

UNITED NATIONS STATISTICAL COMMISSION and
UNITED NATIONS 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)

PROVISIONAL AGENDA AND TENTATIVE TIMETABLE

The aim of this meeting is to exchange experience among statisticians and other experts and to identify best practices in the field of statistical metadata. These cover information on concepts and classifications and the statistical process of data collection, processing and dissemination. Information from some of the papers presented at the meeting will be directly or indirectly incorporated into the METIS Metadata Framework that was developed following the February 2004 meeting.

Registration and Security: The meeting will begin at 9.30am on Monday 3 April 2006 in Salle XVI at the Palais des Nations in Geneva. On the first day of the session confirmed Delegates¹ are requested to present themselves at the Pass and Identification Unit of the UNOG Security and Safety Section (maps online at <http://www.unece.org/stats/documents/ece/ces/ge.40/2006/inf.3.e.pdf>) for the issuance of an identification badge. The Office is open Mondays to Fridays from 8 a.m. until 5 p.m (non-stop). In case of difficulty, please contact by telephone the UNECE secretariat on +41 (0)22 917 2084 (internal extension 73328 or 72084).

Monday, 3 April 2006		
Time	Session	Documentation
9:30	Opening of the meeting	WP.1
9:45	Introduction to session (i) and presentation of invited papers (1hour)	
Session (i) Metadata in a Corporate Context Session Organizer: Bo Sundgren (Statistics Sweden) A statistical metadata system must be an integral part of the strategic direction of a statistical organization. For this reason, it is vital that senior management is directly involved in any metadata project. This topic explores the approaches being taken to inform senior managers of the issues relating to metadata management. Issues addressed should include: <ul style="list-style-type: none">• characteristics and elements of an effective metadata management system;• the metadata needs of different uses / users (internal purposes, end-users of statistical agency output, international comparison purposes);• processes for keeping staff and senior managers informed about metadata issues.		
Invited papers:		
Presentation of a draft of Part A of the Metadata Framework: Statistical Metadata System and its role in a statistical organization Jana Meliskova (Consultant), Graeme Oakley (Australian Bureau of Statistics)		WP.2
The Value Added of a Statistical Office Marleen Verbruggen and Max Booleman (Statistics Netherlands)		WP.3
10:45	Break (20 minutes)	

¹ Delegates must have completed the registration form (available in the Information Note and from the Internet website of the UNECE Statistical Division (<http://www.unece.org/stats/documents/2006.03.metis.htm>)) and sent it to the UNECE secretariat either by e-mail (jessica.gardner@unece.org) or by fax (+41-22 917-0040). On the first day of the session, delegates are requested to present themselves at the Villa les Feuillantines, UNOG Security and Safety Section, for the issuance of an identification badge. In case of difficulty, please contact by telephone the UNECE secretariat on +41 (0)22 917 2084.

11:05	Presentation of session (i) supporting papers (30 minutes)	
	Supporting papers:	
	The statistical and geoscientific IBGE's Metadata System Luigino Italo Palermo, Metadata Project Manager (Instituto Brasileiro de Geografia e Estatística - IBGE, Brazil)	WP.4
	The BNSI Experience in Metadata Collection and Organization Svetlana Ganeva (Bulgaria National Statistical Institute)	WP.5
	Experiences in developing a metadata system for the Czech Statistical Office Ebbo Petrikovits (Czech Statistical Office)	WP.6
11:35	General discussion, summary and conclusions from session (ii) (1hour)	
12:35	Lunch break (90 minutes)	
14:05	Introduction to session (ii) and presentation of invited papers (1h 25m)	
	Session (ii) Metadata Concepts, Standards, Models and Registries Session Organizer: Paul Johanis (Statistics Canada) This topic includes issues relating to: <ul style="list-style-type: none"> terminology - development of standard terminology at agency, national and international levels and vehicles for the promotion of standard terminology; definitions, purposes and taxonomies; metadata standards, guidelines and recommendations at the agency, national and international levels (current and those in the process of being developed); common metadata constructs; metadata models; metadata registries and registration processes; interchange of data and metadata at both the national and international levels. 	
	Invited papers:	
	Metadata Standards and their support of data management needs Daniel W. Gillman (US Bureau of Labor Statistics) and Paul Johanis (Statistics Canada)	WP.7
	SDMX and ISO/IEC 11179 Arofan Gregory and Chris Nelson (SDMX standards Team)	WP.8
15:30	Break (20 minutes)	
15:50	Presentation of session (ii) supporting papers (1 hour)	
	Supporting papers:	
	Conceptual Modelling of Statistical Metadata and Metadata Data Model in CoSSI Heikki Rouhuvirta (Statistics Finland)	WP.9
	SORS thesaurus of statistical terms Joza Klep (Statistical Office of the Republic of Slovenia)	WP.10
	Statistical Metadata Model Developed at Spain: Current and future use and applications Maria del Mar Blanco Frias (Instituto Nacional de Estadística, Spain)	WP.11
	The Nature of Data Frank Farance and Daniel W. Gillman (US Bureau of Labor Statistics)	WP.12
	Using the MCV terminology for mapping metadata from different institutions: the case of Eurostat and OECD Marco Pellegrino (Eurostat) and Denis Ward (OECD)	WP.13
	Recent achievements in metadata systems at the Romanian National Institute of Statistics Lucian Sinigaglia (National Institute of Statistics, Romania)	WP.14
16:50	General discussion, summary and conclusions from session (ii) (70 minutes)	
18:00	End of Business (EOB)	
18:15	Social programme – <i>Cocktail reception at the Palais des Nations (Bar 13-15)</i>	

Tuesday, 4 April 2006		
9:00	Introduction to session (iii) and presentation of invited papers (2 hours)	WP.15
Session (iii) Metadata and the Statistical Cycle Session Organizer: Graeme Oakley (Australian Bureau of Statistics) The session will focus on the role of metadata in different phases of the statistical cycle: <ul style="list-style-type: none"> • survey planning and design; • survey preparation; • data collection; • input processing; • derivation, estimation, aggregation; • analysis; • dissemination; • post survey evaluation. 		
Invited papers Statistics New Zealand's End-to-End Metadata Life-cycle Gary Dunnet (Statistics New Zealand) Statistical cycle focus in metadata captured for surveys Alice Born (Statistics Canada) Using the metadata in statistical processing cycle – the production tools perspective Matjaž Jug, Pavle Kozjek and Tomaž Špeh (Statistical Office of the Republic of Slovenia) Reengineering projects focussing on metadata and the statistical cycle Ashwell Jenneker (Statistics South Africa)		WP.16 WP.17 WP.18 WP.19
11:00	Break (30 minutes)	
11:30	Presentation of session (iii) supporting papers (30 minutes)	
Supporting papers: Dissemination of Statistical Data and Metadata: Process Based on Common Structure of Statistical Information (CoSSI) Harri Lehtinen (Statistics Finland) Quality Infrastructure System - a Case Study of an E2E Application at the ABS Graeme Oakley (Australian Bureau of Statistics) Metadata as a crucial starting link in new statistical cycles Harry Goossens (Statistics Netherlands)		WP.20 WP.21 WP.22
12:00	General discussion, summary and conclusions from session (iii) (1 hour)	
13:00	Lunch break (90 minutes)	

14:30	Introduction to session (iv) and presentation of invited papers (1 hour 30 minutes)	
Session (iv) Implementation		
Session Organizer: Max Booleman (Statistics Netherlands)		
The Statistical Metadata Framework developed following the 2004 METIS meeting includes information about practical experiences of national statistical offices that have recently implemented or re-engineered their statistical metainformation systems.		
The discussion on this topic will focus on:		
<ul style="list-style-type: none">• implementation planning and management;• identification of recommended practice in the implementation of metadata systems, etc;• usability considerations;• infrastructure development options (e.g. build, buy, sharing and collaboration between NSOs, open source software);• updating and improving statistical processes;• change management:<ul style="list-style-type: none">○ influencing corporate culture (includes communication plans);○ transition planning and management.		
Invited papers		
StatLine 4 Metadata Implementation Edwin de Jonge (Statistics Netherlands)		WP.23
Developing a System for Description of Microdata at Statistics Sweden Klas Blomqvist and K.E Kristiansson (Statistics Sweden)		WP.24
Implementation of MetaStore in OECD Russell Penlington and Lars Thygesen (OECD)		WP.25
16:00	Break (20 minutes)	
16:20	Introduction to session (iv) and presentation of supporting papers (30 minutes)	
Supporting papers		
Development of Metadata System at Croatian Bureau of Statistics Maja Ledić Blažević (Croatian Bureau of Statistics)		WP.26
Using SDMX standards to achieve rapid dissemination of key short-term indicators on the European economy Bengt-Ake Lindblad, Leonhard Maqua, Marco Pellegrino and Giuseppe Sindoni (Eurostat)		WP.27
Statistical Metadata in Statistics Norway Anne Gro Hustoft and Jenny Linnerud (Statistics Norway)		WP.28
16:50	General discussion and conclusion of session (iv) (70 minutes)	
18:00	EOB	
18:15	Meeting of the Metadata Framework Task Force (<i>Task Force members only</i>)	
Wednesday, 5 April 2006		
9:00	Information Session: Establishment of an Electronic Data Reporting Task Force Juraj Riecan (UNECE)	WP.29
9:30	Open Discussion (1 hour)	
10:30	Break (30 minutes)	
11:00	Finalising the Framework – summary of proposed actions and timeframes, resolution of any outstanding issues	
12:30	Lunch break (90 minutes)	
14:00	Future work in the field of statistical metadata (1 hour)	
15:30	Adoption of the report and summary of main conclusions (1 hour)	Draft report
16:30	EOB	

**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 (i): Metadata in a Corporate Context

**PART A OF THE METADATA FRAMEWORK:
STATISTICAL METADATA SYSTEM AND ITS ROLE IN A STATISTICAL
ORGANIZATION**

Invited Paper

Submitted by UNECE and Australian Bureau of Statistics¹

I. INTRODUCTION

1. At the February 2004 METIS work session, a Task Force was established to draft a Statistical Metadata Framework for the UNECE region. The framework was divided into four parts: A –D. The contents of which align to the topics being discussed at this work session.

2. This paper is a draft of Part A: Metadata in the Corporate Context. Part A aims to outline issues surrounding the development and implementation of a Statistical Metadata System (SMS). It is intended particularly for senior managers within a statistical organization.

3. This draft is submitted for review and your input is requested. To facilitate this, we pose some questions for discussion:

- a. Does the group agree with the proposed content of Part A? Are any new sections needed, or should some be removed? For example, do we need to introduce 'access and use' metadata considerations?

¹ Prepared by Jana Meliskova (UNECE Consultant) and Graeme Oakley (Australian Bureau of Statistics)

- b. Do any METIS members have contributions to make to any sections? For example, case studies, examples, suggestions for additional principles, benefits, examples of vision statements?
- c. Is any discussion necessary about questions of detail – for example, terminology - use 'vision' or 'model'?; more exploration of use of metadata by external users, etc.
- d. Finally, does the METIS group support, in principle and subject to application of amendments discussed, this draft? What should be the next steps with respect to Part A? Who would be involved in this work? Timeframe?

STATISTICAL METADATA FRAMEWORK DOCUMENT

Draft

Jana Meliskova & Graeme Oakley

20.02.2006

TABLE OF CONTENTS

A. Statistical Metadata System and its role in a statistical organization	5
A.1. Executive overview	5
A.2. Corporate Value Proposition for Metadata Management	7
A.2.1. Statistical Metadata System (SMS)	7
A.2.2. Metadata users	8
A.2.3. Benefits of a Statistical Metadata System	10
A.3. Metadata management strategies and policy framework	18
A.3.1. SMS Vision	18
A.3.2. Vision Goals	18
A.3.3. Metadata objects and metadata resources	20
A.3.4. SMS Planning	22
A.3.5. Management Strategies for Corporate SMS	24
A.4. Core principles for metadata management	34
A.5. Corporate Governance Models for Metadata Management	35
A.5.1. Introduction	35
A.5.2. Lessons for Good Corporate Governance of Metadata	36
A.5.3. Barriers to the Introduction and Use of Statistical Metadata Systems	37
A.5.3.1. Most important challenges to introduction of metadata systems	37
A.5.3.2. Human issues in relation to adoption of metadata systems	37
A.5.3.3. Organizational issues	38
A.5.3.3.1. The degree of central coordination	39
A.5.3.3.2. Tasks of a coordinating unit	40
A.5.3.3.3. The involvement of different types of specialists	40
A.5.3.3.4. Cooperation with other organizations	41
A.5.4. Case Study - Australian Bureau of Statistics (ABS)	41
Glossary of terms used (to be completed)	44
AMRADS	44
COSMOS	44
Designer	44
GESMES	44
Metadata	44
SDMX	44
Senior Management	44
XML	45
XBRL	45
References	46

TABLE OF FIGURES

Figure 1:	Metadata Users	9
Figure 2:	Schematic View of the SMS Vision and its Components	20
Figure 3:	Example of metadata objects for statistical production process	21
Figure 4:	The SMS Life Cycle.....	27
Figure 5:	SMS Management Model - Cross-cutting Strategy	28
Figure 6:	Model for integrated management of SMS	34
Figure 7:	Perceived challenges to introducing or using statistical metadata systems	37
Figure 8:	Perceived barriers to the provision of effective metadata	38

A. Statistical Metadata System and its role in a statistical organization

A.1. Executive overview

This document aims to outline issues surrounding the development and implementation of a Statistical Metadata System (SMS). It is intended particularly for senior managers within a statistical organization.

An SMS is an important tool for safeguarding the internal and external integration of a statistical information system (SIS). The varying needs of a diverse and large number of users of statistical metadata highlights the strategic nature of an SMS project. Knowing who the users are and understanding their needs, is the foundation for effective SMS development. To ensure its efficiency, the SMS must be an integral part of the strategic direction of a statistical organization. For this reason, it is vital that senior management is directly involved in any SMS project.

The management of an SMS project is a demanding task. Statistical metadata management is a developing field with many researchers and experts, both in statistical offices and in universities, continually contributing to its development. In addition, intensive international cooperation is going on in this area. For example, there have been a large number of European Union (EU) projects dealing with different aspects of statistical metadata management, such as [AMRADS](#)², [MetaNet](#)³, [METAWARE](#)⁴, [COSMOS](#)⁵. Standards and guidelines for statistical metadata have been developed, including [GESMES](#)⁶ standard for exchange of statistical data and metadata and the [UNECE Guidelines for Statistical Metadata on the Internet](#)⁷. There is international cooperation on the development of the [Statistical Data and Metadata Exchange](#)⁸ (SDMX) project, which aims to develop standards for metadata and data exchange between international organizations and national statistical offices.

Recent national experiences show that the successful development of the SMS must directly engage the senior management of the organization. Isolated involvement of metadata and IT experts is not sufficient. Furthermore, a shift from needing IT expertise, towards needing expertise in content oriented statistical issues, is clearly evident.

In the past, the prevailing, and very often a unique role of metadata in a statistical organization, was to support the production of official statistics. Today, the SMS should address many other important requirements. It should be a tool for the efficient functioning of the whole SIS; for the organization of statistical services; and for the systematic cooperation with major stakeholders of statistical data and metadata. In this framework, the SMS should be a self-sustainable project, supporting major functions of the SIS, including its further

² Accompanying Measure to Research and Development in Official Statistics (AMRADS) website at <http://amrads.jrc.cec.eu.int/>.

³ MetaNet website at <http://www.epros.ed.ac.uk/metanet/index.html>.

⁴ METAWARE website at <http://europa.eu.int/en/comm/eurostat/research/retd/metaware.html>.

⁵ Cluster of Systems of Metadata for Official Statistics (COSMOS) website at <http://www.epros.ed.ac.uk/cosmos/>.

⁶ GESMES/TS (formerly called GESMES/CB) is the message used by the European Central Bank to exchange statistical data and metadata with its partners in the European System of Central Banks (ESCB) and other organisations world-wide. For more information see the website at <http://www.ecb.int/stats/services/gesmes/html/index.en.html>.

⁷ Available online at <http://www.unece.org/stats/publications/metadata.pdf>.

⁸ Statistical Data and Metadata Exchange website at <http://www.sdmx.org/>.

development. It requires a corporate and systematic management of all stages and activities dealing with SMS design, implementation and use.

The material in this framework concentrates on the major issues important for the corporate management of an SMS project.

Chapter **A 2, Corporate Value Proposition for Metadata Management**, delineates the role and functions of the SMS for a statistical organization. It describes the major users of statistical metadata and the benefits provided by the SMS project.

Chapter **A 3, Metadata Management Strategies And Policy Framework**, is devoted to the management and preparation of the corporate vision of SMS for the statistical organization. It presents potential objects for metadata description and formulates recommendations for the development of metadata management strategy and planning.

Chapter **A 4, Core Principles For Metadata Management**, formulates the most important principles and recommendations for managing the design and implementation stages of an SMS project.

Chapter **A5 -- Corporate Governance Model** outlines methods for good governance of an SMS project. It describes barriers, challenges, human and organizational issues and provides a case study of the governance model used by the Australian Bureau of Statistics.

A.2. Corporate Value Proposition for Metadata Management

A.2.1. Statistical Metadata System (SMS)

What is a Statistical Metadata System?

The definition “metadata is information about information” predetermines that the Statistical Metadata System (SMS) informs about the Statistical Information System (SIS).

In general, metadata has two basic functions. The first is to uniquely and formally define the content and links between objects and processes of the SIS. The second function is to determine all related technical parameters. When designing the SMS, priority should be given to issues relating to content.

In an environment of rapid development of information and communication technologies, developing efficient strategies for the production and dissemination of statistics is a challenge. The growing use of the Internet has caused significant change in the priorities of the SMS functions. In the past priority was often given to technical metadata and IT challenges, whereas now there has been a clear shift to prioritizing content and methodological issues.

Due to these changes, integrated and transparent description of information flows inside and outside statistical offices has become increasingly inevitable. The use of technology for data collection, interactive communication with users and dissemination of statistics, calls for a coherent and well functioning SMS.

SMS should be a self- sustainable project. Its implementation should be independent of the technology employed for the statistical data processing. However, the links between SMS and e-processing systems must be ensured. Processing of statistical data should be driven by metadata stored in SMS.

What is the Role of the SMS?

The success of an SMS can be measured by the extent to which the needs of diverse groups of statistical metadata users are satisfied. The need for metadata is defined by the various activities, tasks and processes carried out inside a statistical organization. All those activities and processes make up the SIS and strategy of the statistical organization. Therefore, the role of SMS should be understood in the framework of processes and activities of SIS.

In this context, the SMS should be a tool enabling a statistical organization to perform effectively the following main functions:

1. Production of official statistics. Management of all phases of statistical data production (data collection, storage, evaluation and dissemination).
2. Planning, designing, implementing and evaluating statistical production processes.
3. Management of methodological activities. To use coherent metadata in statistical methodology is of a primary importance.

4. Management of cooperation with end users of statistical data and information. Facilitation of user feedback.
5. Enhanced availability of statistical metadata and data for clients. Improved discovery and exchange of data between the statistical office (SO) and its users.
6. Improved quality of statistical data. Observing and evaluating the quality of statistical data is one of the most important goals of statistical activities. To this end, national and international SOs have adopted a set of criteria (relevance and completeness, comparability and coherence of statistical concepts, accuracy of statistical estimations, timeliness and punctuality of delivered statistical information, its accessibility and clarity). SMS should offer a relevant set of metadata for all of these criteria.
7. Management of statistical data sources and cooperation with respondents.
8. Dissemination of statistical information to end-users. End users need reliable metadata for searching, navigation, and interpretation. There should be also metadata available to assist post-processing of statistical data.
9. Improved integration of SIS with other national information systems. There is a growing need to use administrative data for statistical purposes. It calls for better integration and sharing metadata among statistics and state administration in order to ensure coherence and consistency of exchanged information.
10. Improved integration of SIS with information systems of international organizations. International organizations (eg Eurostat, OECD, UN, IMF and others) are increasingly requiring an integration of their own metadata with metadata of national statistical offices in order to make the flow of statistical information more comparable and compatible.
11. Management, unification and standardization of the workflows inside the SO.
12. Knowledge base on the processes of SIS. It enables also to share such knowledge among the statistical staff and to minimize the risk related with its migration.
13. Improved administration of SIS encompassing namely responsibilities, legislation, performance, users' satisfaction.
14. Facilitate the evaluation of costs and revenues for the SO.
15. Unified conception of statistical terminology as a vehicle for better communication and understanding between managers, designers, subject matter statisticians, methodologists, respondents and users of SIS.

A.2.2. Metadata users

A primary challenge for the SMS is to cope with the requirements of diverse metadata users. The use of various information and communication technologies has meant that users of statistics have increased and diversified. Effort should be made to understand who the users are, as their requirement for data and metadata may vary substantially. According to the goals of national statistical services the four major groups of statistical metadata users could be specified (see *Figure 1*):

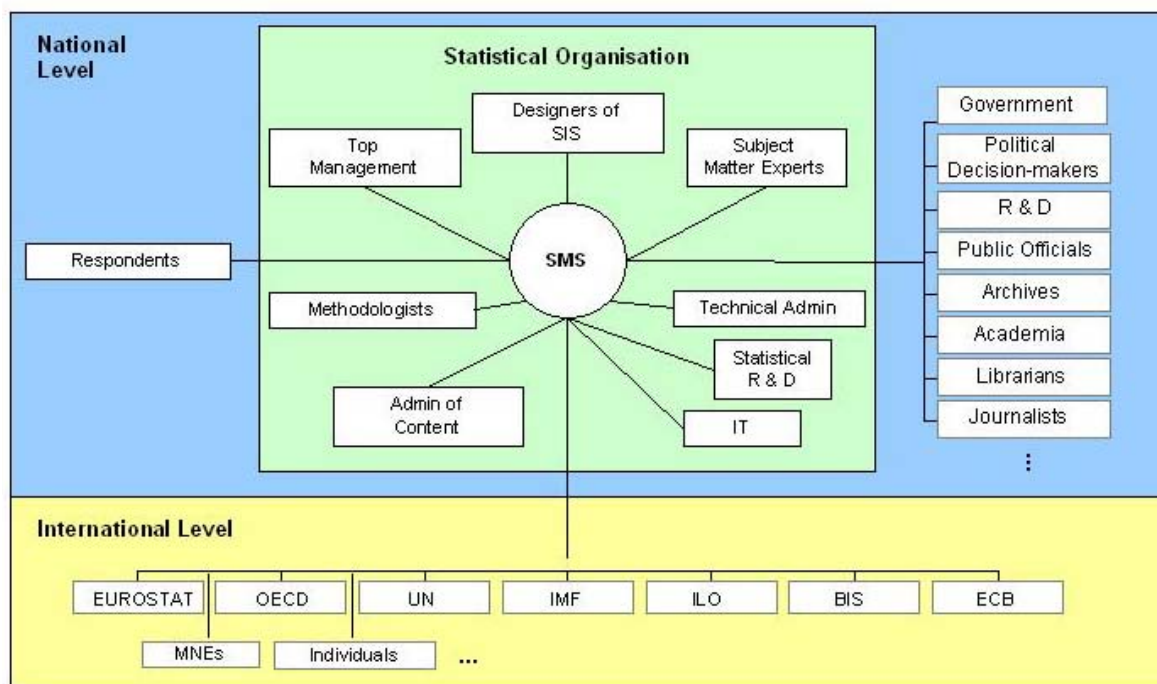


Figure 1: Metadata Users

(i) Users inside SO

This group of metadata users encompasses the many professions involved in the phases of preparation, production and dissemination of official statistics and the functioning of the SIS. These include the following metadata users:

- Senior management;
- Designers and evaluators of SIS;
- Methodologists;
- Subject matter statisticians;
- Statistical research and development;
- Administration of metadata content;
- Technical administration of metadata;
- IT unit responsible for statistical data processing;
- Dissemination specialists.

(ii) Respondents

This group consists of those that supply statistical data to the SIS. Information and communication technologies bring statistical data providers and statistical users closer together. Some institutional statistical activities, particularly in the health, education and justice sectors, require shared access to microdata. In these cases, respondents are both suppliers and users of metadata. Special attention was once given to the suppliers of administrative data on business enterprises. However, with the spreading use of Internet, the growing number of statistical users and their requirements calls frequently for new data sources and thus new data providers. The number of statistical respondents is changing rapidly.

(iii) End users on national level

This group includes: governmental institutions, political decision makers, researchers, public officials, archives, academia, librarians, journalists and the general public. As the audience of users grows it also diversifies. In the past, data dissemination methods typically assumed a certain level of economic and statistical sophistication. Frequently, users' knowledge for a given set of statistics was comparable to the knowledge of subject matter statisticians. This is clearly no longer the case. The audience for economic statistics, for example, can range from professional economists and policy makers, to interested members of the general public, to students working on school assignments. An understanding of economic and statistical concepts can no longer be taken for granted.

(iv) International users

Individuals, multinational enterprises (MNEs), international organizations and others are becoming important users of statistical metadata. As for international organizations at least the following most important should be mentioned: the Organization for Economic and Co-operative Development (OECD), Eurostat, the United Nations Statistical Division (UNSD), the United Nations Economic Commission for Europe (UNECE), the International Labour Organization (ILO), the International Monetary Fund (IMF), the Bank for International Settlements (BIS) and the European Central Bank (ECB). Integration of metadata from national statistical organizations with the statistical metadata of international users is becoming increasingly necessary.

A.2.3. Benefits of a Statistical Metadata System

Statistical organizations and other metadata stakeholders can benefit when metadata exists from creation until archival, rather than as captive to a particular statistical processing system or infrastructure package. The following benefits are valid for all categories of users outlined above:

- Improved statistical information and more efficient operations;
- Improved quality of metadata;
- Better location, retrieval and exchange of data and metadata between organizations to enhance availability to users of statistics;
- Use of a common set terminology, names and descriptions for standard metadata elements to improve communication;
- Central metadata repositories organized to facilitate reuse of existing data;
- Increased use of metadata standards; and
- Improved knowledge of metadata flows.

A.2.3.1. Benefits for Internal Users

Senior Management

SMS facilitates design, planning, decision-making and evaluation processes of SIS. Dissemination strategy, cooperation with end users on national and international level and

data suppliers would primarily belong to those processes. SMS should provide the tools for answering the questions like: to what extent do users actually use the statistical outputs? Are they satisfied with the qualities of data and metadata with regard to content, accuracy, timeliness, availability and coherence? Are there complaints or unmet demands from respondents? SMS should help in giving answer to these questions and should also serve as the administrative management of statistical system. And, last but not least, senior managers of SIS will be interested to learn about the costs and benefits of individual statistical activities.

For these purposes especially the following metadata will be needed:

- End users needs and other stakeholders requirements on a national and international level;
- External information systems related to SIS;
- Suppliers of data into SIS with special attention to the state administration and business enterprises;
- Sources of data for SIS;
- Available statistical services;
- Statistical publications, publication calendar, copyrights and others;
- Statistical production process;
- Responsibilities inside the SO, legislation, performance;
- Cost and revenues of the SO.

Designers and Evaluators

Designers and evaluators of a statistical system are responsible for the design, implementation, maintenance and evaluation of statistical systems.

Planners and evaluators need access to metadata from similar systems, either within or outside the organization, to inform the design, development and implementation of a new system.

For existing systems under their responsibility, they need feedback about performance (qualities and costs), usage and users' satisfaction.

When designing and developing statistical system the following information is required:

- How similar systems have been designed in the past;
- What observation data is already available and how these data can be observed;
- How can this data be obtained; and
- What methods, tools and software components are available and how can they be used.

For maintenance and evaluation of statistical system the following information will be needed:

- Detailed, up-to-date documentation of the system;
- Feedback information, both formal and informal, concerning production and usage of the SIS;
- Experiences from similar systems;
- Knowledge about methods, tools and software components;
- Special evaluation studies performed on an ad hoc basis.

Methodologists

An SMS creates a framework for design and implementation of statistical tasks and surveys to meet statistical obligations in production of official statistics and needs of end users. The SMS provides tools for safeguarding the integration of SIS at national and international level. Furthermore, maintenance, use and further development of statistical classifications and nomenclatures, use of statistical registers, evidence and maintenance about statistical standards, knowledge about statistical methods and relevant research methods, are all activities for which an SMS is indispensable. This group of users will operate namely with metadata relating to the following:

- Content of available statistical data (microdata, macrodata) and associated data concepts;
- Quality of statistical data (relevance, timeliness, accuracy, availability, coherence and comparability);
- Existing statistical tasks and surveys (questionnaires, other sources etc);
- End users and their feedback;
- Requests of international organizations and related standards;
- Data sources and their links;
- Respondents' information systems;
- Administrative data;
- Information systems and their output databases (portals)
- Statistical registers (population, farms etc);
- Statistical classifications, nomenclatures and related international standards;
- Statistical population, statistical units, measurement units time series;
- Statistical methods and relevant research projects.

Subject Matter Statisticians

The subject matter statistician is the expert in a particular field of statistics within a national statistical organization. They have the crucial role of understanding the users information requirements, in the context of the policy and program decision making of the users, and the capabilities of their national statistical office, ie what they can do to provide the required information. Subject matter staff work with other NSO specialists to design and construct an appropriate survey and generate statistics. However, the statistician then has the role of communicating the information to their user community through the creation of statistical products and the provision of associated metadata to assist users in understanding the results. Evaluation is also an important responsibility for the subject matter specialist.

