meet different user's scenarios, also search mechanisms need to be of several types. The traditional way of searching and navigating has been trough hierarchically organised menu systems. This type of search may be satisfactory for many users, especially for casual users with rather "standard" information needs. On the other hand, there are users who know rather precisely what they are looking for and will prefer identification of a particular time series.
On the other hand, there are users who are not very articulated about what they are looking for and they may found hierarchies too rigid, imposing a view of the statistical information that they do not

[1] Prepared by Jozica Klep, jozica.klep@gov.si.

feel “at home” with. Moreover, one hierarchy may not be enough to direct the user to all relevant statistics.

6. Dahlberg (1991) says that thesauri provide the natural-language access to knowledge – that is their task, a “classic” one by now. But what kind of knowledge? On the one hand, we have the knowledge of facts and interrelationships as laid down in defined terms and/or concepts, and on the other hand, the knowledge of interrelationships as reflected in thesaurus relations. Knowledge always come about/trough statements/propositions/judgements. Any definition is such a judgement, as well as the establishment of relation to a super- or subordinated concept, so that all hierarchies can be regarded as definition systems.

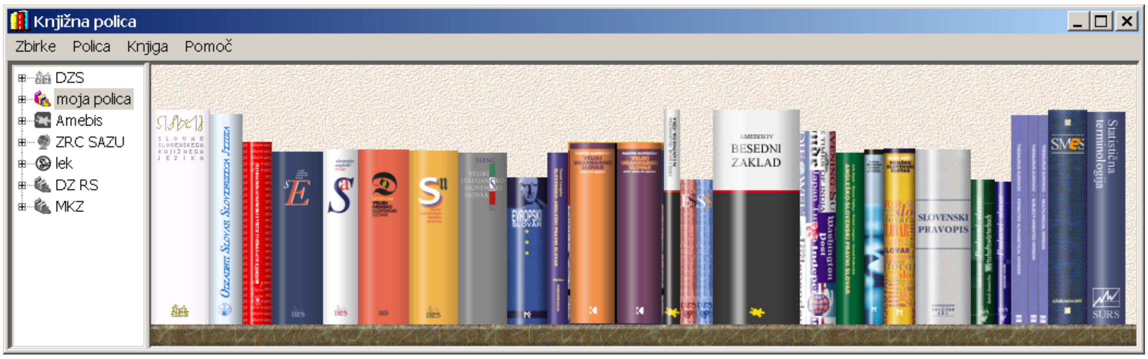


Figure 1: Electronic dictionaries, Statistical terminology as a part of the "bookshelf"

II.B Electronic dictionaries within SORS network

7. Over the last years, SORS bought electronic dictionaries, available to all employees within the office network to support publishing of statistical data. The following dictionaries are available:

Vocabulary of Slovene language
Dictionary of Slovene language
English – Slovene dictionary
Slovene – English dictionary

II. C EUROVOC thesaurus

8. EUROVOC thesaurus of European Union was translated within the Slovene government project of Information Documentation Centre. Library of Slovene Parliament played the role of a coordinator. It is considered as a public good and available free of charge.

9. EUROVOC thesaurus is organized in the way that can be consulted as a dictionary and is available from the same interface.

II. D Statistical terminology

10. From different sources. Altogether 22628 terms and phrases available in March 2006.

III MANAGEMENT AND MAINTENANCE OF STATISTICAL TERMINOLOGY

11. Like any other terminology, statistical terminology is continually evolving and has to be adapted to take account on the one hand of development in the fields in which statistical offices are active and on the other of changes in the language.

12. It has also an important role to convey meaning of statistical (and other) constructs, developed in international organisations to Slovene users. To this end, maintenance procedure must be based on both, internal (within the office) and external needs.

13. Maintaining of Statistical terminology database is the responsibility of SORS Sector I – General methodology and standards.

IV. IMPORTANT ROLE OF IT SUPPORT

14. Support for statistical terminology management was developed by Slovenian company AMEBIS Kamnik. This company is specialised in language technologies. It developed support for electronic dictionaries (ASP 32) that are for sale in Slovenia.