Given these roles, the SMS (in very broad terms) is a knowledge management system for the subject matter statistician. In this information system ideally) they would want to be able to create, update, search, browse and retrieve many different types of metadata entities that would cover many aspects, such as:

- users (customers) requirements;
- standard concepts, data elements and classifications;
- operational information and quality metrics about the operation of their survey system;
- documentation about statistical techniques (methodology) applied to their survey; and
- products created from the statistical data.

The benefits of an SMS to the subject matter statistician include:

- a knowledge base about their statistical collection, including all previous cycles. This is an invaluable resource for new employees coming into a statistical field and for statisticians in other fields who might be researching a new collection - there may be elements in another survey that can be reused.
- access to a consistent store of standard classifications, data elements, process engines that can be used in new survey development with the knowledge that using these elements will assist greatly in ensuring statistical integration.
- as the SMS is a corporate facility, then it would be expected that many tools will be provided that utilize the information repository of the SMS, for example, the product creation environment of the NSO would use the SMS as a source of metadata and so enable the subject matter statistician to more easily create statistical products for the organization's web site with a 'common look and feel'.
- associated with the SMS are standard processes eg registration of new data elements, which would provide a common method across the organization for the subject matter statistician to create and use metadata, thereby reducing training efforts because of various local solutions. There would be better support and consultation services because of a common SMS, and more employees working with the same facilities.

Statistical Research and Development

A science of SIS should be interdisciplinary. It requires a good knowledge and understanding of statistical methodology but this is not enough. At the same time, it requires knowledge about technologies and methods used in information management, information systems and computer science. Furthermore, scientific studies of statistical system would need contributions from behavioral and economic sciences and other disciplines. Researchers will need similar kind of metadata as for designers and methodologists who work on more corporate level (not only on individual surveys and production systems). In addition, SMS should ensure the following metadata specific for the research purposes:

- General knowledge about statistical systems and statistics production (e.g. recognized theories and methods, standards, current best methods, current best practices);
- Specific knowledge and experiences from different statistical organizations;
- Costs and quality aspects in SIS processes.

Administration of Metadata Content

SMS should ensure smooth and systematic update and maintenance of statistical metadata. Maintenance of metadata content will be performed through firmly installed network in which subject matter specialists and methodologists, responsible for metadata content will cooperate. Metadata should be updated by the Administrator of the SMS corporate metadata repository (CMR), once only and in one place. This will help avoid inconsistencies and unnecessary redundancies. All linked updates in CMR have to be performed automatically, without any further human interference. Furthermore, it should ensure errorless metadata navigation of e-production of statistical data. Administrator will need a user-friendly interface, avoiding any special technical skill. To this end the administrator will need the following metadata:

- All metadata related to the content of and links between statistical metadata;
- Information about organization of metadata in CMR;
- Metadata allowing discovery and retrieval;

- Updating methods and procedures;

Technical Administration of Metadata

Technical administrator (IT expert) will use SMS tools for technical maintenance of the CMR. They should cooperate with designers, evaluators and content administrators in solving technological aspects and further development of SMS. The technical administrator will use, oversee and maintain the following metadata:

- Technical metadata related to the CMR, and to the links for e- production systems;
- Information and knowledge about technological aspects of statistical production;
- Information about technical links to other information systems.
- Information about tools and software used by content administrator.

IT Unit Responsible for Statistical Data Processing

Important metadata users are those people operating and monitoring the statistical e-production process.

Metadata driven statistical production creates favorable conditions for standardization and thus efficiency of statistical production system. Metadata on the content of statistical data and associated concepts, including all other delimiting metadata (statistical classifications, statistical units, measurement unit, time series, statistical population etc), are a key condition for the whole throughput of production phases (data collection, storage, evaluation and dissemination). Technical metadata on the organization of CMR and links to the production systems belong to the metadata set needed for fulfilling functions of e-processing.

Ideally, statistical production processes will generate metadata about their own performance, giving producers feedback about functioning and efficiency of metadata driven production. In this respect, producers should cooperate with SMS designers, subject matter specialists and methodologists, content and technical administrators on the design, implementation, evaluation, and further development of the SMS.

A.2.3.2. Benefits for Data Providers

Respondents are important partners of any SIS. Statistical data suppliers are often also the users of statistical data. Their role is becoming more important with the growing number of systems and on-line communication possibilities. In the past it was sufficient for respondents to know requests for statistical data in the framework of the methodological definition of statistical questionnaires, the requests of data suppliers nowadays are more demanding. Bearing in mind the possibility of on-line supply from respondents' information systems to the SIS and the possibility of on-line access of respondents to the SIS it is evident that the requests of data suppliers change. SMS will play a key role in those tasks.

As for the content, there is a growing need to harmonize methodological definitions of data and related metadata from respondents' and statistical information system. The attention should be drawn to the implementation and use of relevant technological metadata standards.

Within the business information systems the standard [XBRL](http://www.xbrl.org/)⁹ (Extensible Business Reporting Language) is frequently introduced as a technical metadata standard. Especially for statistical purposes the metadata standard [SDMX](http://www.sdmx.org/)¹⁰ (Statistical Data and Metadata Exchange) has been developed in cooperation between IMF, OECD, Eurostat, UN, BIS and ECB. Metadata Common Vocabulary developed in the framework of the SDMX will be an important metadata standard used by both suppliers of statistical data and SOs as suppliers of statistical data to international organizations. Data suppliers will require from SMS especially the following information:

- Metadata related to the content (definitions, terminology) of statistical data in the input stage of the statistical production;
- Security and confidentiality of microdata;
- Feedback from statistical surveys;
- Information about the content of statistical warehouses;
- Knowledge about comparability of statistical and respondents data/systems;
- Technical parameters for search and retrieval of metadata in CMR and links to statistical warehouses;
- Knowledge about potential interface between SIS and respondents' information systems;
- Relevant technological standards for metadata and data e-supply;
- Information about software and other tools supporting e-supply of data and metadata;
- Information about strategies for further SMS development;
- Training in use of SMS;

A.2.3.3. Benefits for End Users on the National Level

Understanding different communities of end users and their classifying could help in classifying users requirements. SMS will help users to better discover, understand, interpret and interrogate needed data. The proliferation of information has raised the issue of consistency and comparability of data. Comparability of data is desirable, but not always possible. It is important to know what the differences are and the reason for them, explicated to the different level of users' sophistication. SMS will also assist to convey the credibility of statistical data and recognizing intellectual property.

It is important to monitor users feedback and to embrace the need for metadata in both directions. SMS will offer the possibility to understand how the users search and the terms/terminology that they use. SMS will also support handling access of users to microdata. The fact that users are increasingly requesting access to microdata, calls for tools that allow concerns about confidentiality protection to be overcome.

With spreading use of Internet it is important to provide clients with maximum information about statistical outputs via statistical websites. However, numerous statistical websites are offering diverse metadata to users for identifying and seeking statistical information. There is a potential to flood users with too much metadata. Appropriate communication of metadata

⁹ XBRL website at <http://www.xbrl.org/Home/>.

¹⁰ SDMX website at <http://www.sdmx.org/>.

should be based on principles of 'cognitive psychology', that is, there is a presentational aspect to metadata consumption.

This heterogeneity, together with more visible methodological differences and inconsistencies of statistics disseminated via Internet, poses difficulties for the users. Clearly, there is a need for a harmonization of metadata accompanying statistical information on Internet. Important role in this respect should play UN international standards ("Guidelines for statistical Metadata on Internet", UN-CES Statistical Standards and Studies –No 52).

Last but not least, SMS should support integration of statistical output databases and portals with the portals of other external institutions.

The following metadata is vital for end users of statistical metadata and data at the national level:

- Availability of statistical outputs;
- Metadata related to the statistical outputs (metadata and data concepts and definitions, classifications, aggregations, statistical and evaluation methods, terminology, history, etc);
- Coherence, comparability, explanatory notes;
- Access to microdata;
- Timeliness;
- Time series;
- Updating procedures;
- Statistical revisions;
- Responsibility for individual statistical outputs;
- Links to other information systems both national and international;
- Confidentiality;
- Planned changes in statistical outputs;
- Content related standards, both national and international;
- Statistical websites¹¹;
- Statistical output databases;
- Outcomes from statistical analysis on users feedback;
- Rules for searching, accessing and downloading statistical metadata and data from output databases;
- Technological standards relevant for extraction and transfer of data and metadata;
- Information about software and other tools supporting e-search, retrieval and downloading of metadata and data;
- Users training possibilities.

A.2.3.4. Benefits for International Users

There are more and more demands by international users for greater consistency when interacting with NSOs. In the case of international organizations, the metadata and data requirements (and their collection and exchange) have to be coordinated not to overburden

¹¹ The UNECE published A Guide to the Websites of National and International Statistical Organizations, 2001. Available online at <http://www.unece.org/stats/publications/Webguide.pdf> [accessed 27 January 2006].

countries with duplicate requests. In order to fulfill this task, better integration of metadata at the national and international level is needed.

A lot of metadata is available on websites of international organizations. Links could be inserted from the metadata of international organizations to more detailed metadata on national websites. Coordination of access could be achieved through a single gateway for data and metadata, e.g. through a portal side. To this end, the IMF launched a joint project, together with OECD, Eurostat, UN, BIS and ESB. The Task Force on the project called Statistical Data and Metadata Exchange (SDMX) was created. SDMX standards are at present under intensive development with Version 2.0 released in November 2005.

Another example is dealing with the multinational enterprises. MNEs can be significant in terms of a nation's economy. To understand the behavior and impact of MNEs, it is important to assess the effects of globalization. MNEs' information systems, however, may not correspond to concepts and models of the SIS. Such situations can potentially lead to gaps and anomalies in the measurement of the activities of MNEs by national SOs. National SOs should explore whether there are biases in national economics caused by gaps and overlaps in the coverage of activities of MNEs. To this end, standardization in the following areas will be needed: definitions of forms of organizations, statistical units, charts of accounts and classifications. Fulfilling such requests without existence of a coherent SMS would be very difficult.

Needs of international users increasingly impact the architecture of national SMS. National SOs face new tasks that can be solved only in close cooperation with international organizations and other international users.

Metadata needed by international users are quite identical with those needed by end users on national level (see the subchapter above). Furthermore, the following information would be required:

- Complying with international standards (coherence, comparability, explanatory notes);
- Standards used for e-metadata and data transfer (XBRL, SDMX, GESMES, others);
- Information about other international and national users;
- Indication of needs for revision and/or standardization of statistical data and metadata concepts.

A.3. Metadata management strategies and policy framework

The focus of this chapter is on the preparation of a corporate SMS Vision, related planning and on the major characteristics of a metadata management framework and management strategies.

A.3.1. SMS Vision

This subchapter presents major goals and functions of the Vision. Furthermore, it assists to understand better what could be the objects of metadata description related to the functions defined in the Vision.

The Vision should clearly state the goals or aims of the SMS. It should apply across the entire SIS and be realistic and within the capabilities of the SO. It should also include a statement about scope: what is included in the project and what is not.

A.3.2. Vision Goals

An important prerequisite for successful design, implementation and functioning of the SMS is the development of a corporate Vision of SMS in the statistical organization. The functions of SMS, centered upon metadata and data users, are oriented towards the diverse processes and activities of SIS. Organizational units within a statistical agency, respondents and end users are all involved in the preparation, implementation and use of the SMS tools. The Vision should be developed with the direct involvement of senior management within the statistical agency.

The Vision should be an integral part of the strategic direction of the statistical organization. It is an important task for the SMS management to ensure that not only the development of the Vision but also the SMS design, implementation and further development will be monitored by senior managers. For this purpose a relevant management structure of SMS should be established. Feedback and evaluation, supported by metadata accumulated in the previous processing cycles, should be an integral part of the SMS design.

The Vision should define major goals and functions of SMS for the SO (see “ The Role of SMS in the Chapter A1) and attribute the priorities for implementation. It should clearly identify the users of statistical metadata (inside and outside the SO) and determine their rights and obligations in the phase of design and development of SIS.

The metadata requirements associated with each element of standard business are articulated. That is all the points of contact between the metadata model and business processes, in terms of creation, update and use activities should be described.

Important part of the Vision should be analysis of the state-of-art of the existing statistical metadata objects and services, finishing by clear specification what kind of existing metadata can be used in the corporate SMS, what kind of existing metadata and services should be updated and what kind of existing metadata should not be used at all. Especially when the latest mentioned possibility appears, it is desirable to support SO in its decision to cancel such metadata blocks.

It is advisable, that the SMS is not developed as a purely technical project. It is still quite often the case in the SOs that the subject matter departments do not understand fully the requests formulated by the IT specialists. When developing the Vision, it is essential to express clearly that the first priority in the SMS is given to the safeguarding of the content and methodological integration of statistical data and metadata.

To make SMS a success story, the Vision and its functions should be based on the real existing possibilities of the SO. Effective management of SIS and integration process of information flows on national and international levels should remain one of the major goals of SMS.

The Vision should also encompass cost propositions of the SMS project. Costs should be proposed based on the real possibilities of the SO. Warning signs should be made to a very broad (although theoretically correct) requirements for the metadata functions. Such proposals should be very pragmatic, reflecting ultimate needs and metadata priorities. The experience shows that the human capacities and financial factor in the SMS developments could be quite demanding.

Experience shows, that many SOs implemented some functional blocks of metadata without having a complete SMS Vision at the beginning of the process. It is especially true for the objects dealing with the description of statistical data. It can be observed, that namely the following blocks of metadata have been frequently implemented: statistical variables and values sets, statistical surveys, social-economic classifications and nomenclatures, time series, statistical publications, statistical population, economic subjects, statistical units, aggregation and statistical evaluation methods, output tables and others.

Without having a coherent Vision there is very often a lack of coordination among individual metadata blocs. It causes many inconsistencies, duplications and, last but not least, the low efficiency of metadata tools from both, costs and staff capacities needed. The end users could, because of lack of coordination, struggle with unnecessary diversity of users' roles and related diversity of communication metadata languages. Such situation certainly does not stimulate enough joint cooperation of users with statistics on metadata implementation.

The Vision should contain a metadata model complying with the SMS functions. Such model should encompass metadata about data and processes behind them as well as metadata about other objects and processes of SIS relevant to the SMS functions, Metadata needed for the management and administration of statistical system and statistical organization like metadata about the costs and benefits, cost-effectiveness, satisfaction and complains should be also a part of such model. Metadata objects and links between them should be thoroughly defined.

An agreed conceptual metadata model should be linked to the standard business processes that are the part of the statistical life cycle. This linkage is used to determine what metadata should be collected. Metadata model should take account of and uses international standards where possible.

Figure 2 below provides an overview of the components of the SMS Vision.

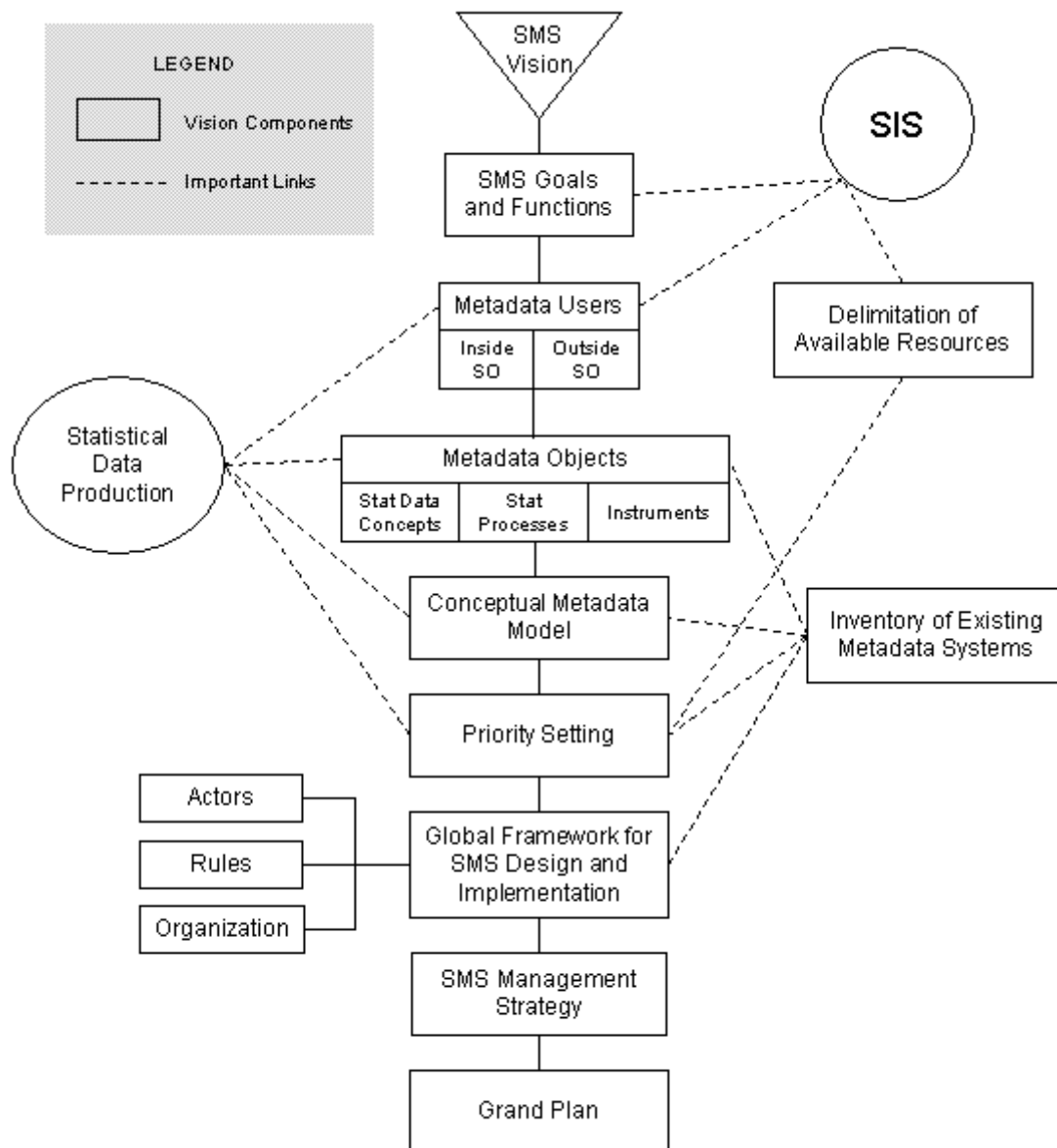


Figure 2: Schematic View of the SMS Vision and its Components

A.3.3. Metadata objects and metadata resources

Metadata should be structured according to the objects and the properties of those objects that they inform about. There are three major categories of metadata objects related to the functions defined by the Vision:

- (i) statistical data and associated concepts
- (ii) statistical processes and associated procedures
- (iii) tools enabling production and usage processes

Different kinds of links exist between individual metadata objects, depending on the task being carried out.

(iii) Tools enabling production and usage processes

SMS should provide tools and vehicles that enable fulfilling the major SMS functions. Instrumental resources can be categorized, according to the functions they are supporting, as follows: (i) search and retrieval tools supporting use and other processes that need access to statistical data and metadata, (ii) production tools supporting statistical production and, (iii) knowledge resources supporting primarily the “intellectual processes” related around statistical system, such corporate management, planning and evaluation, research and development. Instrumental resources should be sharable by multiple processes. They need to be systematized and organized collectively in order to be easy to find and make use of.

In this respect, the Vision should promote the following:

- Development of common terminology for metadata elements across all processes in the statistical life-cycle;
- Development of common and consistent description of metadata elements allowing easy location, retrieval and exchange of data and metadata;
- Development of standard interchange representations allowing sharing of metadata and data between organizations;
- Implementation of consolidated metadata repositories facilitating reuse of metadata;
- Introduction of registration process giving authority to promote use of standard metadata elements and thereby increasing knowledge on metadata flows and statistical integration;
- Improvement of metadata quality;
- Ensure that the production process will be metadata driven.

A.3.4. SMS Planning

The aim of this subchapter is to draw attention to the preparation of a corporate strategic plan for the SMS development. A strategic plan should be an integral part of the SMS Vision, reflecting the goals and functions specified in this document. As a part of the Vision, the senior management of the SO should approve the strategic plan.

The development of a strategic plan needs to be a flexible and adaptive process, possibly with several iterations. The plan should give a visibility, clarity and stability in the development efforts, but aspects are likely to change during its implementation, which may take several years. Certain parts may never be implemented; other parts may be implemented in a different way than originally assumed. Completely new components may appear as a result of new needs, new methodological and technical developments and/or changes of some other basic conditions for the SMS development. Therefore, the plan should be regularly reviewed and revised.

Detailed plans should be developed and approved later on for the design and implementation phases of the SMS development. Such plans should reflect agreed priorities in the solution of individual components of SMS. Last but not least, specific plans should be prepared, of course, for the phases on the SMS use and evaluation.

When preparing a strategic plan, the number of activities, sensitivity of their solution and their priorities for the SO should be taken into the consideration. Links among individual activities and importance of their contribution to the SMS strategic goals should be thoroughly analyzed. Conditions, under which the goals could be carried out, should be clearly specified.

A part of the plan should be establishing of an organizational framework and management strategy.

The strategic plan should be developed and approved by all actors involved in the design, implementation and maintenance of SMS. It is therefore indispensable that such plan is prepared in close dialogue and cooperation with all actors involved in the process of SMS development. The planning could be often made more explicit, so that the whole SO, can discuss the strategies to be used and the choices to be made in the step-by-step development of SMS.

Some Practical Recommendations for Establishing the Strategic Plan:

- When preparing the plan, the SO should consider its current capabilities. Available human and financial resources, as well as organizational and technical feasibility, should be carefully analyzed in order to make the plan realistic.
- Goals defined in the Vision should be transformed into practical steps to which priorities are then assigned.
- Practice shows, that different countries often have similar priorities. This is especially true for the development of databases on statistical classifications and nomenclatures (Nordic countries, Switzerland, France, Australia, New Zealand), aggregated output databases (Nordic countries, Switzerland, the Netherlands, Australia, U.S. statistical agencies, CZSO), and metadata models for the websites (Nordic countries, Switzerland, the Netherlands, Australia). Some countries give priority to the microdata metadata models (the Netherlands, U.S. Bureau of Labour Statistics, Austria).
- Quality of data and metadata should be considered a high priority.
- External cooperation should be clearly defined; categorization and priority setting for external users should be specified. The plan should take the existing working plans of all external partners into consideration.
- The plan should be prepared in such detail that all partners will be able to commit their participation.
- External projects to establish data and metadata warehouses, both on the national and international level, should be considered for potential impact on the SMS.
- External activities on data security and data confidentiality related to the SMS should be considered.

- An integral part of the plan should be activities dealing with the development and implementation of international standards.
- The plan should also consider activities to promote the SMS and create an atmosphere of cooperation with all participating parties. To this end, prototypes for demonstration of SMS functions could be useful.
- Research activities on feasibility studies and analysis of user feedback should be also taken into the consideration when preparing an SMS plan.
- Transfer of know-how and training for participants in the SMS development process should be incorporated in the plan.

A.3.5. Management Strategies for Corporate SMS

The Vision should dictate basic rules for organization and management strategy. Establishment of the management structure for the SMS project should include senior management in a lead role. This senior management involvement should not just be from the technical and methodological areas, but most importantly, should include senior executives from the statistical areas.

A framework for the metadata management strategy should be specified in the Vision. Responsibility for development of metadata policies and procedures and for providing training and advice to developers should be clearly assigned.

An important part of the SMS management strategy should be a systematic cooperation with major metadata stakeholders.

Implementation of the metadata management strategy should follow two broad approaches. They are:

- (i) User orientation – focusing on information relevant to usage such as finding and accessing data, understanding their structure and meaning, assessing their quality and relevancy, and using them correctly. This focus is dissemination oriented; and
- (ii) Producer orientation – metadata driven approach focusing on the needs of information systems and e- processing.

A set of core principles that inform decisions and projects related to metadata should be established. A proposal of core principles for metadata management is presented in Chapter 4.

There are two major dimensions which should be taken into consideration when deciding on SMS management strategy: (i) the crosscutting nature of the SMS role and its functions in statistical organizations and, (ii) the requirement of corporate management during all phases of SMS development and use.

A.3.5.1. SMS management across the whole statistical organization

Diverse organizational units of the SO and external bodies will participate in the SMS project. Senior and middle management, subject matter experts, methodologists, IT units as well as respondents and end users of statistical metadata and data, are all important SMS partners. Functions of SMS partners will differ whether they will participate in the SMS as metadata users, designers, developers, producers, administrators or evaluators.

The points below outline some recommended practices to ensure involvement in SMS management across the whole organization.

- The roles and responsibilities of all partners should be clearly defined, understood and followed. Where possible, automated workflows can be used to enforce agreed role and responsibilities.
- The integration role played by SMS, both inside and outside the SO, and therefore the necessity of senior management involvements, should be clearly recognized when defining the SMS management strategy. The management of traditional statistical activities in the SO is performed in the framework of individual statistical domains, tasks and/or projects in accordance with the organizational structure of the SO.
- Metadata management is part of every project and should be considered alongside resource allocation and accountabilities, in the same way as business processes and data flows are considered.
- Establish SMS management boards to take an ultimate, corporate view on all decisions dealing with the SMS development.
- SMS management strategy should be specified in close alliance with the existing managerial structure of the SO. With the lead role of the senior management in the SMS management model, clear links should be defined also in the middle management level and in the experts' level (methodologists, subject matter statisticians, IT experts). A model showing a crosscutting nature of the SMS management is presented in the *Figure 5*.
- A multidisciplinary team should be the major organizational form for the development of SMS project. The ideal SMS Team(s) will include: statistical methodologists; subject matter statisticians; dissemination specialists; end users; specialists in the implementation of statistical standards; researchers; and IT specialists in data modeling, business process design, architecture and applications development.
- Implementation of a SMS management strategy may highlight the need for change in the organization of statistical activities, particularly where a corporate SMS does not exist. It is especially true for subject matter statisticians. Many critical moments could appear. Such moments should be as much as possible foreseen and reflected when progressing from the definition of the Vision goals and activities to the Vision plans. More detailed considerations about the corporate metadata management are given in the Chapter 5.

A.3.5.2. Good advice for metadata projects

If you are a designer

- Avoid uncoordinated capturing of similar metadata – build value chains instead. Similar, but not identical, metadata may be needed for different purposes. For example, different users of statistics may require metadata of different depth and presented in different ways. Capturing and maintaining similar metadata without sufficient coordination will cause duplications of efforts and may at worst result in metadata that are, or seem to be, in contradiction to each other. A way out of this problem is to create so-called value chains, starting from a basic set of metadata, and then refining these metadata in different directions into metadata end products that are tailor to different needs. For example, a basic documentation of a statistical production system may be refined into different kinds of product overviews and quality declarations.
- Transform data and accompanying metadata in synchronised, parallel processes, fully automated whenever possible.
- Data management processes in statistical organizations have been formalised and automated for a long time. In contrast, the necessary, accompanying metadata have remained manually handled to a great extent, and the manual metadata transformations have not been well coordinated with the automated data transformations. For example, one and the same variable may (more or less unconsciously) have been given different name labels by (a) the designer of a questionnaire, (b) the programmer of data collection and data management software, (c) a programmer of tabulations, and (d) an editor of output tables as presented in statistical publications.

If you are a project coordinator

- Form coalitions around metadata projects.
- All partners of a metadata project should both contribute something and gain something from the project. Preferable each partner should gain more than he or she contributes – a win/win situation. As a collective the coalition should control all resources needed, including the necessary management authority, to make the metadata project a success.
- Make sure that senior management is committed. Most metadata projects are dependent on constructive cooperation from all parts of the organization.
- Organise the metadata project in such a way that it brings about concrete and useful results at regular and frequent intervals.

If you are a senior manager

- Make sure that your organization has a metadata strategy, including a global architecture and an implementation plan, and check how proposed metadata projects fit into the strategy.
- Either commit yourself to a metadata project – or don't let it happen. Lukewarm enthusiasm is the last thing a metadata project needs. There is often scepticism in the organization against metadata projects – for both good and bad reasons. Moreover,

metadata projects are usually strategic projects for the organization. If they should be carried out at all, managers on different levels and in different parts of the organization must be committed to the project. The senior manager is the obvious enabler of this commitment.

- If a metadata project should go wrong – cancel it; don't throw good money after bad money. Metadata projects are often more abstract, more complex, and more difficult to manage than many other types of projects. It is nothing to be ashamed over to fail now and then. But watch out for early signs of fatal problems, and encourage project leaders to admit that problems exist. Discourage idealizing progress reports.
- When a metadata project fails, make a diagnosis, learn from the mistakes, and do it better next time. There is not so much need for finding scapegoats. It is much more important to learn for the future.
- Make sure that your organization also learns from failures and successes in other statistical organizations. Benchmarking and international cooperation is always useful.
- Make systematic use of metadata systems for capturing and organising tacit knowledge of individual persons in order to make it available to the organization as a whole and to external users of statistics.

A.3.5.3. Corporate management of all phases of SMS development and use

The management strategy should encompass all phases of the SMS life cycle. The governance of metadata management and the monitoring of outcomes should be made clear in the SMS management strategy.

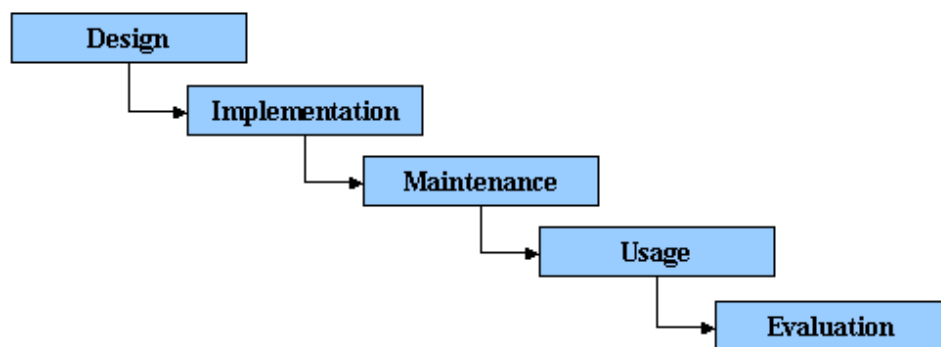


Figure 4 The SMS Life Cycle

As illustrated in *Figure 4*, the SMS life cycle is composed of the following five major phases: (i) SMS design, (ii) SMS implementation, (iii) SMS maintenance, (iv) SMS usage and (v) SMS evaluation. *Figure 5* below presents a model for management strategy of the SMS life cycle.

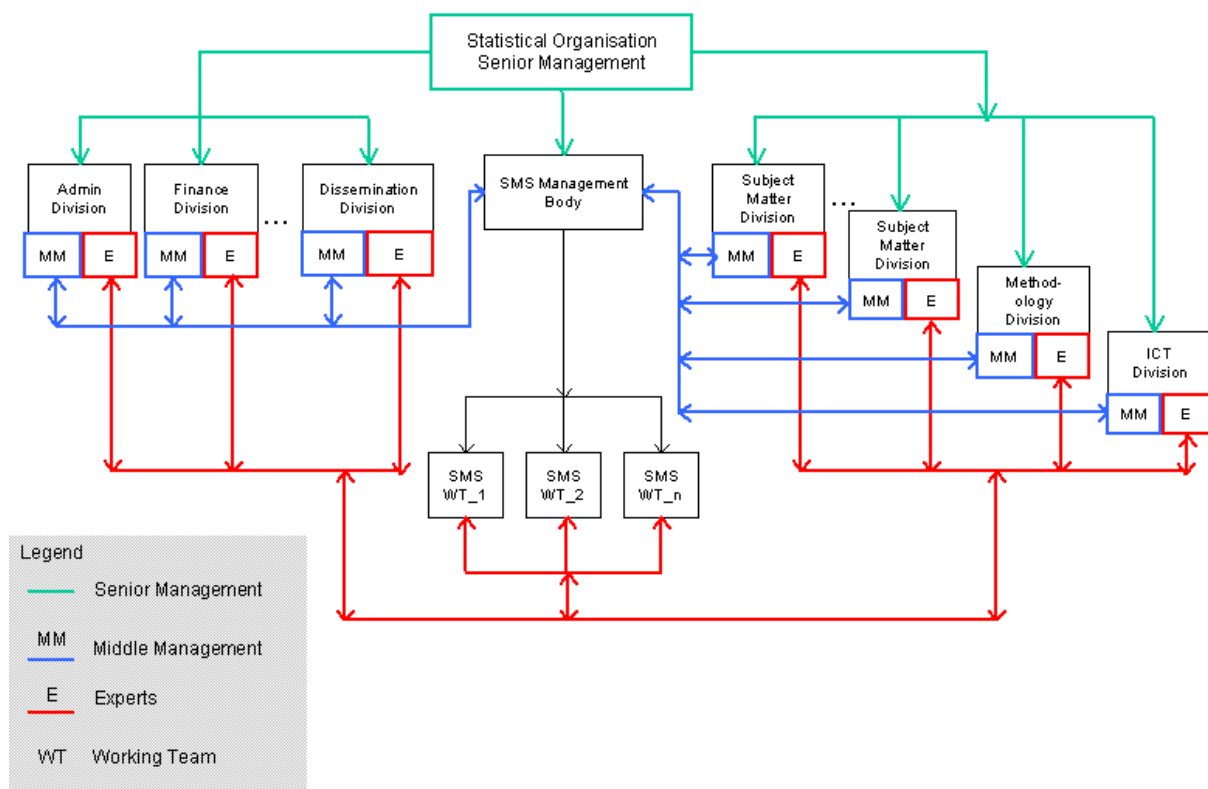


Figure 5: SMS Management Model - Cross-cutting Strategy

The most important activities in the individual phases of the SMS life cycle, which should be taken into the consideration when preparing the SMS management strategy, are outlined below.

(i) Management of SMS design

The role of the design phase is to develop the SMS Vision, global architecture and to establish a management and implementation strategy for the project. The most important functions, tasks and activities for consideration are as follows:

- The development of the SMS Vision is described in the subchapter A 3.1.
- Development of an SMS global architecture. A major strategy for ensuring the overall efficiency and usefulness of the metadata-related work in the SO is to develop a global architecture for all processes that will work with statistical metadata in the foreseeable future. There are several purposes for such exercise. One is to make an inventory of all metadata systems that need to be considered. This inventory should be developed in close cooperation between stakeholders in order to ensure, that nothing important has been forgotten. The inventory could be an excellent basis for putting priorities to different metadata tasks.
- Specification of common components. By analyzing and structuring an inventory it should be clear what metadata components are common for different purposes. For diverse systems, processes and tasks that are designed in a standardized way are able to communicate via standardized interfaces. Such components may be given a higher

priority from a global SMS perspective than they could have got if considered from a local point of view.

- Impact of SMS on existing statistical production system and other processes in the SO is evaluated in the Vision. Processes should be reengineered where necessary in preparation for implementation.
- The metadata requirements associated with standard business processes are articulated, i.e. all the points of contact between the metadata model and business processes, in terms of creation, update, and use activities, is described. For example, metadata associated with understanding user needs, frameworks and standards is acquired and used to inform later phases. To the greatest extent possible, the necessary input and output metadata is captured early on in the collection strategy stage, so that we know well in advance that the desired outputs are obtainable, fit for purpose, etc.
- SMS partners should be committed to participate in all phases of the SMS life cycle. Major partners in the design phase are the users (both, inside and outside SO), methodologists, subject matter statisticians and IT expert.
- Feedback and evaluation is an integral part of the design process and is supported by metadata accumulated in diverse phase of the SMS life cycle.
- Financial requirements for implementation phase should be specified.
- A global plan for SMS development should be established and approved by all participants. Some important recommendations for planning are given in the subchapter A 3.3.

(ii) Management of SMS Implementation

The role of the implementation phase is to implement SMS so that it is ready for use. The following major function, tasks and/or activities should be considered when preparing a metadata management strategy.

- Tools and metadata vehicles specified in the Vision should be developed and tested by all users they were prepared for. Users' manuals and documentation should be developed. Testing should be conducted before making SMS available for the users. Training for all metadata users should be organized.
- Completeness of implementation. The implementation of all SMS subprojects could be a long process. Depending on the links between them, some subprojects can be implemented in parallel and some projects should be completed sequentially. A very important role of the SMS management is continuous monitoring of the implementation from the cost viewpoint.
- Coherent technical implementation. To implement SMS as a technically coherent project should be recommended. It will allow to settle standard links between metadata objects and processes, to develop standard metadata tools for searching, retrieval, exporting and downloading metadata and to harmonize, technical

administration will be easier. Standard operations for administration of diverse metadata can be easily ensured.

- An agreed set of definitions and terminology should be developed. Consideration of national and international terminology standards is of high importance.
- Corporate metadata repository (CMR). A crucial task in the implementation phase is to set up a CMR. This is the physical implementation of the metadata model defined in the Vision and it is likely to be used by all SO projects. The concept of the CMR should be developed, although there could be a number of physical repositories. To develop an appropriate CMR architecture is a demanding task and there is no blueprint for such an exercise. However, many national good practices exist¹² that may be a useful guide.
- Physical loading of metadata into the CMR. Metadata owners should accomplish those activities. This is a resource-consuming task and the impact on subject matter staff should be recognised. For many, capturing metadata is a tedious extra task that brings them no perceived benefit. A characteristic of the 'system' therefore, is that as much metadata as possible is captured automatically, as a result of a computer process or as a result of a required business process undertaken by a person. [Being realistic however, it is inevitable that some metadata will have to be re-entered by humans]. Thorough management and planning of those activities is imperative.
- Existing processes using statistical metadata should be reengineered.
- Outsourcing possibilities for the SMS implementation should be considered.
- Detailed and coordinated plans for all stages of SMS implementation should be prepared and approved by all partners at the beginning of the implementation phase. The basic framework of the SMS plan is defined in the Vision.

(iii) Management of SMS maintenance

The role of the SMS maintenance phase is to ensure that all metadata stored in the CMR is up-to-date for ongoing use. To keep metadata up-to-date is the requirement of primary importance for all metadata users. The following recommendations should be taken into the consideration when preparing a strategy for the management of this phase:

- The major functions to be considered by the SMS management are those relating to the Administration of metadata content.
- Ensure timeliness and coherence of maintenance activities.
- The metadata management should oversee the definition and maintenance of all metadata stored in CMR, although other SO units will contribute to its ongoing enhancement.

¹² The United Kingdom and the United States may provide examples of good practice in developing and maintaining corporate metadata repositories. Recommendations of any other countries that may provide a good example of this would be appreciated.

- SMS management is responsible for definition of policies, procedures and protocols around the CMR maintenance. A 'registration authority' manages all metadata entities in CMR. The major partners for the SMS management are the "owners" of metadata. The owners are, according to the concepts of registration of metadata object specified in the Vision, authorized for keeping up-to-date the metadata that they are responsible for.
- The concept of registration of metadata objects, ownership of metadata, what is the 'standard' for a particular classification or data item, what are the permitted variations from the 'standard' etc should be all clearly defined, agreed and used.
- Rules and guidelines should be developed for the maintenance of each metadata entity in the CMR and responsible metadata owner. It could be recommended that the rules and guidelines will be approved by the senior management and become official documents of the SO.
- Preparation of rules and guidelines requires joint work with owners. Methodologists are also the important partners in this process.
- Management should not only delimitate an organizational background for the maintenance activities, but also should also assure maintenance of metadata history and update links between metadata in the CMR.
- Ensure that all maintenance functions, performed by metadata administrators and diverse metadata owners, use a coherent/standard set of metadata tools and vehicles. Such vehicles should be available especially for the following maintenance functions: search and retrieval, inserting and deleting of metadata objects and related parameters, changes and corrections, presentations and exports, metadata editing and consistency controls, checking and updating of metadata links, maintenance of metadata history.
- Planning is an important instrument for managing the maintenance phase. Everyone participating in the maintenance processes should approve a detailed plan of maintenance activities, which meets required timelines. Such a plan is an indispensable instrument for management to fulfill a smooth and coherent monitoring of the phase of metadata maintenance.
- Training of metadata owners in the Rules and Guidelines prepared for maintenance activities.

(iv) Management of SMS use

The role of this phase is to ensure efficient use of metadata and metadata tools by all users specified in the Vision. Production of official statistics and other internal users in statistical organization together with all groups of external users of metadata specified in the chapter A 2.2. belong to potential metadata users. A great effort should be made by management to monitor and coordinate activities and processes dealing with metadata usage by diverse users. The metadata strategy in this phase should encompass especially the following functions:

- Prepare, maintain and coordinate detailed plans of metadata use by all metadata users. To ensure requested metadata quality within required deadlines. The coordination of plans developed for individual users is a major goal of the management.

- Statistical production process. The units responsible for statistical production are accountable for the preparation and maintenance of plans related to the activities dealing with the production process. In this case, the SMS management should ensure that all activities dealing with the use of statistical metadata and metadata tools are well planned and defined.
- Oversee the availability of metadata and metadata tools. It is important to ensure the links between the metadata maintenance and the metadata use. Metadata users should be sufficiently informed about all changes in the metadata contents.
- Organize a permanent feedback from users about metadata quality and the availability and efficiency of metadata tools. Feedback operations could be integrated in the regular activities of the metadata use. Specially organized surveys on users' satisfactions are useful, but not always fully satisfactory source of information.
- SMS management (in close cooperation with the SMS technical administrator) should be aware of the software and technological environment related to the use of metadata and metadata tools. As it was mentioned earlier, the metadata and metadata tools should be platform independent. However, it could be useful to maintain information about changes in the users' software environment.
- Statistical websites are an integral part of an SMS implementation and the use of metadata. Furthermore, they are a regular part of the dissemination strategy of SOs. The structure and quality of metadata presented on the website are important tool for the satisfaction of the metadata users. The needs for statistical metadata on websites varies according to the needs of the individual users groups'. It is therefore very important to monitor the use of statistical metadata on websites in order to keep track of users' satisfaction and evolution in their needs.

(v) Management of SMS evaluation

The goal of the evaluation phase is to determine the efficiency of existing SMS functions and make proposals for improvement or further development of SMS. There are clear links to the knowledge and experiences accumulated in the earlier phase of the SMS life cycle, namely in the phase of the SMS use. Preparing proposals for further SMS development, the SMS evaluation phase makes a loop between the use and design phase of the SMS.

The management strategy of the SMS evaluation phase should follow especially the following procedures and tasks:

- Specify major targets of SMS evaluation and, based on the targets, to prepare a plan of evaluation activities and procedures. It should be clear which functions and aspects of the SMS are to be evaluated.
- Evaluation of the users' satisfaction should be a permanent part of the SMS life cycle. The most important object of evaluation will certainly be the external user. It should be ensured however, that the satisfaction of other users' groups would also be evaluated.

- Other important aspects for evaluation are cost efficiency, implementation of standards, organization of work, maintenance procedures and technological implementation.
- In principle, there could be three major forms of evaluation: (i) regular long-term evaluations (e.g. at 3 year intervals) that examine overall effectiveness of SMS functionality; (ii) regular short-term evaluations (e.g. annually) that primarily assess user satisfaction; and (iii) ad hoc evaluations as deemed necessary.
- Benchmarks should be established for all defined targets and benchmarking parameters should be defined. Evaluation methods should be specified and agreed. For some cases an efficient benchmarking method is to compare experiences and plans with those of a similar organization. International cooperation could be highly efficient in this respect.
- Appoint evaluators for planned evaluation activities. For evaluating user satisfaction, a team of evaluators should include both staff from the SO and metadata users. For evaluation of the project's efficiency and the overall technological solution, it may be useful to hire external evaluators as they provide an independent view.
- Integrate information on the user feedback collected in the phase of the SMS use.
- Organize a preparation of specific surveys on users' feedback.
- Report to the senior management of the SO on the evaluation outcomes and, based on the conclusions made by the senior management, to organize steps for improvement of and/or further development for the SMS

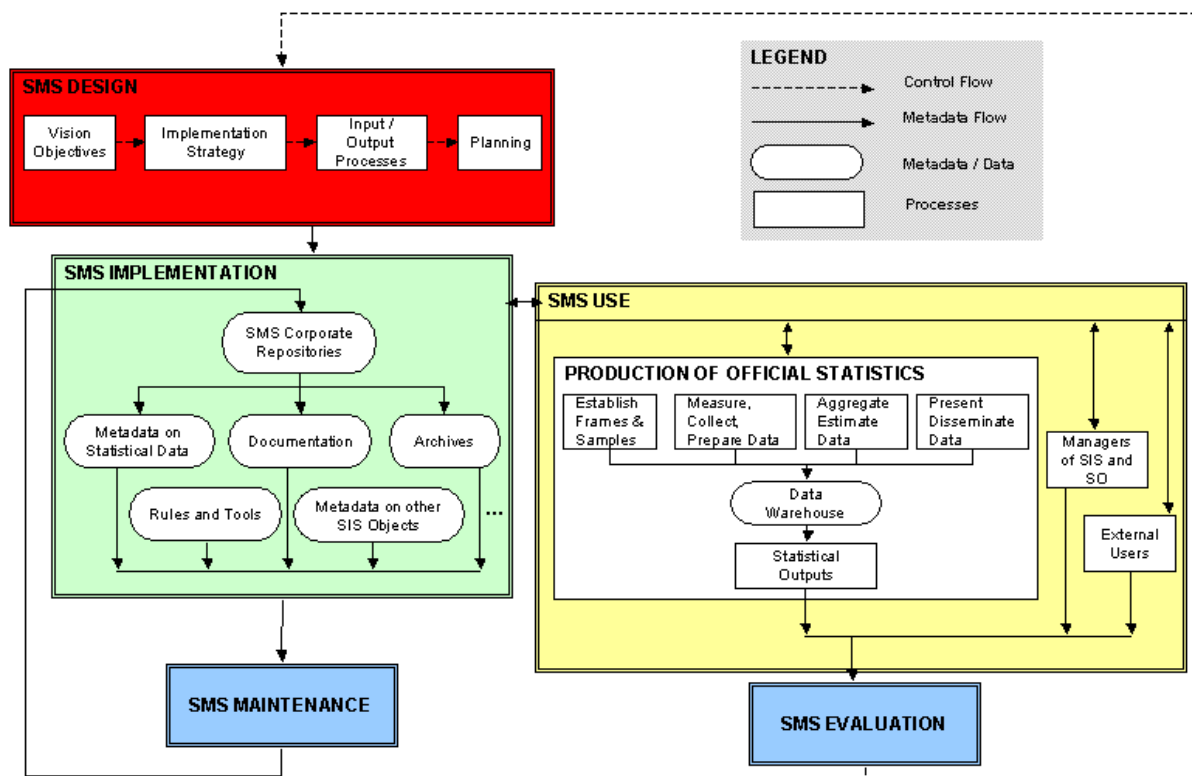


Figure 6: Model for integrated management of SMS

A.4. Core principles for metadata management

This section outlines the core principles and recommendations for effectively managing the design and implementation stages of an SMS project.

1. Make metadata-related work an integral part of business processes across the organization.
2. Describe metadata flow with the statistical and business processes (alongside the data flow and business logic).
3. Ensure that customers are clearly identified for all metadata processes, and that all metadata capturing will create value for stakeholders.
4. Metadata presented to the end-users match the metadata that drove the business process or was created during the business process.
5. Develop SMS as a self-sustainable project, independent of any e-production systems.
6. The SMS is the definitive set of tools, stores and services to support metadata use and further development in SO. If a metadata store, tool or service is not defined by the management to be a part of SMS, then it is not an “approved” metadata facility.
7. The diversity of metadata is recognised and there are different views corresponding to the different uses to which the data is being put. Different users require different levels of detail. Metadata appear in different formats depending on the processes and goals for which they are produced and used.

8. Make metadata active to the greatest extent possible. Active metadata drives other processes and actions will therefore be accurate and up-to-date.
9. Manage metadata with a life-cycle focus (including maintenance and update).
10. Preserve history (old versions) of metadata.
11. Capture metadata at their natural sources, preferably automatically as bi-product of other processes. Minimize errors by entering only once where possible.
12. Exchange metadata and use it for informing both computer based processes and human interpretation. The infrastructure for exchange of data and associated metadata should be based on loosely coupled components, with choice of standard exchange language, such XML.
13. All data and other objects of SMS are well supported by accessible metadata that is of appropriate quality.
14. Ensure that metadata is readily available and useable in the context of client's information need (whether client is internal or external).
15. Single, authoritative source ('registration authority') for each metadata element.
16. Registration process (workflow) associated with each metadata element, so that there is a clear identification of ownership, approval status, date of operation etc.
17. Reuse metadata where possible for statistical integration as well as efficiency reasons (no new metadata elements are created until the designer/architect has determined that no appropriate element exists and this fact has been agreed by the relevant 'standards area').
18. Cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata.
19. Variations from standards are tightly managed/approved, documented and visible.
20. Ensure a systematic training and transfer of know how for all partners involved. Train trainers.

A.5. Corporate Governance Models for Metadata Management

A.5.1. Introduction

It is not sensible to prescribe an ideal model for corporate governance of metadata. This is because every national statistical organization works under different legislation, organizational arrangements, organization culture, business rules, levels of autonomy with respect to central public sector agencies, etc.

Therefore, in this section of the manual, we look at 'good lessons' for governance. Each SO that is implementing a metadata management strategy might evaluate its objectives, strategies, organizational arrangements and plans against this wisdom that is generated from many metadata and other information management projects.

The Eurostat sponsored project, Metanet, concluded in 2003 had one working group looking at Adoption Issues in respect of statistical metadata systems. The third section in this chapter provides some extracts from the report about barriers and organizational issues - both matters are relevant to governance.

The fourth section is a case study of corporate governance at the Australian Bureau of Statistics to give readers a concrete example of the issues associated with governance.

A.5.2. Lessons for Good Corporate Governance of Metadata

What are some of the lessons for corporate governance of data and metadata management that have come from the experiences at national statistical agencies in the implementation of a metadata management strategy?

1. Senior management group, including the Chief Statistician, should be very involved in policy formulation, approval of development projects and monitoring of outcome achievement. It is very helpful when your CEO and other senior executives ask questions about metadata matters.
2. Clearly understood roles and accountability for all organizational units with respect to metadata. The subject matter areas are responsible for the creation, maintenance, re-use, and approval for dissemination of all the data and metadata content for their statistical domain. A 'corporate data management unit' could be accountable to provide client support to SMAs, develop and maintain infrastructure, provide training, etc
3. The organization should develop an information management culture. That is, all staff understand that it is their responsibility to work towards achieving the ideals of statistical integration, comparability of statistics across surveys and time, and to reuse statistical metadata as appropriate. These goals are achieved by adherence to the metadata management principles.
4. Utilise existing governance arrangements to reinforce the metadata messages, that is, do not create new committees. Particular specialist staff eg business and systems analysts, IT architects, statistical standards experts, are more likely than others to come across new opportunities for advancing better metadata integration, so a particular focus is needed on working with these staff.
5. Make sure that your organization has an endorsed metadata strategy, including a global architecture and an implementation plan, and that this strategy is integrated into broader corporate plans and strategies.
6. Either commit yourself to a metadata project – or don't let it happen. Lukewarm enthusiasm is the last thing a metadata project needs.
7. There is often scepticism in the organization against metadata projects. Moreover, metadata projects are usually strategic projects for the organization. If they should be carried out at all, managers on different levels and in different parts of the organization must be committed to the project.
8. Metadata projects are often more abstract, more complex, and more difficult to manage than most other types of projects. These characteristics need to be recognised in project plans, and the importance of communication with the rest of the organization about the project cannot be overstated.
9. Make sure that your organization also learns from failures and successes in other statistical organizations. Benchmarking and international cooperation are always useful.
10. Make systematic use of metadata systems for capturing and organising tacit knowledge of individual persons in order to make it available to the organization as a whole and to external users of statistics.

A.5.3. Barriers to the Introduction and Use of Statistical Metadata Systems

One of the Working Groups in the Metanet Project explored adoption issues with respect to statistical metadata systems. Realisation that there are potential barriers is an important part of the management and governance of such projects. Consideration of appropriate risk mitigation actions is a significant part of project governance. This subsection explores some of the potential barriers to the adoption of metadata solutions - technical, organizational and human - that were identified by a survey conducted as part of the Metanet research.

The Metanet working group included in their survey of national statistical organizations questions seeking to identify in which area each of the potential problems were most important as well as to go into more detail concerning the different aspects of human related issues.

A.5.3.1. Most important challenges to introduction of metadata systems

The respondents were asked to answer the following question: "For each aspect of metadata please indicate what in your view poses the greatest challenge to the introduction or use of statistical metadata systems in your organization". The result of this for all organizations was the following.

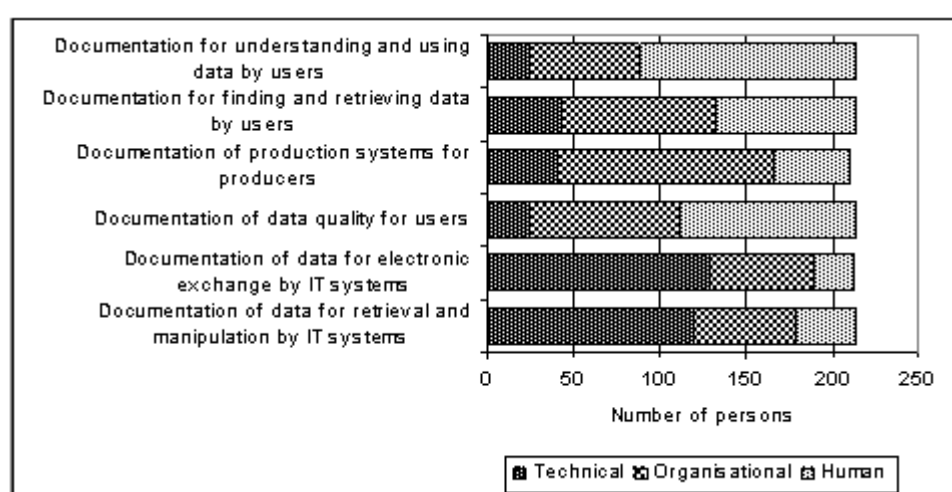


Figure 7 Perceived challenges to introducing or using statistical metadata systems

According to this, the greatest challenges in relation to documentation and retrieval of data are considered to be partly organizational and human. On the other side, the technical challenges are the most important in relation to documentation of data for exchange and retrieval by IT systems.

A.5.3.2. Human issues in relation to adoption of metadata systems

The Metanet working group reported: "The human factor is fundamental to the successful adoption of metadata systems, yet a number of challenges have been identified. At present there might be a substantial gap between some of the more theoretical and abstract contributions on metadata, as presented within the MetaNet project, and what is considered to

be applicable by many practitioners within statistical organizations. Some subject matter specialists tend to dig down into one specific area and not take into account the long time perspective for documentation. Motivation for general metadata solutions might therefore be low. There is a need to acquire input and feedback from subject-matter specialists from different areas regarding metadata/data concepts and methods in order to come up with viable common standards and methods for metadata. However, the viability of metadata solutions presupposes the motivation and commitment of metadata providers. Given that an inability to engage this community constitutes a major concern to many statistical agencies (as identified through initiatives such as AMRADS) it is important that underlying human barriers are fully understood, if they are to be redressed in future.

The respondents in the survey were asked to identify the most significant barriers to the provision of effective metadata within their own organization, as far as human issues are concerned.

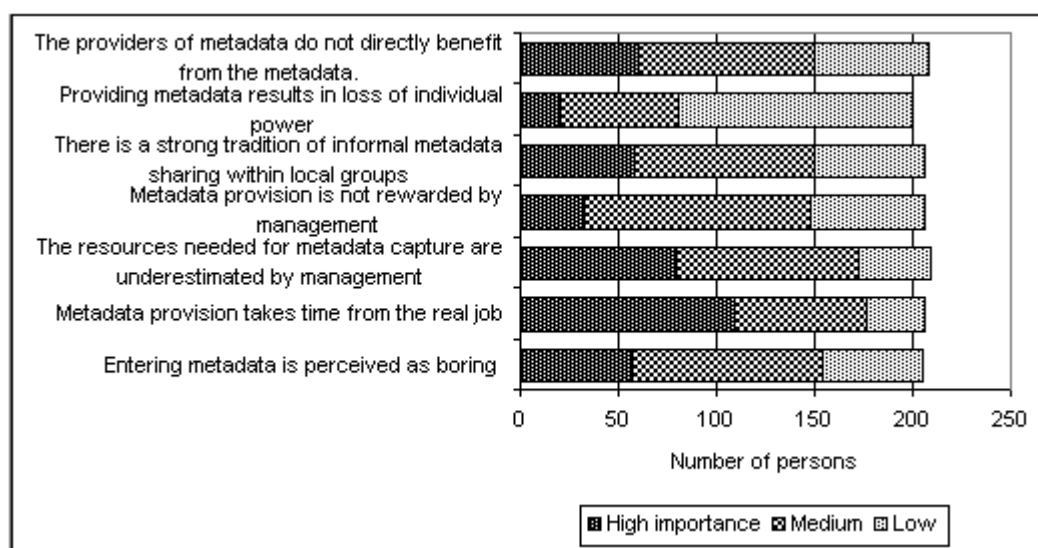


Figure 8 Perceived barriers to the provision of effective metadata

Whilst the result indicates that there is no clear consensus on what the main barriers are, the majority think that loss of individual power as a result of providing metadata, frequently *perceived* as a significant deterrent, is **not** significant. Other possible issues that have often been suggested, namely that metadata provision is boring and does not benefit the provider directly were largely identified as being of only medium importance. Of greater significance was the belief that time spent providing metadata detracts from the real job. This suggests that the importance afforded metadata creation is low and that this activity will inevitably suffer at the expense of traditional work aspects. Moreover, over 25% think that management underestimates resources needed for metadata capture. Thus, if the barriers to effective metadata provision are to be overcome, the status of the activity must be elevated. This demands not only the education and active involvement of would-be providers, but also increased management awareness and support."

A.5.3.3. Organizational issues

The MetaNet report says: "It is often emphasized that it is necessary to ensure the

commitment of top management in order to succeed in putting in place metadata solutions. In addition, organizational issues might be critical when planning and implementing metadata strategies and applications- even if these are not often put high on the agenda for meetings discussing metadata. Thus one should discuss questions such as what type of staff should be involved, and to what degree there should be a central unit, and what tasks should such a unit could cover. The survey tried to address some of these issues without trying to go into detail.