15. The database has two main functions:

- act as a repository for approved terminology, and hence as a tool for the harmonisation and control of professional terminology
- to provide a thesaurus – based means of searching for statistical information on a given subject.

16. The database is MS SQL with interfaces that allow loading, editing, deleting, inserting terms (sources, authors, etc). However, the “semantic” tree (broader terms) can only be managed and edited within the application. Administrators and authors (of translations) are allowed to access this database. Main attributes for each term are seen on Figure 3.
The attributes for synonyms and related terms are foreseen but application does not allow yet their maintenance.

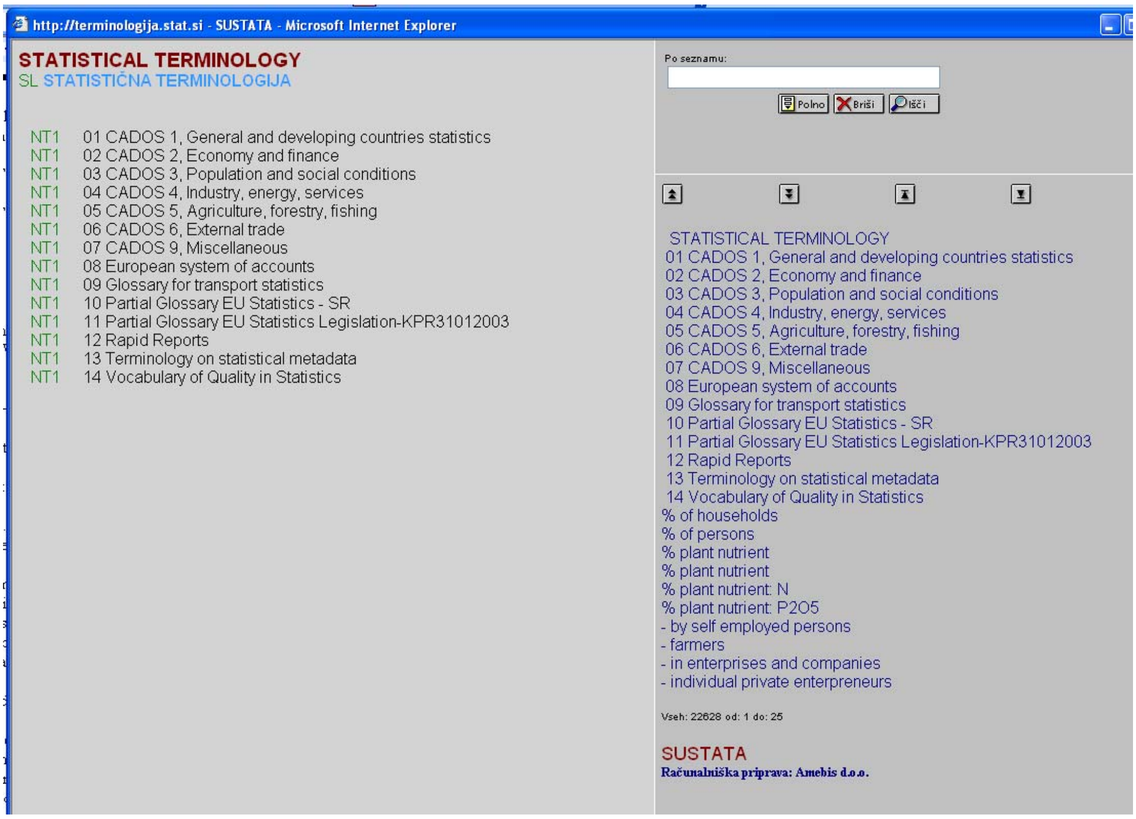
17. There is a fixed number of attributes available for every concept. Several modes of retrieval are possible; concept card is one of them (figure 3).