However, there are several more fundamental issues that need reflection in order to achieve this commitment and implement a proper organization:

- First of all it is necessary to reach a common understanding within the organization of what metadata is and what the functions are. The specification of these functions will have several implications on how projects should be designed and organised and how running tasks should be taken care of.
- Organization of tasks related to metadata should be based on a strategy for the information management of the organization. One reason for the failure of specific metadata projects can be that they are not anchored in a more global view of the information architecture.
- In order to sell the need for basic changes in technology or organization to improve data/metadata management it is necessary to present the benefits and the proposed solutions in an understandable way - possibly based on practical experiences acquired in other organizations. Management is not prone to take decisions involving risk to continuity of production. Once again, management might not be a barrier, but the limiting factor might be the experts' ability to come up with convincing and practical proposals related to metadata. Proposals that reach too far and have a too long time perspective will have difficulties as management normally will ask for quick results within a short time frame.

A.5.3.3.1. The degree of central coordination

One might assume that a central coordinating unit at least is a signal that metadata/documentation is important and that there is a relatively high level of horizontal coordination. According to the overview presented in table 6 only 3 NSIs reported to have a strong central coordinating unit, whereas a majority of the NSIs had a coordinating unit with limited tasks. The data archives apparently have a stronger central organization taking care of this topic. Even if the data archives often are rather small organizations having more limited tasks compared to the NSIs, this supports the impression that the data archives have taken documentation seriously for a long time, both for internal and external purposes.

	All	NSIs	Data archives	Other
A. Strong central coordinating unit -	9	3	5	1
B. Coordinating unit with limited tasks - decentralised organization	20	15	2	3
C. No coordinating unit - distributed organization	5	3	0	2
D. Other/no answer	4	1	1	2

Total	38	22	8	8
-------	----	----	---	---

A.5.3.3.2. Tasks of a coordinating unit

The contact persons were asked to indicate the tasks allocated to any coordinating unit. It is interesting to note that for a majority of both the NSIs and the other organizations having some central coordination, an important task for this unit was to develop common systems and solutions. These coordinating units also have an important role to play in developing common terminology and standards and to ensure general coordination and information in this field of work. Supervision and training apparently is not an important task of many units.

	All	NSIs	Other
General coordination and information	20	10	10
Developing common terminology and standards	22	13	9
Developing common systems and solutions	23	14	9
Supervision and training	13	8	5
Other/not specified	2	2	0

A.5.3.3.3. The involvement of different types of specialists

IT specialists appear in most organizations to have central positions in relation to planning and development of metadata systems and solutions, which is not surprising due to the traditional importance of metadata in computer based systems. Also specialists in statistical methodology have in most organizations a central role in this area. On the other side it is perhaps somewhat worrying that management and subject matter specialists are to a lesser degree involved.

Expertise/Specialist	NSIs		
	Central involvement	Partly involved	Not involved/not relevant
IT specialists	15	5	1
Management	6	11	4
Statistical methodology specialists	12	8	1
Subject matter specialists	14	14	3

A.5.3.3.4. Cooperation with other organizations

Metadata is a field of work where one should expect a large degree of cooperation with other organizations in order ensure harmonisation and exchange of best practise. The survey confirms this. A large majority of the NSIs foresee cooperation with statistical organizations in other countries in this field, whereas many also see a possibility for cooperation with consultants and vendors of IT systems. Cooperation with international organizations is also foreseen. Many NSIs foresee cooperation with data archives/documentation centres.

	NSIs		
	Absolutely	Possibly	Not planned
IT system vendors/consultants	2	13	4
Other stat. org. in own country	6	5	8
Other stat. org. in other countries	14	4	1
International organizations	11	7	0
Data archives/documentation centres	5	9	5

Many large statistical organizations are searching for efficient models for handling data and metadata in an integrated way throughout the production process. Decentralisation of technology, in some cases also leading to loss of central documentation of files and processes, has in many organizations made it even more important to find ways and means for coordinating documentation across the organization.

Thus it is useful to look more into the experiences of different organizational models in order to achieve common and efficient metadata solutions. "

A.5.4. Case Study - Australian Bureau of Statistics (ABS)

What is the situation with corporate governance at the ABS?

The ABS is headed by the Australian Statistician - a statutory office. Administratively, the ABS is included in the Treasury portfolio, along with the Taxation Office. Although the Australian Statistician might occasionally work with the Treasurer (a very senior Minister in the Government), it is more usual for the Statistician to deal with a junior portfolio minister when interaction with the government is needed.

The Statistical Operations of the ABS are divided into two groups: the Economic Statistics Group and the Population Statistics Group. Each group is headed by a Deputy Australian Statistician. The staff responsible for the Technology Services, Methodology, Information Management, and Corporate Services Divisions, report directly to the Australian Statistician (known as First Assistant Statisticians).

ABS corporate governance arrangements ensure transparency in decision making and operation, and accountability to stakeholders by promoting strong leadership, sound management and effective planning and review.

An important element of the ABS governance arrangement is the Australian Statistics Advisory Council, established by the *Australian Bureau of Statistics Act 1975* to assist the ABS to fulfil its role. The Council is the key advisory body to the ABS and provides valuable input to the directions and priorities of the ABS work program and reports annually to Parliament. It is comprised of Federal and State government representatives, along with people from industry, academia and welfare constituencies.

An important feature of ABS corporate governance is the role played by senior management committees, which are active in identification of ABS priorities, ensuring appropriate planning and implementation to address those priorities, and effective monitoring of ABS activities. Those committees relevant to data and metadata management are:

1. ABS Division Heads Committee which includes the Australian Statistician and involves the heads of Divisions (Economic, Population and Social, Methodology, Information Management, Corporate Services, and Technology Services (the CIO)) ie all the 'direct reports' to the Statistician. This group could be considered as the 'Board' and they usually meet weekly. They review and approve all policies related to data and metadata management, approve specific projects that related to metadata infrastructure, and approve all funding proposals.
2. Information Resource Management Committee - the same group as in 1 above, minus the Statistician and including heads of the Technology Services Division branches, namely as Technology Applications (development), Technology Infrastructure (all hardware, software, communications services) and Technology Research (future tools and techniques). This committee focuses on the technology directions and proposals for the ABS, including data and metadata management. This group approved the detailed metadata management strategy, principles, etc.
3. Standing Committees. The two major subject matter groups in ABS - Economic Statistics, and Population and Social Statistics - have standing committees to review, discuss and approve subject matter projects, including the development of metadata content and the standardisation of metadata. These committees provide the strong articulation of the business drivers for data and metadata management work. They comprise the senior executives of each subject matter group, along with the senior executives of support divisions eg technology and methodology.

In addition to the senior management committees, there are a number of other important parts of the governance arrangements. They are:

1. Project Boards. Each major project in the ABS, whether a new infrastructure development eg our Input Data Warehouse, or a new survey, has suitable governance arrangement that would probably involve a Project Board. A Project Board is chaired by the owner of the project, ie the person who is ultimately responsible for achieving the outcomes and objectives of the project. They are assisted by senior people from relevant areas that are able to help the project deal with the tasks, issues and risks that arise. In terms of metadata, one of the roles of the Board is to ensure that corporate policies are followed and that the project

solution follows metadata management principles.

2. Architecture Panels. An architecture panel is usually convened for each project that has a significant IT component, with the view to determining the best technical solution for the project taking into account available development toolsets and the impact on the IT infrastructure eg storage capacity required, server and network load. One function of the architecture panel is to ensure that the solution proposed, in terms of metadata, makes appropriate use of corporate metadata facilities, and if any new metadata facility has to be developed that its potential use for other projects is assessed. Often, the Director of Data Management Section attends architecture panels to make this assessment.
3. Line Management. Responsibility for data and metadata management has been made very clear at the ABS. Corporate units, like the Data Management Section, are responsible for developing policy and practices, as well as specifying, developing and maintaining the corporate data and metadata management systems infrastructure, and providing training and client support in data and metadata management. Subject matter areas are clearly responsible for the statistical data and metadata content - they 'own' the data and metadata and are responsible for the adequate documentation, confidentiality and quality of that data and metadata. Each of our major statistical groups - economic and population/social - has a Standards area which is responsible for defining and maintaining standard classifications and data element definitions. Often they play a role in ensuring compliance with those standards by being part of the approval workflow, for example when collection forms need to be approved before use.

Glossary of terms and abbreviations

AMRADS

Accompanying Measure to Research and Development in Official Statistics (AMRADS) website at <http://amrads.jrc.cec.eu.int/>

COSMOS

Cluster of Systems of Metadata for Official Statistics (COSMOS) website at <http://www.epros.ed.ac.uk/cosmos/>.

Corporate Metadata Repository (CMR)

A database system that stores metadata records for an organization or group of organizations.

Designer

People responsible for the technical design of a statistical metadata system.

GESMES

GESMES/TS (formerly called GESMES/CB) is the message used by the European Central Bank to exchange statistical data and metadata with its partners in the European System of Central Banks (ESCB) and other organizations world-wide. For more information see the website at <http://www.ecb.int/stats/services/gesmes/html/index.en.html>

Metadata

A term used to describe data about data. This may include any information that is stored about the nature of data such as format, source, language, creation date, etc. Metadata may also be referred to as metainformation.

MetaNet

MetaNet was created as a network of excellence to harmonise and synthesise statistical metadata developments. It started in November 2000 and finished at the end of July 2003. See their website at <http://www.epros.ed.ac.uk/metanet/index.html>.

METAWARE

Statistical Metadata Support for Data Warehouses. See their website at <http://europa.eu.int/en/comm/eurostat/research/retd/metaware.html>

SDMX

Statistical Data and Metadata Exchange website at <http://www.sdmx.org/>.

Statistical Metadata System (SMS)

The processes and resources used to manage metadata within a Statistical Information System.

Statistical Information System (SIS)

The processes and resources used to produce statistical information.

Senior Management

The highest level of management in an organization, responsible for ensuring the organization meets its goals efficiently and effectively. May also be referred to as 'Executive' or 'Top' management.

Statistical Organization (SO)

An organization that is responsible for the collection, processing and dissemination of official statistics.

XML

Extensible Markup Language – a markup language primarily used to facilitate the sharing of data across different systems, either within or between organizations.

XBRL

Extensible Business Reporting Language.

References

- ABS, *Strategy for End-to-End Management of ABS Metadata*, (2003)
- AMRADS, EU project, *Final Report of the Working Group on Metadata*, (2003)
- Booleman, M, Statistics Netherlands, *The Dutch Metadata Model*, (2004)
- Czech Statistical Office, *Conception of the Statistical Metadata System in the Czech Statistical Office*, (2004)
- Czech Statistical Office, *Global Architecture of the Statistical Metainformation System in the Czech Statistical Office*, (2005)
- Eurostat, *Eurostat Metadata: Architecture and Strategy*, (2003)
- Eurostat, *SDMX Common Metadata Vocabulary*, (2005)
- Hustoft, A, Statistics Norway, *Development of a variables Documentation System in Statistics Norway*, (2004)
- Lindblom, H, Statistics Sweden, *The Metadata System at Statistics Sweden in an International Perspective*, (2004)
- Meliskova, J, AMRADS project, *External Users of Metadata: a Challenge for National Statistical Offices*, (2003)
- OECD, *Data and Metadata Reporting and Presentation Handbook*, (2005)
- Sundgren, B, Statistics Sweden, *Advice Concerning the Strategic Decisions that have to be made by a Statistical Office when Developing and Implementing a Metadata Structure*, (2002)
- Sundgren, B, Statistics Sweden, *Statistical System: some fundamentals*, (2004)
- Sundgren, B, Statistics Sweden, *Designing and Managing Infrastructures in Statistical Organization*, (2004)
- United States Bureau of Labor Statistics, *Creation and Use of Metadata by Survey Methodologists*, (2002)
- Zeila, K, Statistics Latvia, *Metadata Driven Integrated Statistical Data Management System*, (2004)

**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 (i): Metadata in a Corporate Context

THE VALUE ADDED OF A NATIONAL STATISTICAL INSTITUTE

Invited Paper

Submitted by Statistics Netherlands¹

I. CONTENT

1. The paper focuses on the necessity of good metadata and metadata models from the perspective of a process of reducing primary observations and the increasing use of external registrations. Due to this process the organisation of Statistics Netherlands changed a few years ago from a stove pipe organisation into a process oriented organisation. Increasingly, the same data source will be used by many statisticians to produce their statistics. Therefore, metadata are the corner stone of the production process.
2. Using data, it should be clear what is the content (conceptual metadata), how it is produced (process metadata) and what is the quality (quality metadata). During the production process it should be clear how it should be produced (prescribing metadata) and what has been realised so far (realization metadata).
3. The paper introduces the new business architecture of Statistics Netherlands. The role and position of metadata in the architecture will be discussed. Also organizational issues and problems related to managing conceptual metadata will be presented.
4. Statistics Netherlands is also looking for experiences within other national statistical institutes related to prescribing and realization process metadata.

II. INTRODUCTION

5. In 2000 Statistics Netherlands changed its organizational structure from a stove pipe organisation to a process oriented organisation. This organizational change was believed to better support the mission of Statistics Netherlands, which is:

“To compile and publish undisputed and coherent up-to-date statistical information that is relevant for practice, policy and research”.

¹ Prepared by Marleen Verbruggen (mrvbn@cbs.nl) and Max Booleman (mbln@cbs.nl)

6. The main reasons for this step towards a process oriented organisation were the increasing demands for statistical information and the rapid developments in the area of information and communications technology.

7. Statistics Netherlands is in the middle of important changes. The increased complexity of society and the rapid changes it is going through imply an increasing demand for reliable statistical information. The focus is shifting towards thematically presented information providing more insight how developments and sectors are related. Statistics Netherlands wants to speed up its flexible response to the changing needs for information, which requires combining data from various sources.

8. The data collection is shifting from primary to secondary collection. Developments in ICT have made most administrative data and registrations available in electronic form. This enables public and private sector organisations to compile and publish statistical information based on these administrative records. As a consequence, a National Statistical Institute (NSI) is no longer the only organisation that can produce and publish statistics. Moreover, a NSI will also increasingly be forced to use these administrative records, because this is less costly than collecting the information by surveying. Besides, there is a stringent policy to reduce the administrative burden on respondents. In the near future, therefore the value added of a NSI will not be the dissemination of statistics based on a single source, because each external register owner will be able to do so much faster and in a better way.

The value added of a National Statistical Institute will be the knowledge and skills necessary to combine sources with each other.
--

9. Next to this, due to privacy regulations, a NSI is sometimes the only body that is permitted to do so. Furthermore a NSI is maybe the only one who is able to present a coherent picture of the whole society.

10. In order to be able to combine many administrative sources and additional survey information, a NSI needs the appropriate methodological skills, an appropriate ICT environment and... up-to-date meta information!

11. Currently Statistics Netherlands has a few ICT systems that fit nicely in the process oriented organisation. For example, StatLine is a very sophisticated database that contains all the statistical information published by Statistics Netherlands and that can be approached by the website. There exist however also a considerable number of small ICT systems that fitted nicely in the stove pipe organisation, but are becoming a mill stone in the present organisation because they are very heterogeneous and therefore very costly to maintain. Recently, Statistics Netherlands decided to develop a business architecture with the aims to reduce the number of ICT systems, to make the remaining systems more generic and facilitate the communication between different systems. This should also have a positive effect on the costs of maintaining the technical infrastructure.

12. In chapter 2 the basic principles of the business architecture will be explained. Meta information in all its forms (conceptual metadata, process metadata and quality metadata) plays an important role in this architecture.

13. In chapter 3 we will focus on the organizational consequences of this business architecture and especially the role of metadata in this respect. Key issues are co-ordination and the attribution of tasks and roles with regard to managing metadata. This chapter will be concluded with a brief overview of present experiences with managing metadata.

14. The paper concludes with a few final remarks in chapter 4.

III. THE BUSINESS ARCHITECTURE AND THE ROLE OF METADATA

15. As introduction a few words on the concept of a business architecture. A business architecture is not a well defined concept. There are numerous private organisations that have developed very different approaches to describe a business architecture. But if one tries to describe what a business architecture does it would be more or less the following:

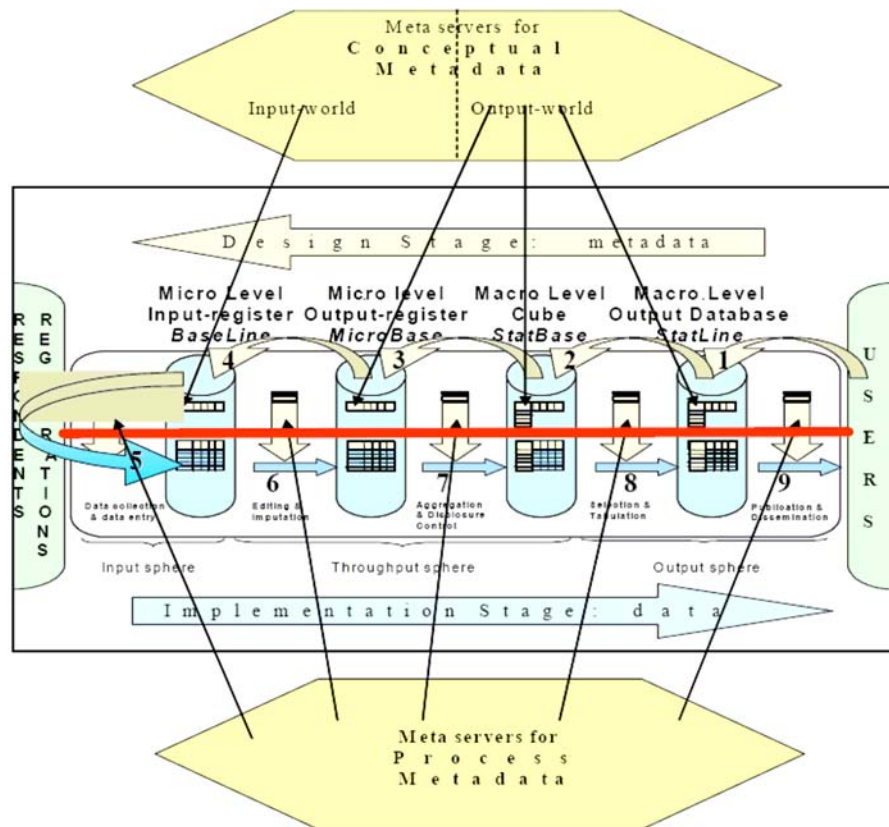
A business architecture tries to translate the strategic goals of an organisation into an optimal organizational structure including the supporting ICT systems.

16. The model that is used by statistics Netherlands is based on the Integrated Architecture Framework (IAF) and should provide the office with a decision making framework. The current view on the future set-up of the organisation shows the following principles for the architecture:

- a. Organisation: a CBS-wide, generic set-up of processes and ICT support. The ICT support must therefore be independent from the organizational structure. Research shows that primary data collection among persons and enterprises can be incorporated in one universal, common CBS model. The ICT solution has to support this.
- b. Product: support of all statistical products and a fast and flexible response to changing demands for new products. This means that ICT processes must support all statistical products, where the processes must be adjustable (relatively) fast in order to meet changes in demand.
- c. Process: preventing duplication of processes (no stove-pipes), implementing transparency by adding process data (meta!) and leaving room for exceptions.
- d. Environment: addressing the increasing diversity of suppliers by supporting a variety of channels and formats, scalability to process large quantities of input data and to add new suppliers in a simple manner, scalability for processing and production of data and preparing these data for direct integration with external sources (for instance direct integration with the basic registrations).

17. According to the current views on the business architecture, the future statistical process can be depicted as shown in figure 1. The databases in the figure, with the exception of StatLine, are “logical” databases; they do not (yet) correspond to actually existing physical databases. The figure is also simplified in the sense that it does not show the possible re-use of data and other dataflow patterns such as a feedback loops for provisional data and data integration in national accounts. However, the figure perfectly illustrates the key role played by metadata, by showing how at the design stage, ideally, user demands are translated to demand for input data (1 to 5), thereby also determining the transformations needed to get from inputs to outputs (5 to 9). For the transformations themselves, design has to be distinguished from realization of course. Although realization is meant to concur with the design, especially if the process is metadata driven, the distinction is particularly important when analysing quality aspects.

Figure 1 – Future statistical process



18. According to this model of the future statistical process, it is the metadata that has to be specified if one wants to know how inputs are processed in order to produce the statistics that are disseminated. Thus metadata (or 'meta') is the key to quality control. More specifically, according to the business architecture meta has the following roles:

- a. Meta describes / models the process of statistical production
- b. Meta is specified in the design phase
- c. Meta is used in the statistical production process
- d. Meta gives a univocal specification of data, process and quality (one interpretation)
- e. Meta gives a uniform specification of data, process and quality (one language)
- f. Meta is used actively in the process of statistical production
- g. Meta defines what and how and with which quality statistics are produced
- h. Meta fixes the process of data transformation in a transparent way.
- i. There are no data without meta: meta precedes the data
- j. The design of meta is the core business of a NSI.

19. In short, meta forms the corner stone in the future statistical process and the business architecture prescribes therefore the following guiding principles:

- a. A strict distinction between the implementation (production) and design (development).
- b. No regular production without meta (workflow, prescribing rules); processes are metadata driven.

In addition:

- c. There is a strict distinction between data and metadata (conceptual metadata, process metadata and quality metadata).
- d. There is a strict distinction between metadata and meta about the metadata. This last meta defines the concepts used in the metadata and can be seen as a first step towards the building of a framework of metadata.
- e. The metadata should be easily accessible and standardised (object types, variables, classifications, time dimensions, quality characteristics, process terminology).
- f. The design and maintenance of prescribing rules are important statistical activities.

20. In order to realise this broad aims and principles, SN aims for the following specific results in the next few years:

- a. The development and approval of a standardised framework of concepts of metadata, including the appropriate definitions, (the meta of metadata). This will facilitate the communication between and among users (statisticians) and software builders. A lot of preparatory work has already been done and the concept framework is almost already for approval.
- b. The design of a standardised model for conceptual metadata, based on the StatLine model (the so-called Cristal model). All statistical processes should be able to use the model and the model will describe the metadata in a univocal, uniform and systematical way.
- c. A list of demands for the ICT tools and systems that have to be bought or built to support the standardised model for conceptual metadata.
- d. Proposals to hook up several other databases to the metadata server of StatLine. This will facilitate the use of variables and classifications that are currently stored in the StatLine metadata server by other statistical processes (input and throughput).
- e. A study on the requirements on process metadata and quality metadata based on the demands of the most representative statistical processes. The study should also assess the applicability of the so called “view model” in this regard².
- f. An up-to-date and approved organizational structure for managing metadata, including a work plan.
- g. Assuring support by the most important stakeholders concerning the feasibility and advantages of a systematic description of metadata, by conducting pilots and building proofs-of-concept.

21. Most activities are strongly research oriented or ICT based, while the scope of this paper is mainly the organizational aspects of metadata. There is one important exception: managing metadata is mainly an organizational issue. The remainder of this paper will therefore focus on the concepts, plans en experiences with managing metadata.

IV. ORGANIZATIONAL ASPECTS OF METADATA

A. Introduction

22. Based on the above mentioned principles, metadata is the corner stone of the production process of statistical output. It is the most important resource of a NSI. Using data, it should be clear what is the content (conceptual metadata), how it is produced (process metadata) and what is the quality (quality metadata). During the production process it should be clear within the NSI how the data should be produced (prescribing metadata) and what has been realised so far (realization metadata).

	Prescribing metadata	Realization metadata
Conceptual metadata: <ul style="list-style-type: none">• endogenous• exogenous	X X	In principle not needed ³
Process metadata	X	X
Quality metadata	X	X

² A view is a selection of information held by any , or any combination, of the logical databases of figure 1. This translates into queries into one or more physical databases. All data processing can be seen as starting with data from one or more databases, selected by applying a view. Transformation rules are applied to this data, the result is entered to one or more databases. The result can also be represented by means of a view. Specifying views and transformations rules are an essential part of process metadata as well.

³ Only if prescribing process metadata is organising the process in such a way that the process results into the prescribing conceptual metadata.

23. In itself, managing realization metadata is not needed. These metadata should be generated automatically during the production process. What kind of realization metadata is needed to monitor, describe and present the quality of the process and the product is predefined.
24. Prescribing process and quality metadata is the result of user and process-owner demands.
25. With regard to prescribing conceptual metadata a distinction can be made between predefined external metadata (exogenous) and endogenous metadata, which can be adjusted given user wishes. Most of the time exogenous metadata equals metadata of input data. This kind of metadata should be stored and linked to (endogenous) output metadata.
26. Managing endogenous conceptual metadata is a task of a NSI and is strongly related to the statistical system. The wish to be coherent with other related concepts is sometimes in contradiction with individual internal and external user wishes. Strong internal co-ordination is needed. Statistics Netherlands distinguishes several roles to manage prescribing conceptual metadata.

B. The roles

27. For managing conceptual metadata the following roles are distinguished:
 - a. **The Theme administrator.** Theme management concerns the substantive responsibility for the conceptual metadata of a thematical area (i.e. demography, employment, national accounts) and the consistency in particular. The Theme Administrator should in detail be informed on internal and external agreements concerning the conceptual metadata belonging to the theme. The Theme Administrator is responsible for compliance with national and international agreements. For a theme different Interested Parties of these metadata can be identified (see below). The Theme Administrator is also responsible for the identification of the list of Interested Parties.
 - b. **The Interested Parties.** Sub roles: Main User and remaining Interested Parties. The Interested Parties are not necessary users of the metadata; an Interested Party could use related metadata. The Main User is designated as the most important Interested Party within CBS.
 - c. **The Supplier.** Sub roles: Supplier of endogenous metadata and Supplier of metadata of secondary data sources. For all clarity: these are all internal roles, also if it concerns externally defined metadata.
 - d. **The Problem spotter.** It occurs frequently that someone observes that there is a problem related with the conceptual metadata in a specific area. A Problem Spotter reports to the Coordinator (see below).
 - e. **The Authority.** The authority is the formal owner of metadata. The Coordinator acts as delegated authority in case of unanimous recommendations of Theme Administrators.
 - f. **The Communicator.** The communicator organises the distribution and accessibility of concepts and approved metadata at Statistics Netherlands.
 - g. **The Editor of output metadata.** Disseminated metadata must meet certain standards. The Editor is responsible for this.
28. In fact with these roles management of conceptual metadata has been totally covered. Beside these roles two other roles can be distinguished to facilitate the process of co-ordination:
 - a. **The Coordinator** of metadata. The Coordinator starts and coordinates the process that leads to authorised metadata. Furthermore, the Coordinator monitors the general requirements to which metadata definitions and - explanations must comply. Finally, the Coordinator is the delegated authority in case of unanimous accepted proposals.
 - b. **The Process Owner.** The production or modification of conceptual metadata from the beginning until the end can be considered as a specific process. This process has an owner, who is responsible for the progress made and the quality of the process.

C. The phases of the process

29. The process to reach authorised conceptual metadata contains different phases:
- a. **Initiation phase:** the process starts with an action aimed at approval of the metadata with a:
 - proposal by a Supplier for addition of metadata or
 - proposal by a Supplier for removal of applying metadata or
 - proposal by a Supplier for modification of applying metadata or
 - b. Signal of a problem by a Problem Spotter.
 - c. The proposal or signal will be addressed to the Coordinator, who addresses the appropriate theme and presents it to its Theme Administrator..
 - d. **Production phase:** If the Theme Administrator accepts the case, he will draw up a recommendation. In this core process the Interested Parties are involved.
 - e. **Authorisation phase:** Authorisation is no commonplace if the recommendation of the Theme Administrator is controversial. An escalation procedure is part of this process.
 - f. **Setting-up phase:** The eventual decision must be communicated within the organization, including starting date and possible accompanying measures.
30. The Coordinator has not been indicated in the process diagram. He accompanies and monitors the whole process. During the process the Coordinator ensures the communication between the Theme Administrators, Interested Parties, Suppliers, Problem Spotter and the Authority.
31. The process assumes the existence of a list of themes (including their content description) each with an overview of the Interested Parties, among which the main users. The theme list is a product of the **Theme Council**. The Theme Council exists of all Theme Administrators, a representative from the methodology department and is chaired by the Coordinator. The Theme Council meets at least once a year and establishes the list of themes (including their contents description) and actualises these. The lists of Interested Parties and Main Users by theme are maintained by the Theme Administrators. Formally there is no special role for the Theme Council in the authorisation process, but it is well conceivable that for difficult, several themes touching problems, the responsible Theme Administrator uses the Theme Council for detailed discussions. It is also possible that the Authority consults the Theme Council. Furthermore, the Theme Council decides on the annual program for the production of conceptual metadata. Due to limited resources a Theme Administrator based on this annual program could postpone or refuse new proposals by Suppliers or signal of Problem Spotters. Finally, the Theme Council is responsible for the preparation of an annual report on the progress of co-ordination of conceptual metadata.
- ### D. Experiences
32. At this moment Statistics Netherlands is running the above mentioned process without Theme Administrators. The Coordinator is acting like a Theme Administrator for all themes. One of the problems is that specialised subject matter knowledge cannot be expected from the Coordinator. It is also important to involve the subject matter departments more in the process of metadata co-ordination. A Theme Administrator should be and feel responsible for its theme. For this reason Statistics Netherlands will establish the role of Theme Administrator this year.

V. SUMMARY AND CONCLUSIONS

33. More and more external registrations become available. It is a challenge of a NSI to enrich secondary sources by combining them into a coherent and consistent set of concepts and figures. Based on this basic set a NSI is more flexible to change its output if there are new information needs. Only changing an organisation from a stove pipe into a process oriented organisation is not sufficient. ICT must also support the new processes in an efficient way. A business architecture tries to translate the strategic goals of an organisation into an optimal organizational structure including the supporting ICT systems. Multiple use of tools within a common context forces the need for a good set of appropriate metadata. This set and the knowledge of combining multiple sources is the main capital of a NSI.

34. Building a set of consistent conceptual metadata is a labour intensive job. Strong co-ordination is needed between the different thematical areas and departments. For process and quality metadata it is even not quiet clear what kind of metadata is needed and what kind of functions metadata should or could fore fill. Statistics Netherlands is looking for best practises within this field.

REFERENCES

- Altena, J.W. and A.J. Willeboordse (1997): Matrixkunde of "The Art of Cubism" (Dutch only). Statistics Netherlands, Voorburg.
- Bethlehem, J.G., J.P. Kent, A.J. Willeboordse and W. Ypma (1999): On the use of metadata in statistical processing. Third Conference on Output Databases, Canberra, March 1999.
- Keller, W.J. and A.J. Willeboordse (2001): New Methods for Statistical Processing in a New Organization Environment
- Keller, W.J. and A.J. Willeboordse and W.F. Ypma (1999): Statistical Processing in the New Millennium. Proceedings of Statistics Canada Symposium 99, Combining Data from Different Sources, Ottawa, May 1999.
- Kooiman, P., A.H. Kroese and R.H. Renssen (2000): Official Statistics: an estimation strategy for the IT-era. XIV Conference of the International Association for Statistical Computation, Utrecht, August 2000.
- Willeboordse, A..J. (2000): Towards a New Statistics Netherlands. Netherlands Official Statistics, Spring 2000, Statistics Netherlands, Voorburg.
- Willeboordse, A.J (2004): A generic framework for the structuring of conceptual meta data, Paper prepared for the meeting of the Neuchatel group, June 2004, Neuchatel

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 (i): Metadata in a Corporate Context

THE STATISTICAL AND GEOSCIENTIFIC IBGE's METADATA SYSTEM

Supporting Paper

Submitted by IBGE, Brazil ¹

Abstract

1. This paper focuses on the Metadata (MetaBD) project of the Instituto Brasileiro de Geografia e Estatística (Brazilian Institute of Geography and Statistics) - IBGE. This project started in the 1990's and is in continuous evolution, due to users' new demands and to changes in the Information Technology infrastructure. The MetaBD project allows an effective data administration and enables the planning and design of the surveys and censuses, supporting most of the production processes (capture, editing, imputation, tabulation and dissemination) on a centralized basis.

2. The MetaBD is actively integrated to other institutional "*ad-hoc*" data processing and dissemination systems. It describes both micro-data and data aggregates, and enables access to the official statistical publications of IBGE. Quality indicators for statistical products and processes are stored according to the recommendations contained in the Program of Statistical Cooperation for the European Union and Mercosul and Chile.

3. In addition to statistical metadata, the MetaBD database also incorporates geographical and cartographic metadata, supporting different technical areas in their need for data descriptors.

Keywords: statistical metadata; geoscientific metadata; corporate metadata systems.

I. INTRODUCTION

4. Metadata are useful to identify, locate, understand, manage and use statistical and geo-scientific data originated from surveys, censuses and projects in general elaborated by the Brazilian Institute of Geography and Statistics (IBGE), the national statistical and geoscientific agency of Brazil.

¹ Prepared by Luigino Palermo, luigino@ibge.gov.br

5. As these data sets grow in number and diversity, are produced by different departments and are made available through communication networks (e.g., LAN, internet), it is essential to describe them in a standard and centralized basis, allowing knowledge about data to be shared among technical areas.
6. The great amount of statistical data at IBGE must be described, in order to allow researchers and the Brazilian society to access it both in an organized and safe way. Besides that, geographical data are of a distinct nature and both kinds of data can be kept in many forms, varying from flat files to relational databases or geographical information systems.
7. In the beginning of the 1990's, IBGE developed an initial version of the Metadata system, basically concerned with helping the task of administration of micro-data files delivered to the Database department, usually at the end of a statistical job. With the advance and progress of IT since then, and at the same time with users demands having been changing considerably, that system has evolved to the present version, so called MetaBD (abbreviation in Portuguese for Meta-Database), with new purposes:
- To offer a centralized metadata system to the data producers areas, which can feed the different kinds of metadata along the production phases of a statistical survey or census;
 - To make the MetaBD database active to the execution of homologated systems at IBGE, mostly available for processing some of the production phases;
 - To become a reference to other institutional products, besides the micro-data files for example: aggregate data, dissemination medias (statistical publications, cartographic documents, CD-ROMs, etc), special collections (Natural Resources area), among others;
 - To allow the retrieval of statistical metadata following varied criteria such as: survey theme (population, health, industry, etc), responsible department or geographic level of dissemination;
 - Provide free-indexing mechanisms, making it easy for inexperienced users to search for metadata;
 - Include metadata provided by geo-scientific production at IBGE.
8. The purpose of this paper is to present, in a concise way, some aspects of the IBGE institutional MetaBD system, and how metadata were incorporated to the daily job of the agency. These aspects will be next organized in 3 on topics: statistical context of metadata, geoscientific context of metadata and implementation issues.

II. STATISTICAL CONTEXT OF IBGE'S METADATA

A. Micro-data

9. The statistical metadata set covered by MetaBD system was the result of a study of the production phases of statistical surveys and census [Silva1997] and also from information obtained in meetings and interviews involving IBGE's technicians.
10. Nowadays IBGE's statisticians count on facilities to store surveys' metadata into MetaBD database along with the survey production, from the planning phase to the dissemination phase (Table 1).
11. Data and respective metadata can be accessed during the whole survey's cycle of life, including Data Collection and Capture, allowing quality to be assured after each stage of production, by means of data analysis. However users are grouped in clusters, depending on the nature of the task they are allowed to do, and their actions are restricted according to the metadata status.
12. The MetaBD system also offers the storing of editing rules plans of surveys, to be used in the Editing and Imputation phase, through writing them using natural text or CryptaX language. CryptaX [Hanono1996] is a system developed in IBGE for the editing and imputation phase. In this case, when running CryptaX apuration programs, the MetaBD database is active, because it validates the editing rules, using data dictionary informations created in previous phases.

Phases	Metadata Group	Metadata Item
Planning	Survey Objective	Objective
		Main variables
		Concepts
	Resources and Deadlines	- - -
	Target Population	Geographic Coverage
		Investigation Unit
		Survey Scope
		Territorial hierarchy
	Process Type	Process Type (census, survey, administrative registers etc)
	Periodicity	Periodicity
	Survey Plan	Methodology
	Data Collection Method	Data Collection Techniques (CAPI, CASI, CATI, PAPI)
Questionnaire Definition	Questionnaire Image	
	Variable Concepts	
Data Processing	Data Collection and Capture	Notes about Survey Occurrence
	Editing and Imputation	Editing and Imputation Plans
		Physical characteristics of data elements - dictionaries (conventional files) and databases
	Data Analysis	Classification
		Cross Tabulation
		Derivation Algorithm
		Quality Indicators
Dictionaries		
Data Dissemination	Data Dissemination	Publication
		Tabulation Plan
		Constraints (e.g., census secret etc.)
		Special Collections

Table 1 - Stored metadata per survey production phase at IBGE

B. Aggregate Data and Tabular Plans

13. In the same way as for micro-data, the MetaBD system can handle descriptions of the institutional aggregate datasets. The main IBGE system for on-line queries of aggregate data, available in the official IBGE web site (www.ibge.gov.br) in the internet [Figueredo2005] is the SIDRA² system, and it uses those aggregate data descriptions, keeping MetaBD active also during the dissemination phase.

14. More recently, this interaction between MetaBD and Sidra systems led to the development of the Sidra-Tabula system [Hanono2005] for generation and publication of tabular plan's tables of a survey, in a standard way, considerably reducing the time spent on tabulation process.

15. Figure 1 illustrates the communication between both systems, that allows the integration of the aggregate data production process with its dissemination process.

² Abbreviation for "Sistema IBGE de Recuperação Automática" (IBGE System for Automatic Retrieval)

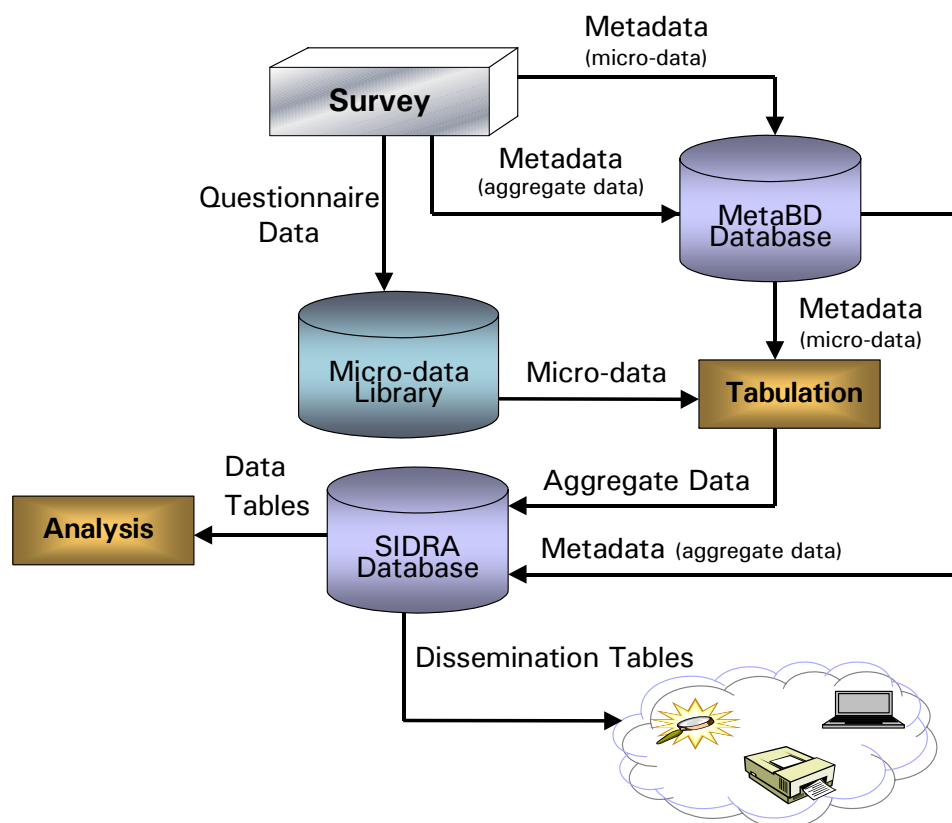


Figure 1 - Integration between SIDRA and MetaBD database

C. Quality Indicators

16. According to the directions of the Program of Statistical Cooperation of the European Union and the Mercosur, IBGE has been assessing and improving the quality of its statistical processes and products [Bianchini2004].

17. A quality methodological study was conducted by a group of experts from European Union and Latin-American agencies in 2002, which was considered strategic subject in the scope of this cooperation program. The methodological study listed nine quality dimensions that should be observed: relevance, accuracy, timeliness, punctuality, accessibility, transparency of product, comparability, coherence and exhaustiveness.

18. Matrices of quality indicators by type of statistical process (censuses, surveys and administrative registers) were created for each phase of the process, in a total of about fifty different quality indicators.

19. A standard report was specified to be elaborated at the end of each statistical process, comprising three parts: characteristics of the process, product quality and process quality. The quality indicators displayed in this written report would permit a comparison of quality between data collections, and would highlight the progress or the strengths and weaknesses in the statistical process and their products.

III. GEOSCIENTIFIC CONTEXT OF IBGE'S METADATA

20. The projects in the areas of Geography and Cartography are of a different nature: the items of the geographical and cartographic metadata depend on the theme/subject being processed, whereas in the statistical context the majority of the items are the same, with their contents varying according to the survey/census being focused.

21. This observation lead us to approach the treatment of these metadata in the new version of the MetaBD with a different strategy, where each project of the geoscientific area of IBGE is individually analyzed, and afterwards has its metadata incorporated, what in general implies in creating new metadata entities.

22. Presently, there are two projects which have been successfully incorporated to the MetaBD, both in Cartography: the 2000 Population Census Digital Municipal Maps and the Digital Index Map (which was already available in CD-ROM). In both cases the goal was to extend and to make more democratic the access to the information contents of IBGE maps.

23. At the moment, the first project in the field of Geography is been implemented, more specifically in the theme of Natural Resources, describing IBGE's accumulated intellectual patrimony about vegetal and animal species - some of them threatened of extinction - kept in an ecological reserve belonging to the institution and located in the Brazilian savanna in the Center-West Region of Brazil.

IV. IMPLEMENTATION

24. The system implementation was performed by analysts and programmers of the MetaBD project. In a few occasions, they counted with the orientation of external consultants, especially when a new technology was being used.

A. MetaBD Database

25. The conceptual data model of the new MetaBD system was conceived after several meetings gathering researchers and database experts of IBGE; from the results of these meetings, emerged the first design of the MetaBD's database, whose macro aspects can be observed in figure 2.

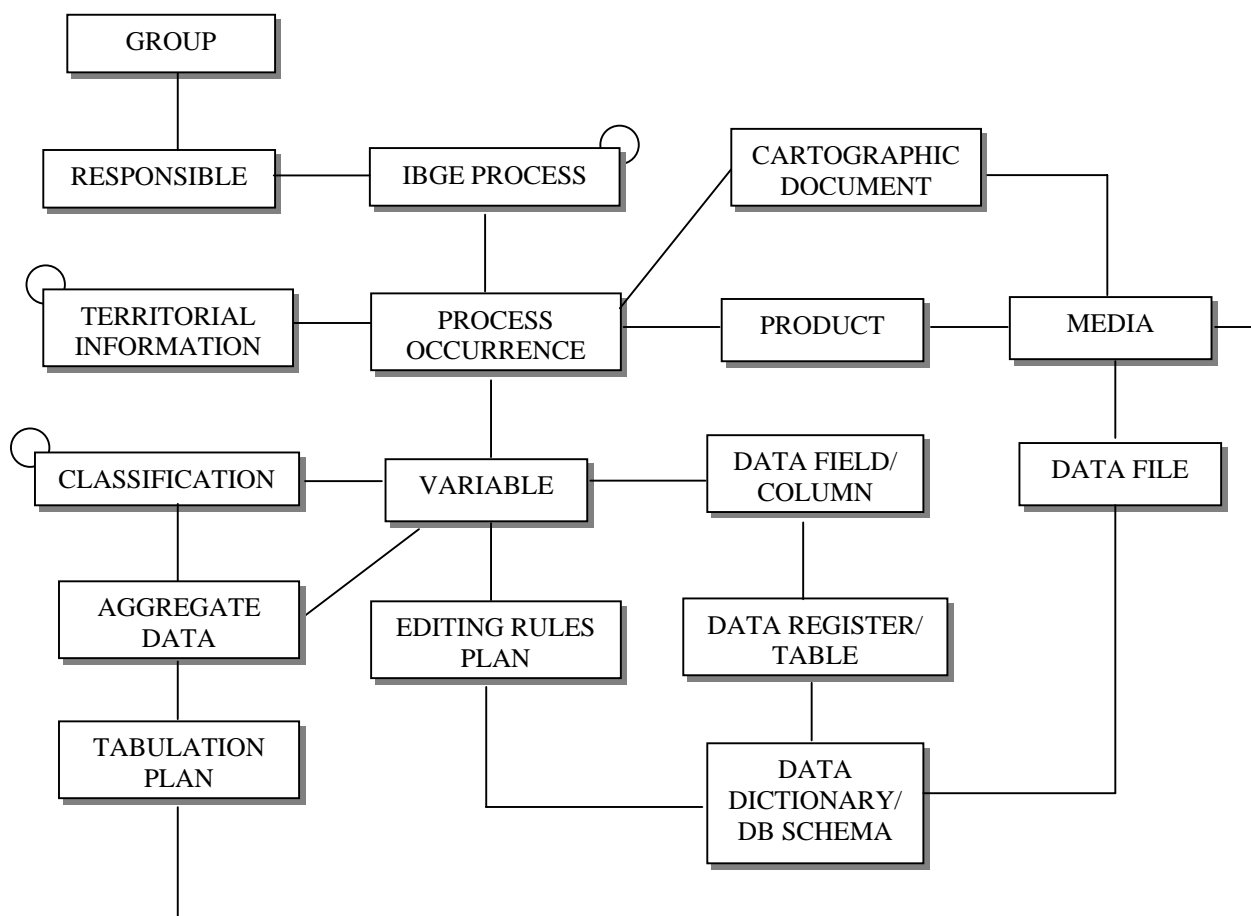


Figure 2 - MetaBD Main Entities

26. The MetaBD database became operational in January of 2000, but its data model still keeps evolving up today, getting new extensions to cover nonexistent contents. Among the most recent extensions, we can highlight the geoscientific metadata for cartographic documents, the tabular plans' metadata, and the quality

indicators' metadata. For this last, a data model extension (figure 3) was conceived, updating some previous entities and creating others, and of course adapting metadata access applications, which will be commented further on this paper.

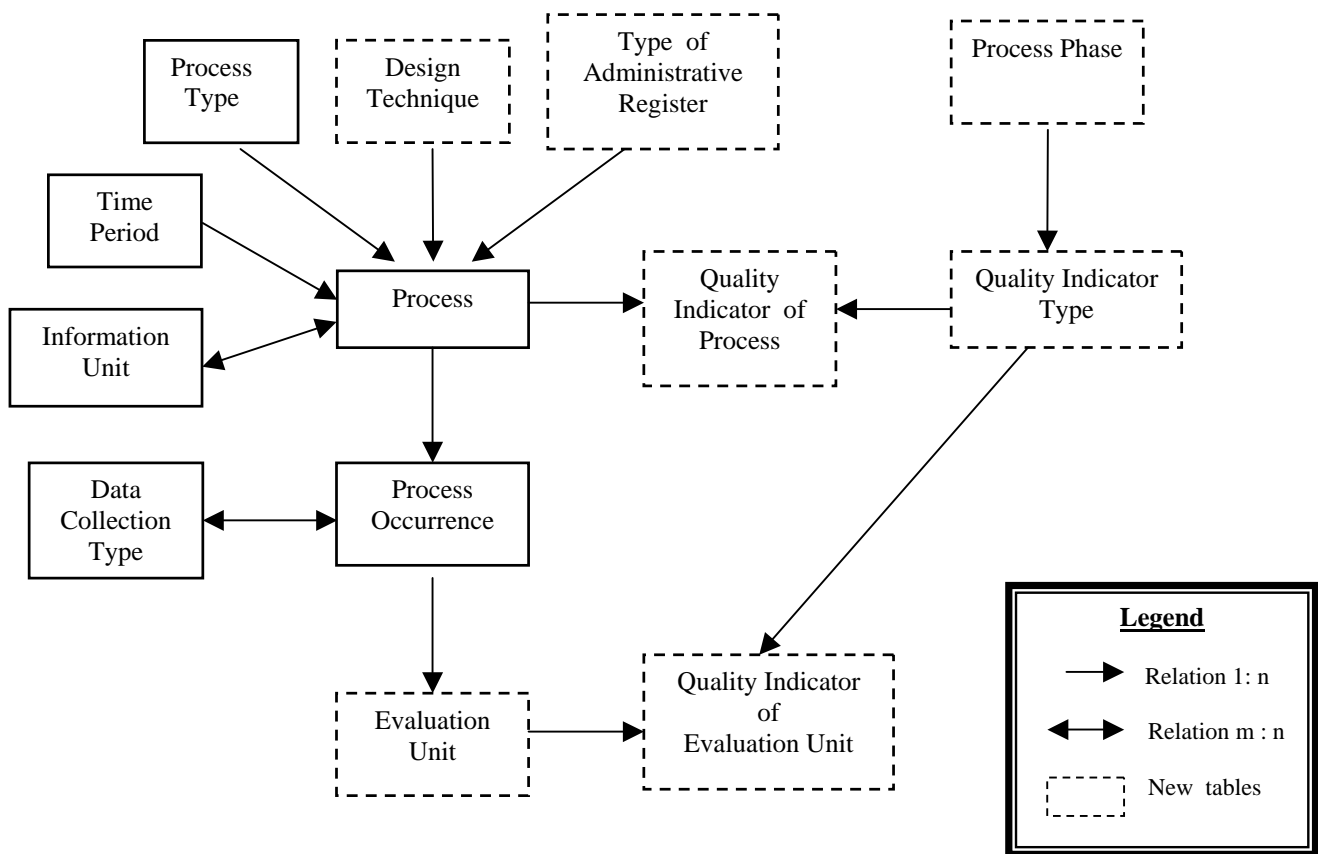


Figure 3 - MetaBD Data Model Changes for Storing Quality Indicators

27. At present, the MetaBD scheme is logically divided into following sub-schema, exclusively for management reasons³:

- IBGE Collection processes (censuses, surveys etc.) and Geoscientific projects - basic metadata that is not time dependent. Examples: name, acronym, periodicity, unit of observation, geographical coverage, background, methodology etc. (12 tables);
- Statistical or geoscientific events - metadata about surveys, censuses, and others but specific on their time occurrences, including variable concepts and pertaining documents (17 tables);
- Questionnaire - statistical metadata (9 tables);
- Editing and Derivation - metadata of editing rules plans and algorithms of derived variables (11 tables);
- Dictionary or Schema - metadata about dictionaries of micro-data files and about relational databases (23 tables);
- Territorial - metadata of geographical levels and items (8 tables);
- Aggregate - metadata of aggregate data (21 tables);
- Publication - metadata of tabular plans (6 tables);
- Cartographic - metadata of cartographic documents (33 tables);

³ The number of tables enclosed in parentheses includes the entities themselves, their relationships and auxiliary control tables.

- j. Results - metadata of products (15 tables);
- k. Media storage - products' metadata (21 tables);
- l. Quality - metadata of the quality of the statistical processes and products (8 tables);
- m. Safety - metadata for users, groups, and their authorizations (12 tables).

28. Following the data processing orientations of the institution, the relational ORACLE DBMS was chosen, which version nowadays ratified is Oracle 9i. The Oracle Designer tool has made database administration easier, for tasks such as the chartering of sub-schema and the semi-automatic generation of SQL scripts - to create about 200 tables and their indices.

29. Considering the exposition of IBGE's metadata to foreign institutes and partners, and foreseeing the availability of the MetaBD in the Internet, there was a concern in storing name and description textual fields in another two languages: English and Spanish. Description fields are stored as Oracle CLOB (i.e., Character Long Object) objects.

B. Metadata Access

30. The MetaBD system offers two applications for accessing metadata, available in the Windows environment: one for metadata loading and updating, and another for querying metadata.

31. The metadata loading and updating application is of the type Client/Server, developed in Visual Basic, which demands an initial set up, besides the needed basic software: Oracle Client and some ODBC driver for Oracle. Its use is restricted to the technical staff members - responsible for the data production, loading and updating the metadata - and to IBGE's Data Administration Team.

32. The Data Administration Team has access to all metadata and is in charge of defining authorizations based on each users group's profile and on sets of operations allowed to that group over MetaBD tables. Profiles and sets of operations are implemented using Oracle roles, in such a way that this safety mechanism is naturally supported by the DBMS.

33. The menu structure of the application is organized according to metadata context: statistical, geoscientific and administrative⁴. Menu options that do not apply to a specific user profile are not displayed. Thus, when specifying a user of the loading and updating application, it is necessary to inform the surveys under his/her responsibility, as well as the group to which he/she belongs (or his/her profile, to be provided in a new group, in the case he/she does not fit any group).

34. Concerning statistical metadata, it is also possible to store images (e.g., the survey's logo, the survey's questionnaire) or still to store links pointing to documentation files written in some textual editing application like MS-Word.

35. The MetaBD database can be viewed through a user-friendly web application, so far only for internal users. The decision-making process to make it available in the Internet depends on institutional policies and includes subject matters such as the definition and quality of the items to be released.

36. The querying application was entirely built on the MS.Net platform, using MS Windows 2000 Server as operating system and IIS as web server software, and can be reached starting from a browser (like Internet Explorer or another) and does not require any software component to be set up.

37. When accessing the querying application's home page, the first screen to be displayed brings information on recent data and metadata incorporated to the institutional databases. The MetaBD project team also uses this space to inform about implementation of new application facilities.

⁴ In general this includes auxiliary tables, for instance: pre-defined attribute values.

38. Among the most attractive facilities of the querying application is the possibility of producing files containing the dictionary layout in different formats/languages, to be used off-line by other data processing and analysis tools, like SAS [Palermo2003], REDATAM and CriptaX.

39. The querying application also offers some reports about MetaBD database entities: Survey and related Variables, Data Dictionaries, Classifications, Quality Indicators.

C. Hardware Infra-structure

40. The hardware infrastructure of the MetaBD system can be seen in figure 4. Database and web servers of the MetaBD system share same equipment of those of the SIDRA system, and are based on Intel Pentium Xeon Dual-processors machines.

41. A fast cross-over channel connects the database and the web servers, streamlining the metadata access for queries made via web. The database server is connected to the local area network of IBGE for execution of both access applications, but limited only to internal use (corporate network). Presently only one external Brazilian organism, belonging to the planning sector of the national government is able to query the MetaBD database over the internet.

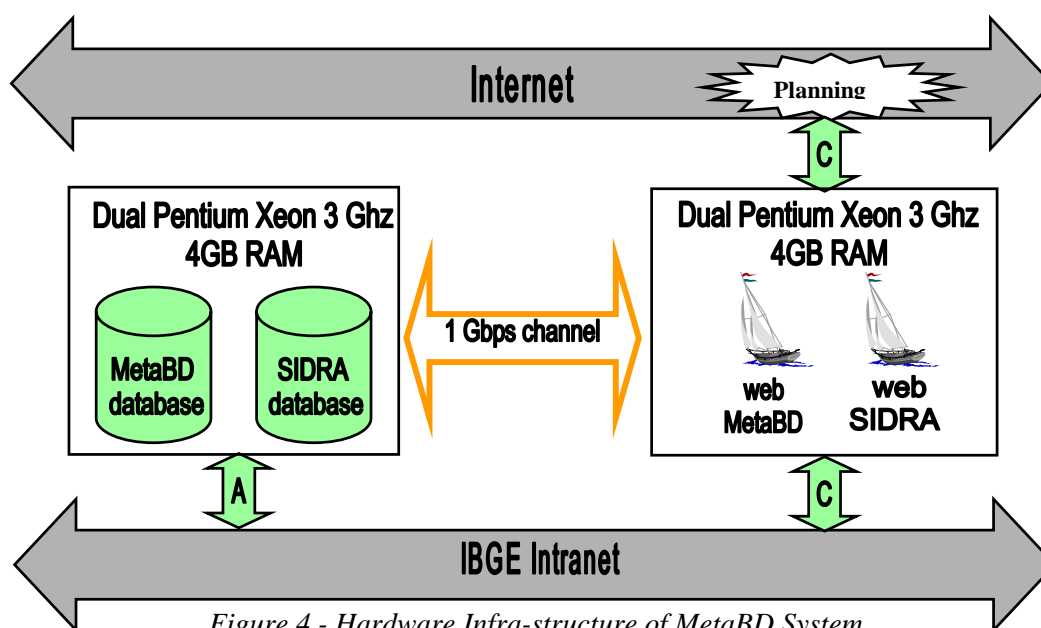


Figure 4 - Hardware Infra-structure of MetaBD System

D. Foreseen Changes and Improvements

- Improvement of the descriptors' quality (e.g., reducing conflict of concepts);
- Development of a new 3-tier levels version of the loading and updating application, suitable for web use;
- Design of a better web interface for the metadata querying application, then allowing internet access to MetaBD database, and perhaps some adjustments for supporting less popular browsers;
- Release of a sub-set of geoscientific metadata to other national geographic agencies, according to the Federal Geographic Data Committee - FGDC, to meet staff members demands;
- Take into consideration free-software options, since the Brazilian government is moving towards this direction. The use of PostGreSQL database management system is under study;
- Integration into new editing/imputation/analysis tools.

V. CONCLUSIONS

42. In spite of IBGE's limited resources, the most relevant motivation of the institutional metadata system has always been to transform data into information - for the benefit of staff members not involved in data production and of society in general. This emphasizes IBGE's institutional mission which is "to portray Brazil with the necessary information to the knowledge of its reality and to the practice of citizenship."

43. To achieve this goal, the Data Administration Team and the staff members responsible for IBGE's data production have been documenting the data, adding - with the help of the MetaBD system - its semantic metadata (concepts, background, methodology, sources, geographical coverage etc.), syntactic metadata (micro/macro data fields', registers' and files' descriptions, relational databases' descriptions etc.) and pragmatic metadata (editing rules plans, tabular plans, derivation algorithms etc.), without which those data would be either incomplete or useless.

44. In the future, the challenge to overcome is the appropriateness of the MetaBD system to metadata patterns, a subject that was not taken into account when the system was designed.

45. Presently the MetaBD database stores, in statistical metadata, approximately 117 surveys and censuses (being almost 6,000 time events of these collection processes), over 50,000 variables described in 824 dictionaries and stored in 2,764 physical files.