18. After completed editing each thesaurus (source) is exported from the database as xml file and stored to predefined folder. When all the desired files are ready in the folder administrator runs the “generation” procedure. Procedure reads all the records and “calculates” the semantic tree from the data on broader terms (parents) and further on generates three files. The first file which follows the rules oaf ASP32 is organised according to the rules and is copied to the ASP32 interface. Copying of the file creates a new version of Statistical terminology (and it becomes instantly available within the SORS network) in the shape of an electronic dictionary. The terms are automatically connected to two dictionaries, namely to English - Slovene dictionary and to Dictionary of Slovene language. Consultation, browsing, searching and use (copy/paste) follow the same rules as in other dictionaries.

19. The second and the third file are created for consultations with browser (one with the tree in Slovene language, the other with the tree in English language). These files are then copied to the internet server and accessible from SORS home page for Slovene language:
http://www.stat.si/

20. For English language
http://www.stat.si/eng/index.asp
From the folder “Methodology, Projects” as “Statistical Terminology”.

The procedures are repeated whenever necessary.



http://terminologija.stat.si - SUSTATA - Microsoft Internet Explorer

STATISTICAL TERMINOLOGY
SL STATISTIČNA TERMINOLOGIJA

NT1	01 CADOS 1, General and developing countries statistics
NT1	02 CADOS 2, Economy and finance
NT1	03 CADOS 3, Population and social conditions
NT1	04 CADOS 4, Industry, energy, services
NT1	05 CADOS 5, Agriculture, forestry, fishing
NT1	06 CADOS 6, External trade
NT1	07 CADOS 9, Miscellaneous
NT1	08 European system of accounts
NT1	09 Glossary for transport statistics
NT1	10 Partial Glossary EU Statistics - SR
NT1	11 Partial Glossary EU Statistics Legislation-KPR31012003
NT1	12 Rapid Reports
NT1	13 Terminology on statistical metadata
NT1	14 Vocabulary of Quality in Statistics

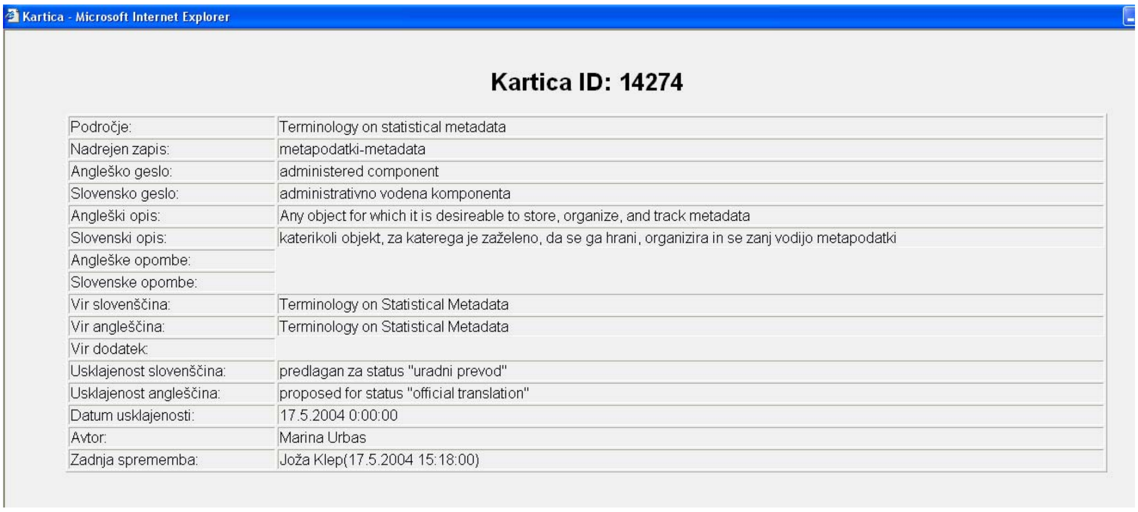
Po seznamu:

STATISTICAL TERMINOLOGY
01 CADOS 1, General and developing countries statistics
02 CADOS 2, Economy and finance
03 CADOS 3, Population and social conditions
04 CADOS 4, Industry, energy, services
05 CADOS 5, Agriculture, forestry, fishing
06 CADOS 6, External trade
07 CADOS 9, Miscellaneous
08 European system of accounts
09 Glossary for transport statistics
10 Partial Glossary EU Statistics - SR
11 Partial Glossary EU Statistics Legislation-KPR31012003
12 Rapid Reports
13 Terminology on statistical metadata
14 Vocabulary of Quality in Statistics
% of households
% of persons
% plant nutrient
% plant nutrient
% plant nutrient: N
% plant nutrient: P2O5
- by self employed persons
- farmers
- in enterprises and companies
- individual private enterpreneurs

Vseh: 22628 od: 1 do: 25

SUSTATA
Računalniška priprava: Amebis d.o.o.

Figure 2: Statistical terminology as seen from the website



Kartica ID: 14274

Področje:	Terminology on statistical metadata
Nadrejen zapis:	metapodatki-metadata
Angleško geslo:	administered component
Slovensko geslo:	administrativno vodena komponenta
Angleški opis:	Any object for which it is desireable to store, organize, and track metadata
Slovenski opis:	katerikoli objekt, za katerega je zaželeno, da se ga hrani, organizira in se zanj vodijo metapodatki
Angleške opombe:	
Slovenske opombe:	
Vir slovenščina:	Terminology on Statistical Metadata
Vir angleščina:	Terminology on Statistical Metadata
Vir dodatek:	
Usklajenost slovenščina:	predlagan za status "uradni prevod"
Usklajenost angleščina:	proposed for status "official translation"
Datum usklajenosti:	17.5.2004 0:00:00
Avtor:	Marina Urbas
Zadnja sprememba:	Joža Klep(17.5.2004 15:18:00)

Figure 3: "Card- view" with the set of attributes for "administered component"

21. ASP 32 (developed by AMEBIS Kamnik) consists of four parts (generation system, database in ASP format, browser for databases in ASP format, auxiliary programmes for enhanced functionalities).

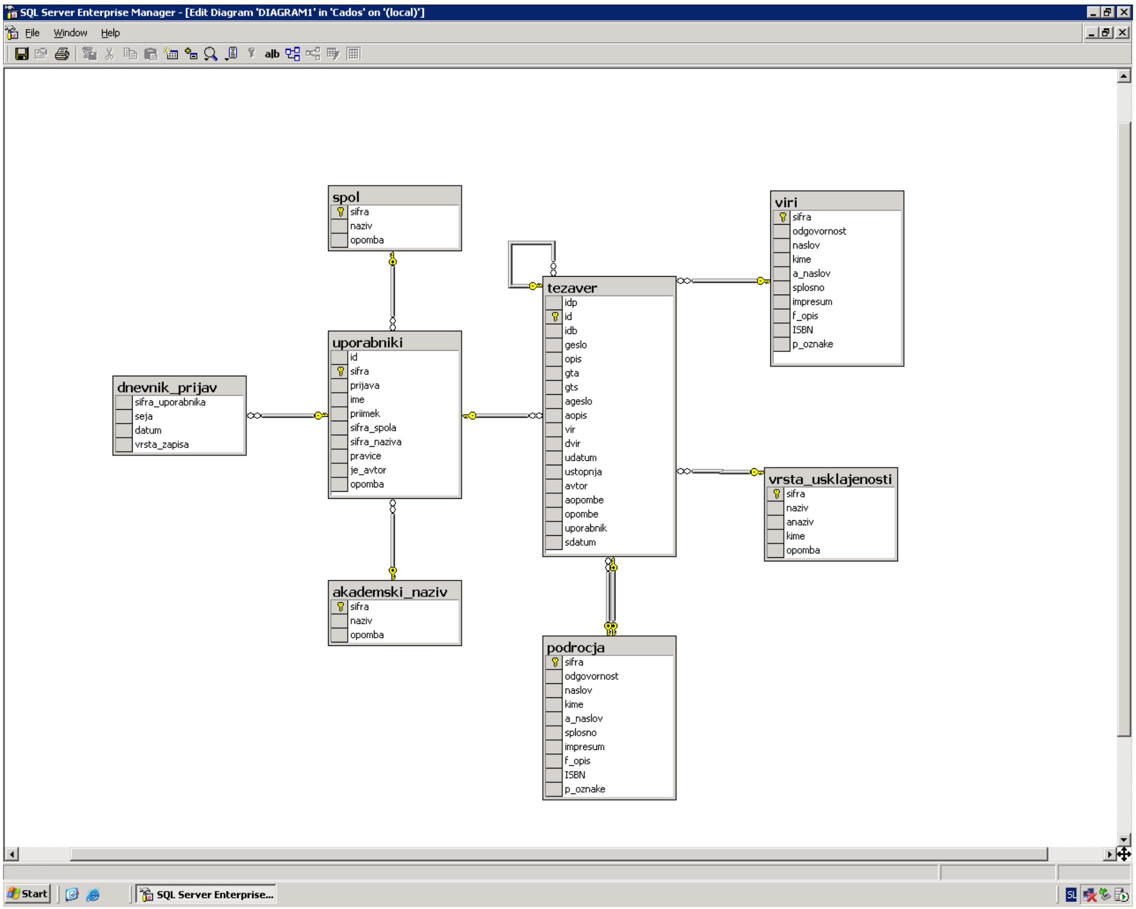


Figure 4: Database schema

V. DESCRIPTION OF CONTENTS OF TERMINOLOGY DATABASE

22. In spring 2006 terminology from following thematic fields is available for intranet and internet users:

01 CADOS 1, General and developing countries statistics
02 CADOS 2, Economy and finance
03 CADOS 3, Population and social conditions
04 CADOS 4, Industry, energy, services
05 CADOS 5, Agriculture, forestry, fishing
06 CADOS 6, External trade
07 CADOS 9, Miscellaneous
08 European systems of accounts
09 Glossary for transport statistics
10 Partial Glossary EU Statistics - SR
11 Partial Glossary EU Statistics Legislation-KPR31012003
12 Rapid Reports
13 Terminology on statistical metadata
14 Vocabulary of Quality in Statistics

23. Only part of this stock will be elaborated bellow.

A. EUROSTAT CADOS THESAURUS

24. What follows is a brief description of history of establishment of CADOS Thesaurus (from CADOS, Thesaurus 3, Population and social conditions, 1989, pages 7-16)

25. In 1986, when the European Parliament and the Publications Office were revising the EUROVOC Thesaurus, the Commission of the European Communities launched the “Common Vocabulary” project in thesaurus form. At first, EUROSTAT looked into the possibility of joining this project in which it was intended to create a domain on statistical information.

26. Since EUROSTAT did not only need a controlled vocabulary but also a collection of specific aids allowing analysis of the logical organisation of data in databases, it was decided to create CADOS thesaurus which would contain this degree of specificity. A thesaurus has been developed for each theme contained in the EUROSTAT standard scheme for publications. The same themes were used to structure CRONOS in domains. Each theme corresponded to a thesaurus.

27. Sources for the methodology were ISO standards and for contents: CRONOS database, publications of EUROSTAT, EUROVOC thesaurus, “Common Vocabulary” and Macro thesaurus of OECD.

28. In March 1989 the CADOS thesaurus included:

- 7 thesauri
- 72 micro-thesauri
- 1299 generic terms, which were divided as follows:
- 93 for theme 1. General and developing countries statistics
- 102 for theme 2: Economy and finance
- 66 for theme 3: Population and social conditions
- 74 for theme 4: Industry, energy and services
- 61 for theme 5: Agriculture, forestry and fishing
- 56 for theme 6. External trade
- 847 for theme 9: Miscellaneous (lists, methods, unclassified words)
- 5608 descriptors
- 4421 synonyms or quasi-synonyms in French
- 2081 synonyms or quasi-synonyms in English
- 3119 synonyms or quasi-synonyms in German.

29. For on-line search we read the following:
Thesaurus consultation provides knowledge of the vocabulary used for indexing and allows “visualisation” of the data available and to prepare a question optimally taking into account the number of references selected.