46. In the geoscientific area, the numbers are also expressive, considering that only two cartographic projects were carried out up to now, in a total of 12,363 stored maps in 8,194 physical files.

VI. REFERENCES

- [Bianchini2004] Bianchini, Z.M. & Albieri, S. - "A prototype of the Quality Indicators System at the Brazilian Institute of Geography and Statistics". *European Conference on Quality and Methodology in Official Statistics*, Mainz, Germany, 2004.
- [Figueredo2005] Figueredo, L.A.G.A. & Masello, J. - "SIDRA - Banco de Dados Agregados - Arquivo de Dados Agregados - Definição e Carga" (SIDRA - Aggregate Database - Definition and Loading). *Technical Note 01/05, Diretoria de Informática, IBGE*, Rio de Janeiro, Brazil, 2005.
- [Hanono2005] Hanono, R.M Figueredo, L.A.G.A. and Masello, J. - "Processo de Tabulação SIDRA-Tabula" (SIDRA-Tabula Tabulation Process). *Technical Note 02/05, Diretoria de Informática, IBGE*, Rio de Janeiro, Brazil, 2005.
- [Hanono 1996] Hanono, R.M. and Barbosa, D.M.R. - "CRIPTAX - A Generalized Editing Application Generator", *Work Session on Statistical Data Editing*, Voorburg, Netherlands, 1996.
- [Palermo2003] Palermo, L.I. and Masello, J. - "Metadados e Geração Automática de Programas" (Metadata and Code Automatic Generation). *Work presented in the Brazilian Annual SAS Users Group Meeting*, São Paulo, Brasil, 2003.
- [Silva1997] Silva, A.F. - "Aporte da Ciência da Informação à Organização de Metadados para Acervos Estatísticos" (Contribution of the Information Science to the Metadata Organization for Statistical Data Libraries). *MSc Dissertation, Pos-Graduation Course in Information Science, UFRJ/ECO and CNPQ/IBICT*, Rio de Janeiro, Brazil, 1997.

**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 (i) Metadata in a Corporate Context

THE BNSI EXPERIENCE IN METADATA COLLECTION AND ORGANIZATION
Submitted by NSI - Bulgaria¹

I. INTRODUCTION

The National Statistical Institute of the Republic of Bulgaria has begun developing an Integrated Statistical Information System (ISIS). The lack of ensured funding for the complete development of ISIS along with the architecture defined in the Strategy for development of Integrated Statistical Information System and the IT Strategy² requires applying a gradual system of development component by component. The paper presents the approach and organization of ISIS metadata collection at the NSI.

**II. ESSENCE AND GOAL OF ISIS AND PRINCIPLES OF DEVELOPMENT
COMPONENT BY COMPONENT**

1. ISIS development aims at setting up mutually connected object and process oriented information systems using common data bases with controlled access, allowing shared use of the NSI source by experts at the NSI and external experts. A key ISIS component is the Metadata base that is also the basis for the ISIS driving. The architecture of ISIS is defined in accordance with the UNECE Standard 51³. The BNSI top management and experts understand the advantages of the complex ISIS development as an unified integrated system, however the lack of funding is the reason for looking for technological innovation by gradual system development component by component.

2. The gradual development of ISIS could be realized in different ways depending on the consecutiveness in the development of functional and infrastructural components. The advantage of this approach towards a higher technological SIS is conformity with the

¹ Prepared by Svetlana Ganeva, e-mail sganeva@nsi.bg.

² The Strategy for ISIS development and the IT Strategy were developed within the framework of the National Phare Programme project BG 00.06.04 with the technical assistance of Swedish and German experts, as well as the local experts. They are a part of the complex Strategy for Statistics development, which is currently being elaborated and discussed at the BNSI.

³ [Information Systems Architecture for National and International Statistical Offices](#); Statistical Standards and Studies No. 51, Geneva, 1999.

available sources. The disadvantage is the risk of the NSI having to develop a separate project for integration of the already developed ISIS components. The reasons could be the necessity to apply contemporary technology or the specific features of tender procedures under the Procurement law.

3. ISIS development in the BNSI is realized by adhering to the following requirements as a guiding principles:

- The strict adherence to the ideology and architecture of ISIS defined in the Strategy for ISIS development in the work on developing separate ISIS components.
- Common metadata model implementation in the application work on the development of separate ISIS components.
- Providing interfaces between information systems if possible, especially for the metadata exchange. This requirement is compulsory for infrastructural components providing metadata for other information systems-ISIS components.
- The components exploitation must follow developed and approved internal standards.
- The ISIS components must include built in the compulsory data and metadata exchange standards of Eurostat and other international statistical organizations.
- Consolidation of the IT environment and IT elements to be achieved by developing the ISIS components in accordance with the IT Strategy requirements.
- Common development tools to be used in the development of the system.

4. An important requirement for the implemented approach is defining the order for the development of components, while retaining the continuity of the existing statistical information system exploitation without disruption so that every following component be integrated with the already existing ones. This means that the stove-pipe organization of the SIS is maintained. The priority development of the infrastructural components is a possibility to use them in the functional components development. Thus the implemented approach determines the integration of following components to the system until its development in accordance with the approved principles.

5. The BNSI is presently working on development of the following information systems as ISIS components:

- Register of the statistical units (RSU);
- Information system “Statistical classifications” (ISSC);
- Information system “Planning and design of the statistical survey” (ISPDSS).

6. BNSI has developed a bilingual electronic vocabulary of the statistical terms and their definitions to be used internally in the NSI Central office.

III. STANDARDS FOR THE METADATA DESCRIPTION IN INFORMATION SYSTEMS

7. The standard object approach in IS development is implemented through the definition and usage of common objects in all IS. The model used is the MetaNet Reference Model, recommended by EU experts in National Phare project. The BNSI experts discussed the information objects and their attributes one by one and the accepted objects were included in the elaborated technical specifications for information systems (the revision of the technical specifications was required by the contract).

8. The information objects of the ISPDSS are as follows:

- Statistical survey;
- Statistical observation;
- Observation unit (object);
- Global variable;
- Object variable;
- Value domain;
- Question;
- Table;
- Questionnaire block;
- Questionnaire;
- Statistical output;
- Statistical infrastructure task;
- Document;
- Procedure;
- File from an internal source;
- File from an administrative source;
- Population.

These information objects provide the statistical activities description from the planning and design stage to the statistical products production, including design and output of the documents going along with the statistical survey – questionnaire, instructions, methodology, technical specification for the data processing and so on. The information objects attributes allow keeping the link between objects and storing information for the object's history and changes.

9. It was necessary to define new information objects in accordance with the statistical practice traditions in Bulgaria as follows:

- Statistical observation –the national practice shows that it is possible to have one survey realized by more then one observation. At the same time one observation could be useful for more then one survey. By using this information object is possible to describe in a proper way the existing traditions in the statistical survey organization and to manage them.
- Statistical infrastructure task – the information object is intended to collect the meta information for the activities of the statistical office servicing conducted statistical surveys.
- File from an internal source and File from an administrative source –these information objects allow studying the information flows, providing information for the statistical surveys. They are also a tool for the macroeconomic surveys description, which – as a rule – use the information from the statistical surveys (micro databases) for macroeconomic data production.
- There is also an information object „Procedure” intended for procedures description – technological sequence of activities – with a view of the future driving the processes in ISIS.

10. Standard outputs are defined in the ISPDSS, some of them are extremely important for the management of the statistical information system. Such outputs are the National Programme for the Statistical Surveys, the Calendar presenting the results of the statistical surveys and the List of the standard statistical indicators. The ISPDSS provides introduction of data concerning the time and funding needed as a prerequisite for better processes management.

11. The standard for the classifications description in the IS “Statistical classifications” is a combination between MetaNet Reference model and CLASET model. The information objects in the ISSC allow for both the development and maintenance of “standard” classifications and the establishing and maintenance of the statistical classifications for separate surveys, as well as for extraction of information for classifications used in a definite point in time. It is foreseen to load main classifications used after 1990 in the system. The classifications, the relations between them and the dynamics of their development are described using the information objects as follows:

- Classification family;
- Classification sub-family;
- Classification;
- Classification version;
- Classification variant;
- Classification level;
- Classification item;
- Best classification practice (Case law);
- Table;
- Relation;
- Dynamics;
- Document.

12. Information objects in the Register of the statistical units have an important place in ISIS. Objects in this system include the administrative and statistical units, forming the structure of the national economy. These units are defined in accordance with national and European legislation as follows:

- Administrative units:
 - Legal Unit;
 - Local Legal Unit.
- Statistical units:
 - Enterprise;
 - Local Unit;
 - Kind of Activity Unit;
 - Local Kind of Activity Unit;
 - Enterprise Group.

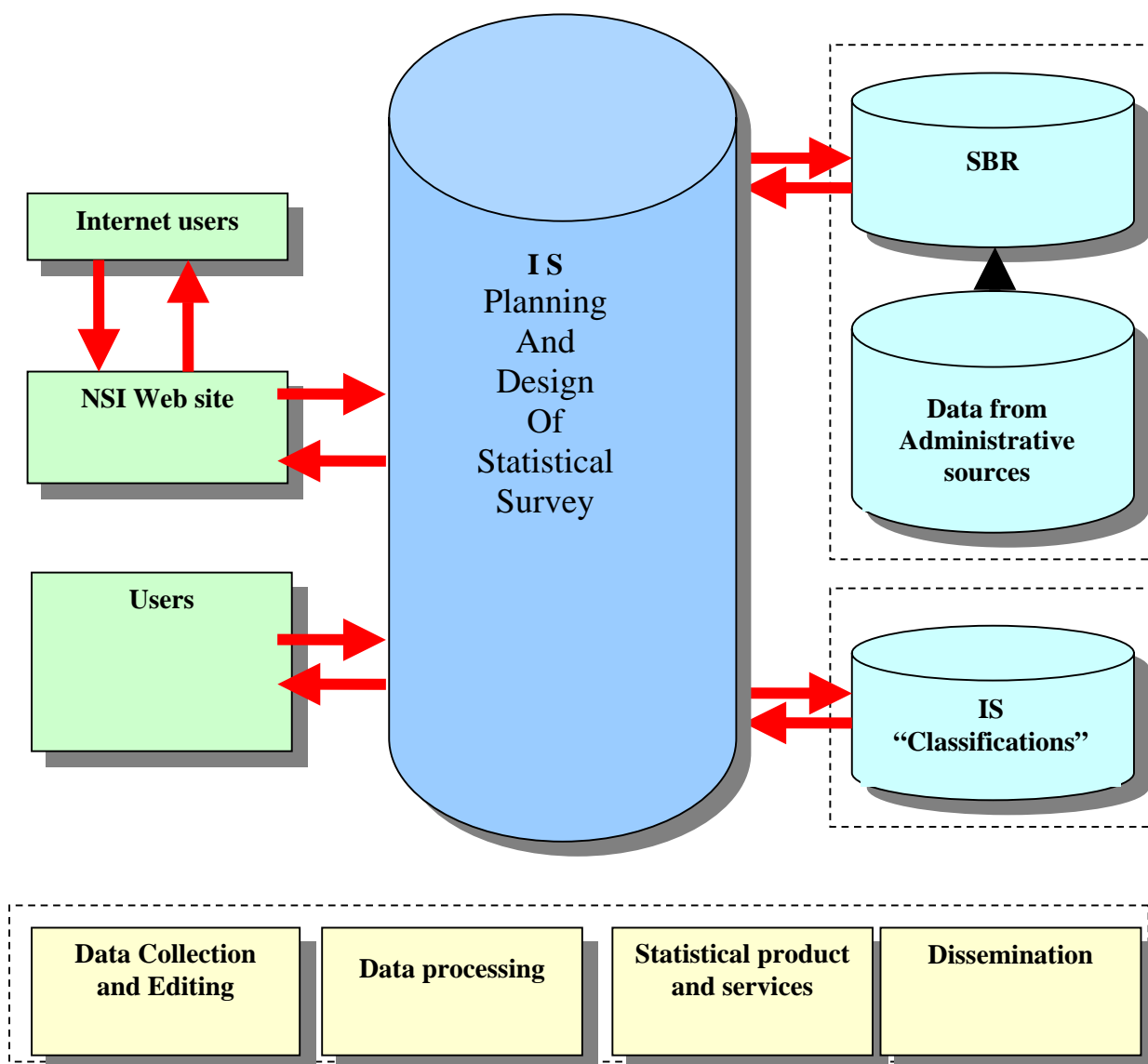
13. The attributes of the statistical units and the statistical units themselves are described in the ISPDSS as the population and sample are defined based on the objects of the Register of the statistical units.

IV. THE PLACE OF THE DEVELOPING COMPONENTS IN THE ISIS ARCHITECTURE

14. The place of the presently developed information system in the architecture of the Integrated Statistical Information System is defined in accordance with their functionality:

- Register of the statistical units is an infrastructural component including the information for the statistical units, their characteristics, their demography and development their “participation” in statistical surveys;
- IS “Statistical classifications” is intended to collect and store meta-information;
- The Information system “Planning and design of the statistical survey” is a functional component developing as an application tool for collection and storage of data describing statistical surveys.

The relations between developing systems are presented by the following scheme:



V. ORGANIZATION OF METADATA COLLECTION – REALIZATION THROUGH THE SI SYSTEMS

15. During the implementation phase of the information system “Planning and design of the statistical survey” it is envisaged to create an organization and through the description of statistical surveys to load information on statistical surveys. Thus the ISPDSS becomes an application tool for metadata collection for surveys.

16. Work on the development of information systems at NSI is organized by the project leader, who is exclusively entrusted with task. Coordination is ensured by the Coordination Council, headed by the NSI vice-president and including all directors. This allows work to be monitored by the NSI management and experts on all levels. A task force for the development of each information system was set up, consisting of representatives of all directorates participating actively in the discussion of the relevant issues as well as in the design and development phase and the implementation phase. Project management is done by means of a project management timetable - ISO 9001 for project management.

17. During the implementation phase of the information systems it is envisaged to enter the metadata on statistical surveys carried out by NSI in accordance with a programme adopted by the NSI President. The programme shall be based on decisions taken concerning the consecutiveness of implementation of the data collection and editing component as well as the possibilities for migration of existing microdata databases in survey registers.

VI. FUTURE TASKS

18. Work is in progress and a tender procedure is under way for the development of a Technical specification for development of ISIS as well as on a detailed Technical specification for the development of the Information system “Statistical observation” as a follow-up functional component of ISIS. Based on this detailed specification this functional component has been commissioned as a separate phase design under the contract for the PHARE 2004 National Programme. As a minimum task, it is envisaged to develop the functional component “Data collection and editing” along with the creation of the Observation registers and the Final observation registers as a Microdatabase. The national programme also envisages development of the Data from administrative sources database with a view of creating a database of existing administrative registers usable for statistical purposes with a view of a description of their constituent units and their characteristics and their future integration as component of ISIS. As a next phase, it is envisaged to create follow-up functional components – the processing of statistical data with a view of development of the Macro database and the Statistical products dissemination component, along with development of the Dissemination database as a copy of the Macro database, concerning data which can be disseminated. The consecutiveness of setting up the components of ISIS shall be defined precisely in the technical specification for their development. There are options for developing the Metadata base as a key component of ISIS intended as the integrated system management component. One of these options involves development of the Metadata base and gradually adding to it metadata from newly created components. Another option is to develop the components separately and finally integrate them, followed by the creation of the Metadata base as a common management component. In both cases, the process of building up the components must go in a direction of applying common standards and rules, in view of the possibility for their integration in ISIS.

**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 (i): Metadata in a Corporate Context

**MANAGING METADATA SYSTEM PROJECTS:
EXPERIENCES OF THE CZECH STATISTICAL OFFICE
Supporting Paper**

Submitted by the Czech Statistical Office, Czech Republic¹

I. AIM OF THE PAPER

1. The goal of this paper is to inform about past failures in developing and managing statistical metadata projects, and to demonstrate new management principles and approaches in creating a statistical metainformation system.

II. LESSONS LEARNED FROM THE PAST

2. The first statistical metadata project in the Czech Statistical Office began in the early 1980s. It concentrated on statistical classifications and statistical variables. The goal of the project was to improve e-production of statistical data.

3. The need to develop systematic metadata was initiated by information technology (IT). Subject matter statisticians and methodologists did not systematically cooperate in project development. The small metadata unit created for this purpose did not have sufficient authority to decide or solve all the cross-cutting issues accompanying the development and implementation of the metadata project.

4. Several attempts have been made to introduce this project into the broad statistical practice, but all of them have failed. The major reason was that the content part of the project was never approved by the top management, and thus there was no commitment from the subject matter statisticians and methodologists to use metadata in their activities. Partially positive results in IT production were not sufficient and very often led to keeping duplicated metadata descriptions of statistical classifications and variables.

¹ Prepared by Ebbo Petrikovits, ebbo.petrikovits@czso.cz.

5. The reasons for such dissatisfying results were:
- a. The work on metadata project started in the ICT department without a clear and accepted vision.
 - b. The subject-matter statisticians were not involved in the topic; support from their side was completely missing.
 - c. The organizational framework of the project was not clearly specified
 - d. The top management was not well informed about the project and their support of it was very low.
 - e. There was limited general knowledge of the project inside the office.

III. NEW MANAGEMENT POLICY

6. The top management of the CZSO introduced a new policy in 2003 for the development and implementation of key statistical tasks. It defined the main development topics for statistics and launched new management methods of those projects.

7. In 2004 the top management launched a new project called *Reform of Statistical Survey System*. A project team and steering committee were established. Another key project - The *Public Database* - was approved, including its project team and coordination committee.

8. In early 2005 the top management approved a Conception (vision) of the *Statistical Metainformation System* (SMS). Several SMS project teams and steering committees were created. The development of the project was launched.

9. A management board for all above-mentioned projects was established. The top management directly supervises all activities of the management board. Their direct control of the development, follow-up and progress achieved in the projects, brought positive results in a relatively short time. Coordination in the development of all projects was ensured. Systematic involvement of CZSO's middle management was established.

IV. MANAGEMENT OF THE SMS PROJECT

10. The following 3-level management hierarchy was established for the development and implementation of the SMS project:

- a. The first level: the project teams established for individual subprojects defined in the Conception of the SMS. Project teams are composed of subject-matter statisticians, methodologists and IT experts.
- b. The second level: the SMS Task Force supervising and coordinating activities of the project teams. The Task Force is composed of the heads of the project teams, selected representatives of the middle management and head of the newly established SMS unit (head of the TF).
- c. The third level: The SMS Steering Committee. It is appointed by the top management and is composed of the directors of the subject matter statistical departments, a representative of the top management responsible for the methodology and SMS and the head of the SMS Task Force. The Vice-President of the CZSO chairs the Steering Committee and reports the outcomes of their meetings to the top management board.

11. **The SMS project teams** prepare basic documents (subject-matter and technical specifications) for the development of individual SMS subprojects. Each project team coordinates and organizes the work on the subproject. It also coordinates cooperation with subject matter departments and organizes workshops, seminars and training for statisticians and other staff of the CZSO.

12. **The SMS Task Force** coordinates the work of individual project teams and discusses and approves the documents prepared by them. They also prepare regular progress reports on results reached during last three

months, which are submitted to the SMS steering committee. Every progress report contains the evaluation of the previous 3-month period of work, indicates problems arising and outlines decisions expected by the steering committee.

13. **The SMS Steering Committee** meets every three months. It controls the progress achieved in SMS development, considers any problems, decides on solutions and approves composition of the project teams and the timetable for further work. They review the progress reports, approve the results of the work and consider and approve proposed changes in the individual SMS subprojects. The committee submits its draft conclusions to the top management board meeting and submits the progress report and minutes of their meetings to the top management.

V. CONTENT OF THE SMS PROJECT

14. The SMS Conception defines major functions of the SMS in the statistical organization. It determines the subjects of metadata description and defines the SMS sub-projects. The following SMS sub-projects are planned:

- a. Statistical classification;
- b. Statistical indicators (variables);
- c. Statistical task and statistical surveys;
- d. Administrative data;
- e. Statistical data repository;
- f. Respondents;
- g. External users;
- h. Dissemination;
- i. Knowledge base on statistical information system.

15. Furthermore, a special subproject – Global Architecture of the SMS (GA-SMS) – is under the development. Its goal is to define a unified architecture for all system and application functions.

16. When developing a GA-SMS, an urgent need to standardize major processes conducted by the CZSO arose. Special research work was launched for this purpose. One of the key processes identified was the collection, production and dissemination of statistical data. For effective functioning of this process the SMS tools are desperately needed. The SMS functions should support all activities of this key process. User friendliness of SMS tools should be ensured.

VI. MAJOR FINDINGS

17. The recent experience in the SMS development allows to make the following findings and/or recommendations:

- a. Permanent supervision of the SMS by the top management is a necessary precondition to make the project a success story.
- b. Regular follow-up of the SMS development and reporting on the results of the SMS subprojects are an important part of the project management.
- c. Systematic cooperation with statistical subject-matter experts and methodologists is inevitable.
- d. Focus on the subject-matter topics and use of SMS tools in the statistical practice is advisable.
- e. Importance of training and transfer of SMS know-how. The SMS methodology and organization of the work must be addressed. Users benefits should be clearly presented.
- f. Sharing information and knowledge between the SMS project teams (the intranet proved highly effective for this); and wide availability of information about the SMS development to statistical staff brought positive results.

VII. STATE- OF- ART OF THE SMS

18. In the time being the following SMS sub-projects are under the development: (i) GA-SMS, (ii) Classifications, (iii) Statistical Indicators and, (iv) Statistical Tasks and Surveys.

19. **Sub-project GA-SMS.** The preparation of functional specifications is under way.

20. **Sub-project Classifications.** Functional specifications including the classification model were approved. Technical specifications have been developed and approved. Revision and transition of the old classification system into the new one is on the way. Guidelines for all partners involved in the content administration of the subsystem are under preparation. Development of application software was launched.

21. **Sub-project Statistical Indicators (variables).** Functional specifications including metadata model for the description of statistical variables were developed and approved. Preparation of training workshop for subject-matter statisticians is under way. The aim of the workshop is to ensure sound understanding of the methods and techniques used by statisticians for defining and describing statistical variables, before starting such an exercise.

22. **Sub-project Statistical Tasks and Surveys.** Functional specifications are under preparation.

Prague 15 March 2006

**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)

Session (ii) Metadata Concepts, Standards, Models and Registries

METADATA STANDARDS AND THEIR SUPPORT OF DATA MANAGEMENT NEEDS ¹

Invited Paper

Submitted by Statistics Canada², Canada, and the Bureau of Labor Statistics³, United States
Session (ii) Metadata Concepts, Standards, Models and Registries

I. INTRODUCTION

1. For many years, metadata management has been an important concern in national and international statistical offices around the world. Statistical metadata, metadata for statistical data and processes, is used to enhance users' search and understanding of statistical data, improve and automate survey processing within each office, and facilitate statistical data harmonization, among many others. As a result, the area is a fertile ground for research and development. Many offices are using metadata driven systems to automate parts of the survey process (Johanis, 2000; Oakley, 2004, add references -- Netherlands, Norway, Slovenia, Sweden, Eurostat, OECD?)

2. Several things need to be understood and developed before metadata management and metadata driven systems can be built. Foremost, an understanding of what constitutes metadata for the problem at hand. Metadata is not an absolute concept. Data are not

¹ The opinions expressed in this paper are those of the authors and do not necessarily reflect the official policies of Statistics Canada or the Bureau of Labor Statistics.

² Prepared by Paul Johanis at Statistics Canada. Contact paul.johanis@statcan.ca.

³ Prepared by Daniel W. Gillman at the Bureau of Labor Statistics. Contact gillman.daniel@bls.gov.

metadata because of some inherent properties, they are metadata by use. So, metadata is a relative idea. Data become metadata when they are put into a descriptive relationship with something else (Gillman, 2005; Farance and Gillman, 2006).

3. Once the required metadata elements are understood, a model can be built. This is a data model of the metadata to be used. The model is a framework for how the metadata will be organized in a database, and the structure is often optimized in some way to enhance the uses of the database (Date, 2003). The common constructs among models and their attributes are the focus of the discussion in this paper, as metadata constructs are components of models.

4. Most situations require some amount of modeling work. A reasonable question when designing a model is "Has anyone else thought about this problem, and is there a solution I can borrow that will work for my situation?" Already existing models may not work at all, may work for some purposes but not others, or may work completely. For the models that fit partially, they can be made to work if they can be modified. This is often the case.

5. Where does one look for appropriate models? There are 4 possible answers: other statistical offices, commercial software vendors, published papers or books, and standards. Other statistical offices are a great source for metadata models, as several good metadata models have been developed there (Johanis, 2000; Sundgren, ???). Commercial software usually does not have appropriate metadata models, as the needs of statistical offices are too specialized, and commercializing specialized products does not pay off. Metadata models in books and papers are too high level, so not so useful for building systems. However, they are useful for conveying a conceptual framework, which is shared. Finally, standards are a good source for metadata models, because they contain much detail and are based on consensus among a wide group. Standards are often built by a community of practice, people in similar businesses, or otherwise having like concerns. This leads to the development of standards that appeal to specialized groups, e.g., the Data Documentation Initiative (ICPSR, n.d.).

6. Standards and other statistical offices seem to be the best sources for finding appropriate metadata models, and we will analyze several metadata schemes, which arose from these sources. Part of the analysis will include a discussion of common constructs. Regardless of the specifics of any given scheme, there are common metadata constructs used to describe statistical data. This paper will give an overview of these common constructs.

7. The paper is organized into several sections. We begin with a section on the theory of terminology. This provides a framework for commonality. Next, a discussion of statistical data based on terminology theory is provided. Then, we show that the ISO/IEC 11179 standard is an implementation of the theory. Therefore, the standard is common to all descriptive frameworks for statistical data. The Corporate Metadata Repository (CMR) model, an extension of the ISO/IEC 11179 standard into the statistical survey domain, is discussed. Following, is a description of the most important constructs for each of five metadata schemes: Common Warehouse Model (CWM), Data Documentation Initiative (DDI); eXtensible Business Reporting Language (XBRL), Neuchâtel Variables and Classification models; and Statistical Data and Metadata Exchange (SDMX). Finally, a comparison between the models and a framework for using the models together in a statistical office are provided.

II. THEORY OF TERMINOLOGY

A. Basic Definitions

8. Terminology is the study of concepts and their representations in special language. It is multidisciplinary, drawing support from many areas including logic, epistemology, philosophy of science, cognitive science, information science, and linguistics. Work in the area dates all the way back to the ancient Greek philosophers.

9. To begin, we describe some useful constructs from the theory of terminology. These come from several sources (Sager, 1990; ISO, 1999; ISO, 2000). The constructs and their definitions follow below:

- **object** - something conceivable or perceivable
- **property** - observation, used to describe or distinguish an *object* (e.g., "Dan has blue-gray eyes" means "blue-gray eyes" is the property of Dan. It is abstracted to a characteristic, color of eyes, of people - see *characteristic*.)
- **characteristic** - abstraction of a *property* of a set of *objects*
- **essential characteristic** - *characteristic* which is indispensable to understanding a *concept*
- **delimiting characteristic** - *essential characteristic* used for distinguishing a *concept* from related *concepts*
- **concept** - unit of knowledge created by a unique combination of *characteristics*
- **intension** - sum of *characteristics* that constitute a *concept*
- **extension** - set of *objects* to which a *concept* refers
- **definition** - expression of a *concept* through natural language, which specifies a unique *intension* and *extension*
- **concept system** - set of *concepts* structured according to the relations among them
- **designation** - representation of a *concept* by a sign, which denotes it
- **general concept** - *concept* with two or more *objects* that correspond to it (e.g., planet, tower)
- **individual concept** - *concept* with one *object* that corresponds to it (e.g., Saturn, Eiffel Tower)
- **generic concept** - *concept* in *generic relation* to another that has the narrower *intension*
- **specific concept** - *concept* in *generic relation* to another that has the broader *intension*
- **generic relation** - relation between two *concepts* where the *intension* of one of the *concepts* includes that of the other *concept* and at least one additional *delimiting characteristic*
- **subject field** - field of special knowledge

10. Designations come in three types: A term is a verbal designation of a general concept; an appellation is a verbal designation of an individual concept; and a symbol is any other designation. Signs, through which designations are represented, are left undefined, but a sign is what a person perceives and interprets as designating some concept. Basically, however, a sign is a concept whose extension is a set of perceivable objects. Examples of signs are each of the lines and dots on this page we interpret as words, letters, and punctuation. So, what we see and interpret is not really a sign, but an object in the extension of the sign. The objects **F** and **F** are in the extension of the same sign.

11. Characteristics are used in concept formation. They are abstracted from properties of objects and are used to form the intension of concepts. The objects whose properties are abstracted into the characteristics that form the intension of some concept make up its extension. Characteristics may be concepts in their own right, too. They are used in concept analysis, concept modeling, formulation of definitions, and even term formation.

12. The term *specialization* is often used to denote the creation of a specific concept in generic relation to a given, generic, one.

13. The ancient Greek philosophers began the study of terminology and concept formation in language (Wedberg, 1982), and they discovered a useful relationship between designation, concept, object, and definition, that is illustrated in Figure 1 (CEN, 1995). This diagram, minus the definition part, is often referred to as Ogden's Triangle (Ogden and Richard, 1989).