30. “Generic posting” at the index enables one, during interrogation, to receive all data which is hierarchically inferior in level. On interrogation with “cereals” for example one would receive data concerning wheat, barley, rye, etc. This form of indexing takes into account the hierarchical structure of statistical data and enables direct access to a large amount of data.

31. If, because of this “generic posting”, the answer to a question is too detailed one can reduce the number of references selected by way of keywords which give the level of detail desired.

32. The CADOS thesaurus is worth of a particular attention for many reasons, among the most important being:

- It is a thesaurus of statistical expressions (terms, indicators) of the European Statistical System.
- It is trilingual (French, English, German).
- It was designed as a part of an overall scheme to document all databases in a consistent manner.
- It is a base on which a general catalogue of available data could be founded. In this context it provides a form of a “road map” around statistical information.
- It can give information on the logical organisation of information and how it could be improved.

33. Further “negative priority” development in EUROSTAT regarding thesaurus of statistical terms is revealed in Norre, Groenez and Pellegrino (2004).

B. Terminology on statistical metadata

34. Among the first files loaded in the database was Terminology on statistical metadata. It was translated to Slovene language even before it was officially published in the year 2000. The file was received from Dan Gillman together with recommendations how to deal with hierarchies.

C. Terminology from Rapid Reports

35. Files from Rapid Reports are representing a very important source of bilingual terminology. Rapid Reports is the most comprehensive series of statistical data that is published by SORS. It is bilingual (Slovene – English).

36. The contents of series are structured in 29 fields of statistics and further divided into 190 subjects.

37. The files for loading were prepared as a by-product during the exercise of developing strategy on disclosure control. All the publications were carefully read to find sensitive variables. Variables (and valueset) were recorded together with table titles. These files were then loaded into the database.

38. Terms and phrases from Rapid Reports have the highest validity codes – status of official translation. In the process of translation: MCV – loaded; translation to Slovene started.

VI. PLANS AND FUTURE DEVELOPMENT

39. In the near future we will further elaborate PC-AXIS files and suggest the PC AXIS reference group to consider of giving priority to a new keyword, namely a “search keyword” to the structure of PC files.

40. Then the proposal of international organisation regarding best practice in representing a subject will be applied, f.i. http://dublincore.org/documents/usageguide/elements.shtml