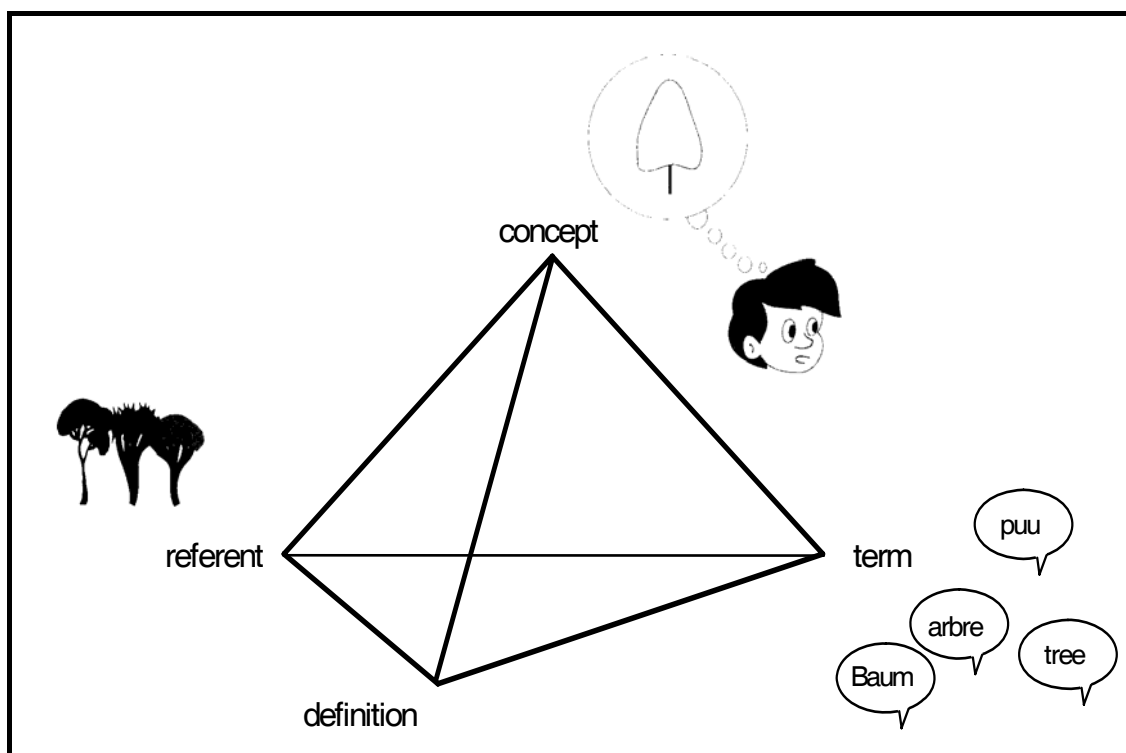


Figure 1. Relationships between referents (objects), concepts, terms (more generally designations), and definitions.

14. Figure 1 shows how terms, concepts, objects, and definitions are related. From the definitions above and Figure 1, several important observations need to be made:

- For any concept, there may be many designations (synonyms)
- For any concept, there are one or more objects in its extension
- For any concept, there may be more than one definition (especially in multiple languages)
- For each term, more than one concept may be designated (homographs)

15. Concepts are human constructions (Lakoff, 2002). No matter how well we define a concept, a complete description is often impossible. Identifying the relevant characteristics is culturally dependent. So, some objects in the extension of a concept, called prototypes, fit the

characteristics better than others (Lakoff, 2002). For example, a robin fits more of the characteristics of a bird than a penguin does.

B. Relationship to Data

16. Statisticians view a datum as a value representing a class in a partition of a population of objects, where the partition⁴ is defined for some characteristic of the population (Froeschl, Grossmann, & Del Vecchio, 2003). Here, we treat the population as a concept. Data are collected on the set of objects, the extension of the population, by measuring some characteristics of the population. For a given characteristic, the corresponding property for an object is assigned a value corresponding to one of the allowed classes in the partition.

17. In the finite case, usually for categorical data, the partition is often called a classification, e.g., sex categories. In the infinite or unbounded cases, usually quantitative data, the partition may not have a finite number of classes. Instead, the values denoting the classes come from a range of values.

18. In any case, the classes in the partition are concepts. In the sex classification example in the preceding paragraph, the classes are male and female. In the case of a range, e.g., all real numbers between 0 and 1, then we might say each value represents a probability. Because each class is a concept, then the value representing the class is a designation, in the terminological sense.

19. Therefore, a datum is a designation (Farance and Gillman, 2006). The concept associated with the designation is, at least, a combination of the population, the characteristic (as a concept) under study, and a class within the partition.

III. ISO/IEC 11179

A. Overview

20. The ISO⁵/IEC⁶ 11179 - *Metadata registries* - standard is a metadata specification devoted to data semantics. It also contains a model and an overview of a procedure for registration, hence the "registries" in the name. However, the main focus is the semantics of data.

21. The standard is divided into six parts, each of which describe an aspect of the standard. A short description of each part follows:

- Part 1 - *Framework* -- an overview of the standard and the methodology behind data semantics
- Part - *Classification* -- presentation of a model for managing a classification scheme, especially as it relates data elements (variables) to each other
- Part 3 - *Metamodel and basic attributes* -- presentation of the full model for data semantics, classification, and registration
- Part 4 - *Formulation of data definitions* -- principles for writing good data definitions
- Part 5 - *Principles for naming and identification* -- provides a naming convention for each of the principal parts of data semantics
- Part 6 - *Registration* -- procedures for registration

⁴ A partition is a non-empty set of mutually exclusive and exhaustive subsets of some other set. The number of subsets is not necessarily finite.

⁵ International Organization for Standardization

⁶ International Electrotechnical Commission

22. The last published version of the standard is the 2nd edition, completed in 2005. All the latest published parts of ISO/IEC 11179 are freely available on the web⁷. The 1st edition of the standard, published in 2000, was superseded by the 2nd. It was called *Standardization and specification of data elements*. The change in focus away from just data elements in the 1st edition necessitated the change.

23. The basic unit for describing data in ISO/IEC 11179 is the data element (variable). The model specified in the standard shows how one should describe a data element. It is concept based and follows the general framework of the terminological theory of data described above.

24. However, the standard does not address statistical data *per se*. It contains a general description of data, and does not go any further than that. Even the idea of a data set is not described in the standard.

B. Implementing Terminology Theory

25. The ISO/IEC 11179 standard implements the terminological theory of data in a very straightforward way. For each of the constructs in the theory, there is one in the standard (ISO, 2005)

26. Without going into details about the model described in the standard and defining all the terms there, we list the mapping between the terms in the standard and the terms from the terminological theory. For more details, see ISO/IEC 11179-1: *Framework* (ISO, 2005).

27. Here is the list, with the terms mapped:

Terminology	ISO/IEC 11179
Concept (population)	Object Class
Characteristic	Property ⁸
Partition	Conceptual Domain
Classes	Value Meanings
Designation (values)	Permissible Values

28. As stated before, the main data construct in ISO/IEC 11179 is the data element. Data elements may be abstract or implemented in some information system. Either way, a data element is a container of data (imagined or actual) where each datum has the same semantics with the possible exception that the value meaning may differ. This corresponds to the situation described in Example 1 and Example 2 below. Example 2 describes a data element with only one value.

29. Two other important ISO/IEC 11179 constructs are conceptual domain and value domain. A conceptual domain is a set of value meanings. A value domain is a set of permissible values. Since a permissible value is actually a pair consisting of a value (designation) and its value meaning (concept), the conceptual domain and value domain are closely related. If one sex codes value domain contains these permissible values

<M, male>

<F, female>

and another contains these permissible values instead

⁷ Information Technology Task Force (ITTF) under ISO and IEC
(http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/ITTF.htm).

⁸ This was a most unfortunate choice. The term will be changed to characteristic in the next (3rd) edition.

<0, male>
<1, female>

then the two value domains are linked through the value meanings (male, female), which are shared.

C. Statistical Data

30. There are two senses in which the term object class is used. First of all, an object class is a concept represented by a definition or description. Also, it is the extension of the concept, the set of objects about which we collect or observe data. Both senses are used, and the context is intended to identify which.

31. Now, in a way that departs from traditional statistics theory a little, we say the object class may be either a general or individual concept. Two examples illustrate the ideas:

Example 1: Data element with object class as general concept (microdata)

Object class:	Adults age 16 and older in Switzerland
Property:	Sex
Value meanings:	{Male, Female}
Permissible values:	0 for Male 1 for Female

Example 2: Data element with object class as individual concept (macrodata)

Object class:	The set of adults age 16 and older in Switzerland
Property:	Proportion of females
Value meanings:	{ $x \mid 0 \leq x \leq 1$ }
Permissible values:	Real numbers between 0 and 1, with precision to 3 decimal places

32. In the case of microdata, object classes are general concepts - see Example 1. In the example, the object class is all adults age 16 or older in Switzerland. Since there are many such people (more than one), this is a general concept. However, aggregate data, or macrodata, requires an object class with one object. In Example 2, the property⁹ "proportion of females" applies to the set containing all adults age 16 or older in Switzerland, not to each of the elements, i.e., each individual. The set consisting of "all adults age 16 or older in Switzerland" is a single thing. It is this aggregate that has the property "proportion of females", not the individual people. "Proportion of females" is not a property of people, "sex" is. Likewise, "sex" is a property of people, and it is not a property of the aggregate, "proportion of females" is. The point is the object classes in the two examples have different intensions, because they are different concepts, even though they are closely related. This does not mean they cannot have some properties in common, but they must have some different ones.

33. There is an exception to the rule that the object class for macrodata is an individual concept. This is the situation where the object class describes multiple instances of the same aggregate, rather than describing a single one. This arises in time series and tables, where an aggregate is reused over time or over multiple specializations. See the discussion and example in the next section.

⁹The term property is used in this paragraph and the rest of section 3 in the ISO/IEC 11179 sense, i.e., a characteristic of a concept.

D. Tables and Time Series

34. Tables are used to present cross-tabulated and time series data in an easy to read format. As the number of dimensions in a cross-tabulation goes up, so does the number of cells. It is tempting to describe each cell as a different data element; however this adds a huge burden to those responsible for recording the metadata. An easier method and equally as effective for describing tables is illustrated next.

35. Consider the simple Table 1 given below. It is a cross-tabulation of sex by age by population for all grizzly bears¹⁰ in Jellystone Park¹¹.

	Population
Male	105
Age 0 - 16	60
Age 17 - 32	45
Female	95
Age 0 - 16	50
Age 17 - 32	45
Total	200

Table 1: Sex by Age by Population for Grizzly Bears in Jellystone Park

36. Table 1 is described using 3 data elements, specified as follows:

Data Element Name	ISO/IEC 11179 Constructs	Values
Sex of Bears, codes	Object class	Grizzly bears in Jellystone Park
	Property	Sex
	Value domain	<M, male> <F, female>
Age of Bears, categories of years	Object class	Grizzly bears in Jellystone Park
	Property	Age (years)
	Conceptual domain	<1, age 0 - 16> <2, age 17 - 32>
Jellystone Park bear population, counts	Object class	Grizzly bears in Jellystone Park (aggregate)
	Property	Cross-tabulated population
	Value domain	<non-negative integers; counts>

37. These 3 data elements completely describe the semantics of Table 1. All the cells are understandable from the semantics of the three data elements. For instance, to understand the cell labeled by "population of female grizzly bears age 0 to 16 in Jellystone Park", one must look at the semantics of the 'age' and 'sex' data elements to understand the conceptual domains

¹⁰ We assume the average life span of a grizzly bear is 32 years. The data are made up.

¹¹ Jellystone Park is an invention of the Hanna-Barbera cartoon syndicate. The cartoon character Yogi Bear lived in Jellystone Park.

used to classify the cell. Finally, one looks at the 'counts' data element to understand what the counts mean.

38. Notice, that the object class for the third data element called "Jellystone Park bear populations, counts" is an aggregate for all the bears, but it is a general concept. Each of the cells in the table corresponds to a different Jellystone Park grizzly bear population. They are specialized populations (e.g., only females age 0 - 16), however, as stated above, one finds the semantics of the cell to understand the modification.

39. There isn't an automatic rule for deciding when an object class is a general concept versus an individual concept. It depends on use, which should be reflected in subtle differences in the properties of the concepts. When data are aggregated, the object class is clearly an individual concept. This is because a single object is under consideration. However, the aggregates in Table 1, considered as a collection, are described by an object class without recourse to some specializations, i.e., "Grizzly bears in Jellystone Park". Each object, the cells, are not described individually, they are described collectively. One could define an object class for each cell, e.g., "Female grizzly bears aged 0 - 16 in Jellystone Park" corresponds to the cell " females age 0 - 16". Then, this object class is an individual concept.

40. The same holds true for time series. There, the specialization is usually based on time. So, if we estimate the population of grizzly bears in Jellytone Park each year, then an object class that is an individual concept must include a time property, e.g., " Grizzly bears in Jellystone Park in 2006".

E. Metadata

41. When one conjures a particular object in the mind, the conception of that object is an individual concept. This is because there can be only one object in its extension, the conjured object. This means every object has an individual concept associated with it. Data associated with a particular object is descriptive of that object, and this means those data are metadata. Data are only metadata when they are used to describe some object.

42. This implies all data are metadata at the point of collection!

F. Registration

43. Registration is the set of rules, operations, and procedures that apply to a metadata registry. The three most important outcomes of registration are the ability to monitor the provenance (the source of the metadata), quality of metadata, and assigning an identifier to each object described.

44. Registration also requires a set of procedures for managing a registry. The rules cover submitting metadata for registration of objects and maintaining subject matter responsibility for metadata already submitted. For actual implementations of a metadata registry, additional requirements may be necessary.

45. Provenance refers to the source of the metadata. Registration handles this in several ways:

- Naming the subject matter specialist responsible for the content of a registered item
- Naming the organization (or person) who submitted the metadata for registration
- Maintaining identifiers and version numbers

46. There are several purposes to monitoring metadata quality. The main purposes are as follows:

- Monitoring adherence to rules for providing metadata
- Monitoring adherence to rules for forming definitions and following naming conventions
- Determining whether a description still has relevance
- Determining the similarity of related data constructs and harmonizing their differences
- Determining whether it is possible to ever get higher quality metadata for some data constructs

47. Every data construct registered in a metadata registry is assigned a unique identifier. Identifiers are a means to keep track of descriptions for administration purposes, to refer to descriptions by remote users of the registry, and aid in metadata transfer between registries.

48. The registration authority is the organization responsible for setting the procedures, administering, and maintaining a registry. The submitting organization is responsible for requesting that a new description be registered in the registry. The steward is responsible for the subject matter content of each registered item. Each of these roles is described in ISO (2005).

G. Application for documenting data elements in statistical agencies

i. Introduction

49. Both the Bureau of Labor Statistics and Statistics Canada are using ISO/IEC 11179 to document data elements, or variables. The work is in different stages of development, but many of the experiences are similar. In this section, we will discuss some of the practical considerations in using the standard in statistical offices.

50. To document variables, the object class, the property, and the value domain are specified, named, and defined. Each of these components can stand on its own and is reusable in the construction of other data elements. A general strategy can be adopted therefore to be very economical in the creation of these constituent parts and to use combinations and permutations of these elementary components to represent the diversity of variables for which data are published by statistical agencies.

51. All of the building blocks, and the links between them in the context of given statistical datasets or surveys, can be stored in a metadata repository. From there it is possible to produce, dynamically and on request, the complete definition of every variable, according to the specifications of the standard.

52. Consider the data element named "type of expenses of business location" as an example. It is analyzed as follows:

- "Expenses" refer to decreases in economic benefits or service potential, during the reporting period, in the form of outflows or consumption of assets or incurrence of liabilities that result in decreases in equity, other than those relating to distributions to owners. "Expenses" is the name of a property.
- "Business location" refers to a statistical unit defined as a producing unit at a single geographical location from which economic activity is conducted and for which, at a minimum, employment data are available. "Business location" is the name of an object class.
- "Type" refers to the reporting of "Expenses of Business Location" using the classification called Expense Categories, Annual Survey of Manufactures (ASM). "Type" is a representation term, naming the kind of value domain.

53. Most data element documentation can be approached in this straightforward manner but some situations arise where conventions or consistent approaches need to be developed and applied to deal with the varied ways used by data producers to present statistical data. Some practical choices need to be made in applying the standard and conventions adopted to deal with some of the more common situations. Suggestions for common applications are presented for each major construct in the standard: object class, property, and value domains.

ii. Object Class (Statistical Unit)

54. Statistical units are defined in Statistics Canada's Policy on Standards as: "The unit of observation or measurement for which data are collected or derived".¹² With reference to the standard, this makes it a statistically relevant type of object class, defined as "a set of ideas, abstractions, or things in the real world that can be identified with explicit boundaries and meaning whose properties and behavior follow the same rules." In applying the standard, it is desirable to try to limit the number of object classes by using fundamental statistical units, if possible. Fundamental statistical units are defined as those that are not types of any other statistical unit and cannot be derived as grouping of any other statistical unit.¹³ The following types of fundamental statistical units were identified:

- Agents: Entities that act and whose actions are reported on by statistical agencies. In social statistics, "Person" can be considered as an agent.
- Events: Actions of (or by) agents as reported by statistical agencies. Events are discrete in time (occur in a time period) and finite (can be counted). In social statistics "Birth" can be considered an event.
- Items: Things that are generally either produced or managed by agents. In economic statistics, "Product" can be considered as an item.

55. Fundamental statistical units are identified as a means of keeping the number of object classes to a minimum. However, some statistical units that can be derived in some way from the fundamental statistical units are so commonly used that they should be identified as separate object classes. Common derivations of the fundamental statistical unit include:

- Subsets of fundamental statistical units based on an inherent characteristic. An example of this is "Person age 15 and over" used as an object class. This is a subset of the "Person" object class based on the Age property. In this way, data elements such as "Type of Occupation of Person age 15 and over" can be defined.
- Subsets based on roles that the statistical units may assume. Examples of this are "Student", "Mother" and "Employee", subsets of the "Person" object class based on various role properties. In this way, data elements such as Category of Major field of study of Student can be defined. Subsets based on roles differ from those based on inherent characteristics in that the same statistical unit can take on more than one role at the same time. It can both assume and discontinue a role over time.
- Supersets of fundamental statistical units. For example, "family" is a group of persons according to certain grouping rules.

56. Starting with fundamental statistical units and only identifying additional statistical units defined according to certain properties when these are commonly encountered, national statistical agencies can represent their data holdings with a fairly limited set of about 80 object classes. These are shown in the following table.

¹² Statistics Canada, Policy on Standards, <http://www.statcan.ca/english/about/policy/standards.htm>

¹³ See Mechanda, K., Johanis, P., and Webber M. (2003) *Conceptual Model for the Definitional Metadata of a Statistical Agency*, Paper for Open Forum 2003 on Metadata Registries, Santa Fe, New Mexico.

Table 2 – Object classes defined in Statistics Canada’s metadata registry

	Social	Economic
Agents	Family Child Crime Victim Criminal Accused Criminal Charged Emigrant Homicide victim Household Household Head Immigrant International Migrant Interprovincial Migrant Intraprovincial Migrant Mother Non-permanent Resident Person Person 12 Years or over Person charged Smoker Student Suspect - Chargeable Woman	Business Entity Business Location Earner Economy Employed Person Employee (LFS) Employee – private Employee – public Employee (SEPH) Employment Insurance Beneficiary Enterprise Establishment Farm Operation Institutional Unit Labour Force Participant Paid Worker (labour income) Paid Worker (labour market) Person 15 Years or over Self-Employed Worker Traveller Unemployed Person
Events	Birth Community Admission Correctional Service Admission Criminal Incident Custodial Admission Death Divorce Homicide Marriage	Person-trip Person-visit
Things	Case Charge Criminal Offence Dwelling Legal Aid Application Legal Aid Plan Probation Order Sentence Shelter	Building Permit Crop Employment Insurance Claim Farm Input Help Wanted Ad Job Passenger-Kilometer Product Security Transaction Vehicle Vehicle –Kilometer Visit-night

57. Under the standard, almost anything can be an object class. In practice, certain choices must be made. A test that can be used in trying to isolate the object class is to answer the question “What is being counted here?” For example, we may be tempted to identify “industry” as an object class. However, statistical agencies do not report meaningful statistics on the number of industries (i.e. in 2004, there were 212 active industries in Canada). Rather, we generally report on the number and size of businesses, classified by industry.

iii. Properties

58. Properties are simply the characteristics of interest of the unit of observation (object class). These include sex, income, industry type, and number of employees, among many others. Having defined the object class, it is relatively straightforward to identify which characteristics are being measured.

59. The application of the standard leads us to sometimes define “compound” properties. Limiting object classes to fundamental statistical units, only occasionally further qualified by a property, shifts more of the meaning to the property and value domain definitional space. As a result, it may be necessary in some cases to define a property with more than one dimension. For example, in the data element “number of production workers’ hours paid of business location”, the property is “production workers’ hours paid”. “Production workers” and “hours paid” can be considered object classes each on their own. They can be counted, they have similar characteristics, etc. However, in this dataset, the “business location” is the unit of observation and “hours paid” and “production workers” have been combined to form a compound property of the business location.

60. Another example where this is the case is the data element concept named “race of reference person of household”, where “reference person” refers to the person in the household used to define all the relationships between members. Adding to the confusion, the US uses such data to apply “race” to the household in order to perform imputations for missing data. So, “reference person” has a subservient role in the semantics. Applying the question “what is being counted here?”, the answer is not immediately clear. The object class could be “consumer unit” or “reference person of consumer unit”, and the property contains the remaining semantics in each case: “race of reference person” or “race”. The question reduces to whether the meaning refers to the data at data collection or during the analytical stage. Applying the list of available object classes (see table 2) solves the problem.

61. Based on the experience of Statistics Canada, it appears as if the specification, naming, and definition of about 500 properties will be sufficient to cover all statistical data published by a national statistical agency.

iv. Value domains

62. Value domains come in two types: enumerated and non-enumerated. They are the set of allowed values a data element may take. In ISO/IEC 11179, a value domain is defined as a set of permissible values, where a permissible value is a pair containing a value and its meaning. The set of value meanings is called a conceptual domain. Conceptual domains also come in the two types: enumerated and non-enumerated. A simple example of an enumerated value domain is “sex codes”, which may contain the following permissible values: See ISO (2003) for a more complete discussion of value domains and conceptual domains.

<M, male>
<F, female>

63. The corresponding conceptual domain is the set of value meanings:
{male, female}

64. The enumerated and non-enumerated types correspond usually to the standard statistical datatypes of categorical and quantitative data, respectively. There are exceptions, but mostly these have to be manufactured. The natural way to use the types of value domains lends itself easily to the statistical datatypes.

65. In particular, non-enumerated value domains are used to represent continuous variables, variables that can assume any numerical value within a range. For example, the non-enumerated value domain for the data element "value of income of person" might be the set of integers between 0 and infinity. All that is needed to understand and interpret such a value is to know the unit of measure (i.e. Canadian dollars), and the precision (for example, two decimal places). So, the value meaning for each value in a non-enumerated value domain is the unit of measure.

66. List of units of measures recorded in Statistics Canada's metadata registry are

Area in Acres
 Area in Square Feet
 Areas in Hectares
 Basic Price in Current Dollars
 Canadian Dollars
 Chained 1997 Dollars
 Constant Dollars
 Count in Dozens
 Count in Metric Bundles
 Count in Metric Rolls
 Count in pairs
 Counts in Whole Numbers
 Current Prices
 Quantity in Megawatt hours
 Set of Numbers Expressed as Indexes
 Set of Numbers Expressed as Rates
 Set of Numbers Expressed as Ratio

Time in Days
 Time in Hours
 Time in Years
 Volume in Bushels
 Volume in Gallons
 Volume in Kilolitres
 Volume in Litres
 Volume in Quarts
 Volume in Tonne-kilometres
 Volumes in Cubic Metre-kilometres
 Volumes in Cubic Metres
 Volumes in Cubic Metres Dry
 Weight in Hundredweights
 Weight in Kilograms
 Weight in Metric Tonnes
 Weight in Pounds
 Weight in US tons

67. Rather more information is required, however, to interpret values taken from an enumerated value domain. An enumerated value domain is a set of categories, represented by codes or labels, or both, each having a meaning unrelated to its actual value. The number 325410 could be the (slightly out of date) population of Victoria, BC in Canada. As a code, it is actually a NAICS code meaning Pharmaceutical and Medicine Manufacturing. It is impossible to know this without reference to metadata. In the standard therefore, enumerated value domains are made up of pairs: values (codes) and value meanings (labels), which can also have a definition.

68. Value domains can also be considered standalone building blocks, which can be associated with appropriate data elements as required. Managing, registering and maintaining these value domains is in fact a common task of national statistical agencies, where they usually take the form of statistical classifications. In this application of the standard, statistical classifications are re-created from value domains specified according to the standard. As a “classification” entity does not exist in the standard, a number of conventions are therefore developed for this purpose.¹⁴

69. First, every value domain is given a top value domain, containing only one permissible value and value meaning, which is the parent of all subordinate permissible values. This value domain is a place holder or organizational device designed to be the container for the classification of interest. This value domain is given the name of the classification that it is intended to represent (for example, NAICS Canada 2002). Under this value domain, one or more value domains are hierarchically identified. Each value domain is a set of permissible values, with associated value meanings, which are mutually exclusive and exhaustive of the universe of observations to be classified. Each value domain is assigned a level within the hierarchy. Every permissible value is assigned to a parent permissible value from a higher level value domain and its order among siblings is recorded. Each permissible value can be the child of one and only one parent permissible value and is thus exclusive in aggregation. With these conventions, the full structure of any classification can be reconstructed.

70. In certain cases, classifications are altered by data producers by grouping certain classes together. This in effect introduces a new value domain, or a new level in a classification, that is not exhaustive of the universe of observations to be classified.

Table 3: A Value Domain for Current Account

Current Account									
Goods	Services				Investment Income			Current Transfers	
Goods	Travel	Transportation	Commercial Services	Government Services	Direct	Portfolio	Other	Private	Official

¹⁴ Part 2 of ISO/IEC 11179 deals with classification, but this relates to the classification of data elements and their constituent building blocks in a metadata registry for ease of organization and search, which will be covered in section 9 of this paper, not the classification of observations in an enumerated value domain, which is the issue here.

71. Table 3 provides an example of a classification comprised of three value domains (levels), in this case the Current Accounts Classification¹⁵ used in Canada. Table 4 shows how the classification has been altered by a data producer by grouping together “Goods and Services” on the second level and Commercial services and Government services on the third level.

Table 4: Alternative Value Domain for Current Account

Current Account (with incomplete levels)								
Goods and Services								
Goods	Services			Investment Income			Current Transfers	
			Other Services					
Goods	Travel	Transportation	Commercial Services Government Services	Direct	Portfolio	Other	Private	Official

72. This has in effect introduced value domains that are not exhaustive of the universe at levels 2 and 4 of the altered classification. If the classification is presented one level at a time to users, which could well happen in many applications, the user will have incomplete information concerning all the values that the data element being represented by this value domain could assume. To correct for this, every level is made to be exhaustive of the universe of observations to be classified by “promoting” classes from the level below (see arrows in Table 4). This results in a new 5-level classification, as shown in Table 5.

Table 5: A Corrected Alternative Value Domain for Current Account

Current Account (Incomplete levels filled in)								
Goods and Services				Investment Income			Current Transfers	
Goods	Services			Investment Income			Current Transfers	
Goods	Travel	Transportation	Other Services	Direct	Portfolio	Other	Private	Official
Goods	Travel	Transportation	Commercial Services Government Services	Direct	Portfolio	Other	Private	Official

¹⁵ This example is taken from Johannis, P., Brooks B., Dunstan T., and Lévesque, J-S. (2003), Statistics Canada’s Implementation of the Data Element Model, paper for the Metadata Registries Open Forum 2003.

73. The outcome is a well configured classification, rectangular, exhaustive at every level, with classes that are mutually exclusive and exclusive in aggregation. Another advantage of this approach is that it preserves the relationship between the original classification and its variants. This promotes the reuse of standard value domains (in the example above, levels 1, 3, and 5 in the variant are identical to levels 1, 2, and 3 in the original) and clearly shows how one relates to the other.

74. Original classifications, which might be standard classifications (and recorded as such in the registration status – see Administration and identification region of the standard), and their variants are treated this way consistently in the IMDB. The original classification is considered an “umbrella” value domain and is flagged as such in the metadata registry. In this way, potential targets for future harmonization are easily identified.

75. Some value domains have a large number of variants. For example, Statistics Canada currently publishes data according to 13 different variants of the North American Industry Classification.

North American Industry Classification System Canada 1997 (standard)

NAICS 1997 Durable / Non-Durable Manufacturing Industries

NAICS 1997 Energy Sector

NAICS 1997 GDP

NAICS 1997 GDP Finance, Insurance, and Real Estate

NAICS 1997 GDP Special Industry Aggregations

NAICS 1997 Goods and Services

NAICS 1997 ICT

NAICS 1997 ICT (Manufacturing/Services Split)

NAICS 1997 Industrial Production (GDP)

NAICS 1997 IOFD

NAICS 1997 Labour Income

NAICS 1997 LFS

NAICS 1997 Trade Groups

76. Some value domains also change over time, but these are considered as versions rather than as variants. For example, NAICS Canada 1997 was the standard classification for type of industry. It had many variants under the same umbrella. When the original was replaced by NAICS Canada 2002, this was considered a new version of the same value domain. Similarly, any variants of NAICS Canada 1997 that were updated under the 2002 version were considered new versions of these variants. There are also other classifications used for type of industry, for example the International Standard Industry Classification (ISIC) and the European industry classification (NACE). These are all related and this is represented in the model by having the value domains making up all of these related classifications use value meanings drawn from a common pool of value meanings, which is the industry conceptual domain. Conceptual domains are containers of value meanings, which are re-used in value domains.