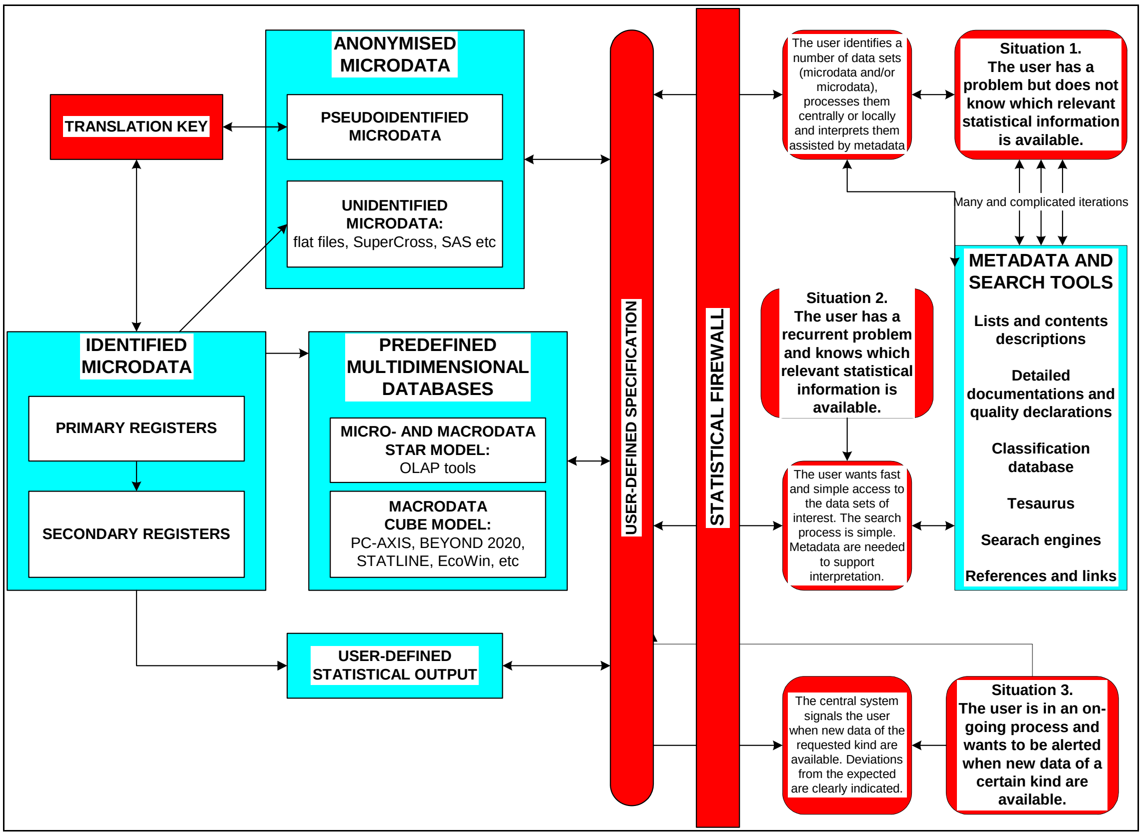

ANONYMISED MICRODATA
TRANSLATION KEY
PSEUDOIDENTIFIED MICRODATA
UNIDENTIFIED MICRODATA: flat files, SuperCross, SAS etc
IDENTIFIED MICRODATA
PRIMARY REGISTERS
SECONDARY REGISTERS
PREDEFINED MULTIDIMENSIONAL DATABASES
MICRO- AND MACRODATA STAR MODEL: OLAP tools
MACRODATA CUBE MODEL: PC-AXIS, BEYOND 2020, STATLINE, EcoWin, etc
USER-DEFINED STATISTICAL OUTPUT
USER-DEFINED SPECIFICATION
STATISTICAL FIREWALL
The user identifies a number of data sets (microdata and/or microdata), processes them centrally or locally and interprets them assisted by metadata
Situation 1. The user has a problem but does not know which relevant statistical information is available.
Many and complicated iterations
METADATA AND SEARCH TOOLS
Lists and contents descriptions
Detailed documentations and quality declarations
Classification database
Tesaurus
Search engines
References and links
Situation 2. The user has a recurrent problem and knows which relevant statistical information is available.
The user wants fast and simple access to the data sets of interest. The search process is simple. Metadata are needed to support interpretation.
The central system signals the user when new data of the requested kind are available. Deviations from the expected are clearly indicated.
Situation 3. The user is in an on-going process and wants to be alerted when new data of a certain kind are available.

Figure 5. An example of user access to statistical data (Sundgren 2003)

41. Valuable new functionality will therefore be added while building a well-interconnected data services over the Internet; overall strategy explained in Jug 2004.

VII. REFERENCES

Bargmeyer B., Establishing Ties Between Metadata Registries and Terminology, Eighth Open Forum on Metadata Registries, Berlin, 2005

Dahlberg I., Knowledge organization, Thesauri, and Terminology, Editorial, in International Classification, Vol. 18 (1991), No.3

CADOS, Thesaurus 3, Population and social conditions, 1989

http://dublincore.org/documents/usageguide/elements.shtml

Jug M., Web-supported statistical dissemination process serving statistical data users, Joint UNECE/Eurostat/OECD Meeting on the Management of Statistical Information Systems, Geneva 2004

Norre B., Groenez D., Pellegrino M., Integrating Statistical Terminology Tools within Eurostat's Dissemination Policy, Joint UNECE/Eurostat/OECD work session on statistical metadata, Geneva 2004

Sundgren B., Developing and implementing statistical metadata systems, MetaNet WG3 Deliverable D6, 2003-06-30

UN/UNECE, Terminology on statistical metadata, Geneva 2000