77. This application of ISO/IEC 11179 to document statistical classifications is consistent with the approach developed by the Neuchâtel group (see section V.B) but has not been reconciled in detail. Uncovering the meta-models underlying the various “classification servers” in use around the world and developing a common approach to documenting

classifications, based on this approach or another, would be a major, and attainable, achievement for the world statistical community.

v. Naming

78. For every item we care about the meaning, we give it a name. Names are the means by which people remember things and first infer some meaning. The ISO/IEC 11179 devotes an entire part of the standard to naming: Part 5.

79. The naming convention for data elements provided in the standard is quite simple. Data elements names consist of main parts:

- Object class term
- Property term
- Representation term

80. The first two of these are understandable from the previous sections. The representation term refers to the value domain and other attributes associated with the representation of the data element. Here "representation" includes the allowed values, their datatype, and the unit of measure (if necessary). Representation is the "form" of the data as they appear on paper or the screen.

81. Terms such as count, value, and number are used as representation types in the case of data elements with non-enumerated value domains. "Value of expenses of business location" is an example of such a data element name, using the representation type "value". In the case of data elements with enumerated value domains, representation types such as name, type, and category have been used, as in the data element name "category of age of person". In the end, a relatively small set of representation types is sufficient to cover all of the statistical output of a statistical agency (see list below).

Table 6 – Representation classes defined in Statistics Canada's metadata registry

Enumerated	Non-enumerated
Category	Amount
Code	Area
Level	Average
Name	Duration
Status	Index
Type	Length
	Mean
	Number
	Percentage
	Proportion
	Quantity
	Range
	Rate
	Ratio
	Value
	Volume
	Weight

vi. Data Elements

82. With these elementary building blocks defined, it is possible through combinations of object classes, properties, and value domains to specify, name, and define all data elements produced by a national statistical agency. For Statistics Canada, this approach has resulted in the identification of around 900 data elements, covering the entirety of its statistical output disseminated through CANSIM.¹⁶ This is where the consistent application of ISO/IEC 11179 could yield major harmonization gains across national statistical agencies.

83. The list of object classes and representation types presented in this paper is almost surely applicable to all national statistical agencies. Harmonization efforts would therefore concentrate on common names and definitions of properties, which is an achievable goal. This would allow users at least to locate common data elements across national statistical agency data holdings and be secure in knowing that underlying definitions are the same.

84. The problem of interoperability, unfortunately, is not that simple. There are considerable disharmonies in the value domains, or classifications, used to represent these data elements. So, work to harmonize classifications (and value domains in general) across statistical agencies is required to make data interoperable.

IV. CORPORATE METADATA REPOSITORY (CMR) MODEL

85. This section contains a partial description of the CMR model. Each of the sub-sections contains a model represented in the Unified Modeling Language (UML). These models are less detailed versions of those found in the CMR model. The attributes depicted in each class are meant to signify greater detail in the actual model.

86. The model presented is a conceptual model. Datatypes are not provided with the attributes, and there are not enough details to insure consistent implementations. Nevertheless, the US Census Bureau, Bureau of Labor Statistics, and Statistics Canada each have implementations of this model.

A. Data Dimension

87. The data dimension describes the semantics of data. It is a copy of part the model specified in ISO/IEC 11179. All the provisions needed for consistent implementation are described in the standard. If one were to implement the standard as part of a CMR implementation, the result would be a conforming¹⁷ implementation of ISO/IEC 11179 (ISO, 2005).

88. The following diagram is a high level overview of the data model. The classes in the model are described in section III.

¹⁶ This understates the actual count slightly as there are a few data elements for which metadata are still in a staging area and have not been loaded into the IMDB.

¹⁷ The term conforming is defined in ISO/IEC 11179-3: Part 3 - *Metamodel and basic attributes*.

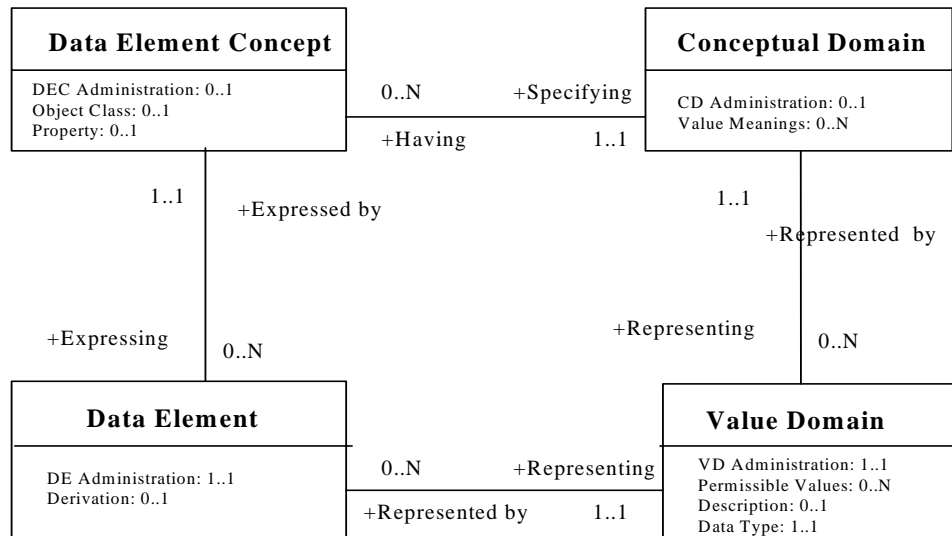


Figure 2: Data Dimension Model

B. Business Dimension

89. The Business Dimension describes the business of statistical organizations. It is composed of classes, attributes, and relationships that describe data that the organization needs to keep about surveys. The model supports the storage of metadata as single attributes or as documents. Figure 3 shows the Business Dimension model.

90. The model describes survey designs, processing, analyses, and data sets. It contains classes for each of the important parts of a survey. The model supports organized storage and complex searches for metadata describing a survey, and it supports searches for metadata across multiple surveys. The model also provides several other features:

- A list of all current surveys conducted by the agency
- Comparison of designs, specifications, or procedures across surveys
- Reuse of designs, specifications, or procedures
- Categorizing and classifying documents
- Assembling complete documentation for a survey
- Attributes to support embedded metadata

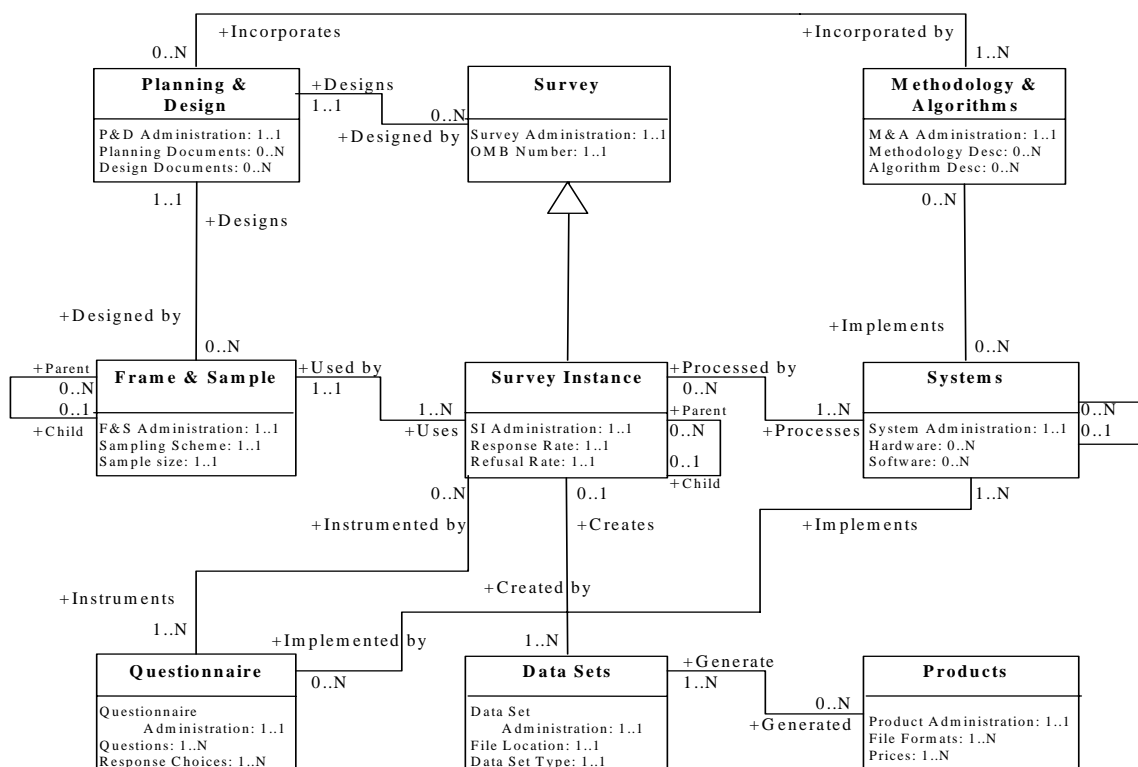


Figure 3: Business Dimension Model

i. Survey Life Cycle

91. The model supports the survey life cycle. Content, Planning, and Design are captured in the Planning and Design, Frame and Sample, Methodology and Algorithms, and Survey classes. Collection is captured in the Questionnaire, Survey, Survey Instance, Data Set, and Systems classes. Processing and Analysis are captured in the Survey Instance, Methodology and Algorithms, Systems, and Data Set classes. Finally, Dissemination is captured in the Data Set, Systems, and Product classes.

ii. Questionnaire Model, Linking Business – Data Dimensions

92. The CMR model contains many links between the various dimensions within the model. Using the detailed questionnaire model, Figure 4 illustrates links between questionnaires and data elements. Data Element Concepts and Questions are linked, because they each express concepts describing the same data, albeit from a different perspective. Value Domains and Response Choices each describe the valid values some data can take.

C. Administration and Document Dimensions

93. The Registration Authority (ISO, 2005) establishes the rules under which the repository operates. Monitoring metadata quality, monitoring the life cycle of the described objects, and maintaining paths of accountability for metadata are important functions.

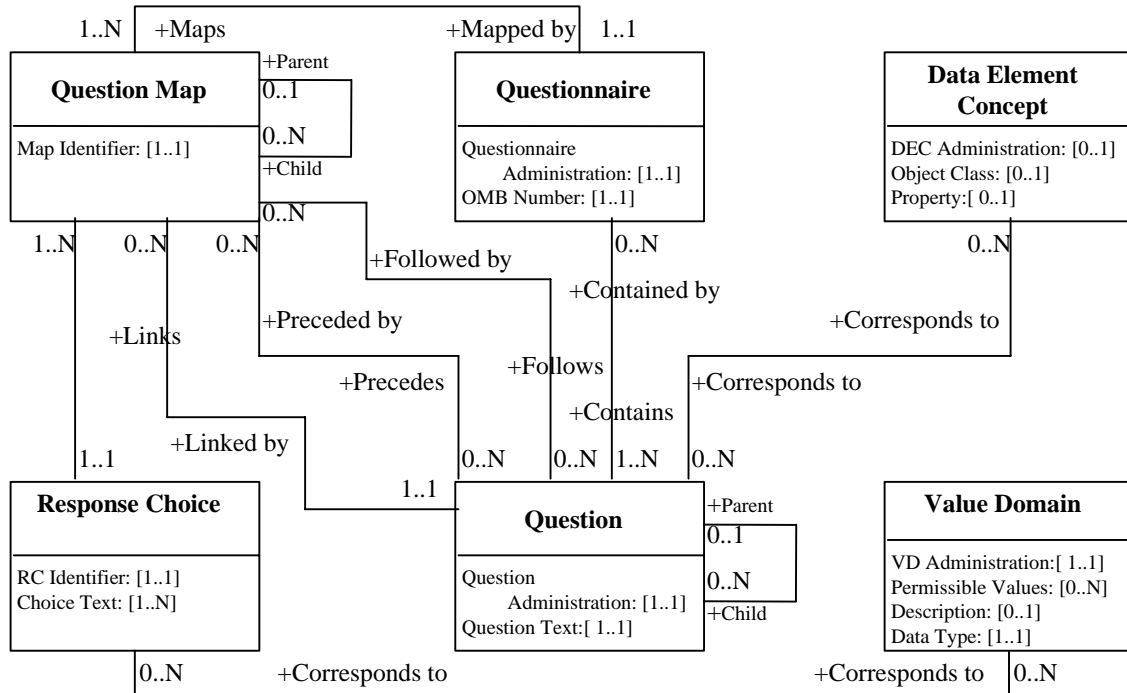


Figure 4: Questionnaire Model

94. An Administration Record is established each time an object is described, or *registered*. The common attributes are provided along with specialized attributes for each object. Metadata is often provided in the form of documents, so URL's to relevant documents are critical metadata. The model allows links to as many documents as necessary. Each document may be linked to many objects. Figure 5 provides a data model for the registration process.

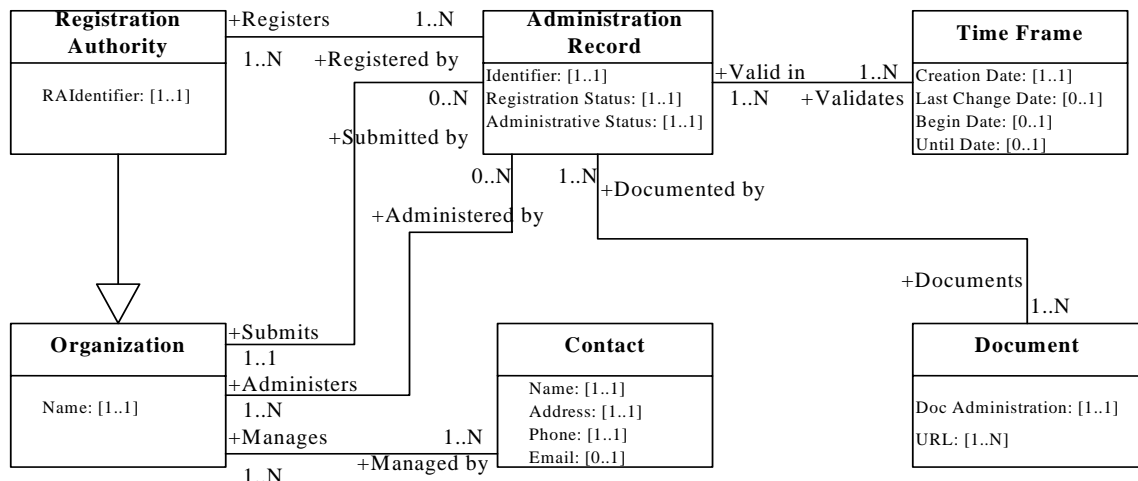


Figure 5: Administration and Documents Dimensions Model

D. Terminology and Classification Dimensions Model

95. The CMR model contains a dimension for managing classification schemes used to classify objects the CMR model describes. The term "classification scheme" is slightly

misleading. The model supports representing a concept system. Classification is achieved by relating objects (administration records) to other concepts. The classification of objects ties them to particular subject fields, such as the concept system determined by the variables of interest for a survey. In other words, the concepts a survey is trying to measure make up part of the subject field for that survey. For instance, assigning an object class and property, as described in section III.C, is a form of classification.

96. The CMR model also supports terminology management as it applies to the concepts, terms, and characteristics that describe registered objects. One of the most important metadata elements for any registered object is its definition. This is crucial for understanding the meaning of the object. The CMR model supports *semantics*. Terminology management is a fundamental part that aim. Figure 6 provides the terminology and classifications model.

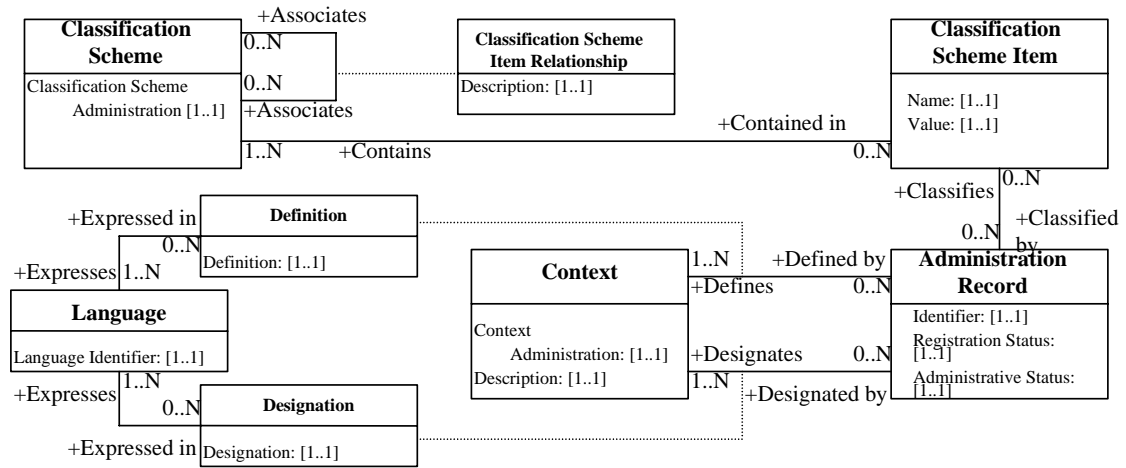


Figure 6: Terminology and Classification Dimensions Model

V. METADATA SCHEMES

97. In this section, five metadata schemes are described, and comparisons made between them.

A. eXtensible Business Reporting Language (XBRL)

98. XBRL (eXtensible Business Reporting Language) is an XML-based, royalty-free, open standard for business reporting. XBRL was developed by XBRL International, a not-for-profit consortium of over 400 leading companies and organisations around the world. Its main focus is to provide a standard format for recording financial data and associated metadata for financial statements and other business data reports. As these are frequently collected by statistical agencies as part of their economic statistics programme, XBRL provides a useful, and perhaps eventually universal, format for electronic data reporting for businesses. In this event, it would be very desirable if the metadata associated with the reported business data could automatically be converted from the XBRL format to the format used in the data repositories of statistical agencies.

99. In XBRL, the main organizing concept is the “instance document”, which represents one instance of some kind of financial report. In a statistical application, it is equivalent to a completed questionnaire or collection instrument. It therefore fits the CMR questionnaire model as described in figure 4.

100. Each instance document refers to a taxonomy, which describes the elements used in the document (for example, fixedAssets, totalAssets, subscribedCapital, and totalLiabilities), as well as the relationships among the elements. XBRL provides a structure to record the metadata for the concepts to be reported and cross-concept relationships, expressed according to the rules of XML syntax in the form of linkbases.

101. An XBRL taxonomy contains the following:

- **Schema:** A group of structured elements that may be used in instance documents. It is a dictionary of defined terms.
- **Label linkbase:** Labels or text associated with the elements in the dictionary may be created in different languages and used for different purposes.
- **Reference linkbase:** References to legal texts or accounting standards on which the concept is based.
- **Presentation linkbase:** Rules to specify the hierarchical relationships between elements.
- **Calculation linkbase:** Rules for calculations (additions and subtractions) between elements in the taxonomy.
- **Definition linkbase:** Rules to document other types of relationships between elements in the taxonomy.¹⁸

102. The challenge is to extract and re-interpret the metadata available in an XBRL taxonomy, that is the schema and all the referenced linkbases, according to the metamodel of ISO/IEC 11179/CMR. Fundamentally, this means identifying all the data elements and associated value domains that are embedded in a taxonomy.

103. The approach for doing this follows the procedure described in section III.G of this paper. What is the object class? For business reporting, it is always the enterprise. What are the properties of an enterprise being measured? First, there are identification properties such as geographic location, and industry. These can be defined in a straightforward way as per section III.G, i.e. name of geographic location of enterprise, type of industry of enterprise, etc. For the properties that describe the financial position and performance of the enterprise, however, we have different choices, depending on how economical we wish to be in identifying data elements. Every element in an XBRL taxonomy could be defined as a property, so we might define data elements such as value of fixed assets of enterprise, value of financial assets of enterprise, value of intangible assets of enterprise, value of total assets of enterprise. This would yield a very large number of data elements, many of which would by definition be interdependent (for example, total assets must be the sum of fixed, financial and intangible assets). An alternative is to consider each such element to be a permissible value within a value domain associated with a more aggregated data element, such as Type of asset of enterprise. This would yield two data elements, Type of asset of enterprise and Value of asset of enterprise. The former would have an enumerated value domain (with three values: fixed, financial, and intangible), the later would have an unremunerated value domain.

104. With this approach in mind, it is possible to consider the mapping between the metadata structure within an XBRL taxonomy and a metadata standard such as ISO/IEC 11179/CMR.

B. Neuchâtel Group

105. The Neuchâtel Group is currently composed of representatives from Statistics Netherlands, Statistics Norway, Statistics Sweden, Swiss Federal Statistical Office, and the

¹⁸ This section draws heavily on a White Paper on XBRL Concepts and Recommendations, published by the Technology Working Group of XBRL Spain, September 2005; available on xbrl.org

US Bureau of Labor Statistics. In addition, a small German software company, run-Software AG, is part of the group, and they build software to implement the specifications.

106. An earlier version of the group developed a model for managing classification systems used in statistical offices. This is called the Neuchâtel Group Classification model. It is in use by many statistical offices in Europe, North America, Australia, and New Zealand.

107. The Neuchâtel Group Variable Model is still under development. Obviously, the main construct described is the variable or data element. Many of the same components of data contained in ISO/IEC 11179 are contained in the Neuchâtel Group Variable Model. The first version of the specification is due to come out later this year. It is very much based on ISO/IEC 11179, but it has some major differences, too. First, it is developed using statistical terms and contains detail for statistical data that ISO/IEC 11179 lacks. Also, it describes databases, files, and formats in addition to basic variables.

108. The Variables Model does not contain the capability for registration, and it does not contain all the flexibility that ISO/IEC 11179 has. However, for the description of a single data element, they are very similar.

109. Since the specification is still in draft, the group does not want the document distributed at the time of writing this paper, so no reference to it is given. Also, because of limited resources, the group does not have a web site. This is also under development.

C. DDI

110. The Data Documentation Initiative (DDI) is an international project to establish an XML-based metadata standard for the content, presentation, transport, and preservation of documentation for datasets in the social sciences. Social scientists need to record and communicate all the important characteristics of the empirical data for which they are responsible in a straightforward way. The DDI endeavors to do this.

111. The DDI metadata specification originated in the Inter-university Consortium for Political and Social Research and is now the project of an alliance (<http://www.icpsr.umich.edu/DDI/org/index.html>) of about 25 institutions in North America and Europe. It is based on the idea of the electronic "codebook," retaining its capabilities, but growing the possibilities by improving the rigor and expanding the scope. The DDI is in use by many social science data archives and statistical offices around the world.

112. The DDI is represented as an XML DTD and an XML-Schema (W3C, 2004). The DDI-DTD is divided into 5 main chapters:

- Document Description - description of the XML document itself
- Study Description - description of the study behind the data
- File Description - physical layout of the described data set(s)
- Data Description - conceptual description of the data
- Other Material - descriptions of related data and documents

113. The main focuses of the DDI are the study and the data set from the perspective of social science statistics. The study is the only required chapter, and it represents a high level description of the data. This means that archives can maintain individual descriptions of each data set they manage. However, this also means that some of the metadata for a series of data sets from the same survey or program must be repeated.

114. The DDI has a rich set of elements devoted to the needs of statisticians and other users of statistical data. It is the only one of the five schemes that is specifically engineered for describing statistical surveys. The Neuchâtel Group, described above, also produces standards directly related to statistical offices, but they are much more specialized.

115. Under the variable description section of the data description chapter, there are some elements for capturing concepts, but there is little in the way of concept management. Part of this is due to the design. XML is hierarchical, and it is hard to model complex relationship structures. Revisions of the DDI are expected to address some of this.

116. The relationship between the ISO/IEC 11179/CMR standard and DDI has been frequently addressed in previous papers.¹⁹ More recently, the feasibility of automatically generating DDI compliant metadata sets from an ISO/IEC 11179/CMR repository has been examined, which concluded that it was possible to generate almost all the metadata required by the DDI standard in this way.²⁰

D. Statistical Data and Metadata Exchange (SDMX)

117. The SDMX project is jointly sponsored by seven international statistical and financial organizations: World Bank, Bank of International Settlements (BIS), International Monetary Fund (IMF), Organization of Economic Cooperation and Development (OECD), Eurostat, European Central Bank (ECB), and UN Statistics Division. Each organization has the need to describe, share, and transfer statistics and their metadata.

118. The project was begun in 2001, and now the 2nd version of the work is available at the project web site: <http://www.sdmx.org>. Also, the work is being developed as an international standard, ISO 17369.

119. The SDMX model is very sophisticated and general and, while it has been developed for the exchange of statistical data and metadata, the generality of the model could satisfy the needs of many kinds of businesses to transfer many kinds of data and metadata.

120. SDMX provides a metamodel for documenting and structuring data sets and associated metadatasets. Provision is made for datasets in the form of time series, cross-sectional tables or data cubes. From the data perspective, the main construct in SDMX is the *data structure definition* and related classes. Many of the components of data are described here. A *data structure definition* in SDMX corresponds to a dataset in the ISO/IEC 11179/CMR approach. Related attributes such as concepts and code lists in SDMX can be related to constructs in ISO/IEC 11179 such as data element concept, data elements and value domains. The detailed mapping between the two models has not been checked in a rigorous way, but there is a strong relationship between the standards.²¹

¹⁹ See Bradley, W.J., Colquhoun, G. and Ryssevik, J., **Integrating ISO/IEC 11179 and the DDI in a Single Application**, Presentation to Open Forum on Metadata Registries, Statistics Section, Santa Fe, January 2003; Bradley, W.J., Gillman, D.W., Johannis, P. and Ryssevik, J. **Is your Agency ISO Compliant? Standardizing Metadata for Improved Knowledge Delivery in National Information Systems**, Bulletin of the International Statistical Institute 54th Session Proceedings, Berlin, August 2003; Ryssevik, J., Glover T. and Colquhoun, G., **Relationship to ISO/IEC 11179**, Data Systems and Standards Division, Health Canada, 2003

²⁰ **Discovering Microdata Variables: Comparing DDI compliant documentation to an ISO/IEC 11179 metadata registry** Tim Dunstan, Statistics Canada, Chuck Humphrey, University of Alberta, Canada, Paper presented at Symposium 2005, Statistics Canada, October 2005.

²¹ For a first attempt at such a mapping, see SDMX, ISO 11179 and the CMR, Arofan Gregory and Chris Nelson, METIS 2006

121. Registration is also a part of SDMX. The developers followed the design of the ebXML registry specification defined by the Organization for the Advancement of Structured Information Standards (OASIS, 2002). This registry specification was generalized from the ISO/IEC 11179 registration idea and is conceptually similar.

E. Common Warehouse Metamodel (CWM)

122. The Common Warehouse Metamodel (CWM)²² is a standard developed under the auspices of the Object Management Group (OMG). Version 1.0 was published in February 2001. The CWM is among an integrated family of standards including Unified Modeling Language (UML, ISO/IEC 19501), Meta-Object Facility (MOF, ISO/IEC 19502), and XML Metadata Interchange (XMI, ISO/IEC 19503).

123. The CWM is a framework for integrating software and tools in the "information supply chain" (ISC). The ISC is a generalization of the survey production life-cycle in statistical offices, where the typical stages are conception, design, collection, processing, analysis, and dissemination.

124. In the ISC, parts of the business data process are linked together by data flows as data move between systems. Metadata are managed by each system independently, in proprietary ways, without regard to any interoperability. The question of how to make metadata sharable throughout the ISC is the purpose of the CWM.

125. The CWM contains an extensible model (a metamodel, in the terminology for OMG)) for modeling metadata. The common modeling framework allows then for mapping between metadata models. One then maps metadata across the ISC, using the metadata output of one process to help drive the next.

126. The modeling facility has 5 layers, called Object, Foundation, Resource, Analysis, and Management. The Object layer contains core modeling elements used to construct metadata models, all based on a special subset of UML. The Foundation layer defines services such as datatyping, business information, and key indexing to provide the means to define modeling constructs (e.g., datatypes, business mappings, and keys) needed to model any kinds of metadata. The Resource layer defines the kinds of data resources in the ISC, for instance relational, multidimensional, or XML. This allows descriptions, say, of relational databases. The Analysis layer provides additional models used to support analysis tools, such as data mining and information visualization. Finally, the Management layer provides for processes and operations that one might use to interact with a data warehouse.

127. The major benefit to the CWM is the possibility for a statistical office to tie together all the systems used for a survey or groups of surveys into one integrated system. Normally, metadata interchange is done between pairs of systems, and a mapping must be built for every pair. The CWM provides for a single point of communication, so each component of a system must map its metadata to a CWM implementation. The CWM then provides for all the mappings between metadata schemes. Therefore, the complexity of an architecture based on CWM is much simpler. There are fewer metadata mappings that must be made.

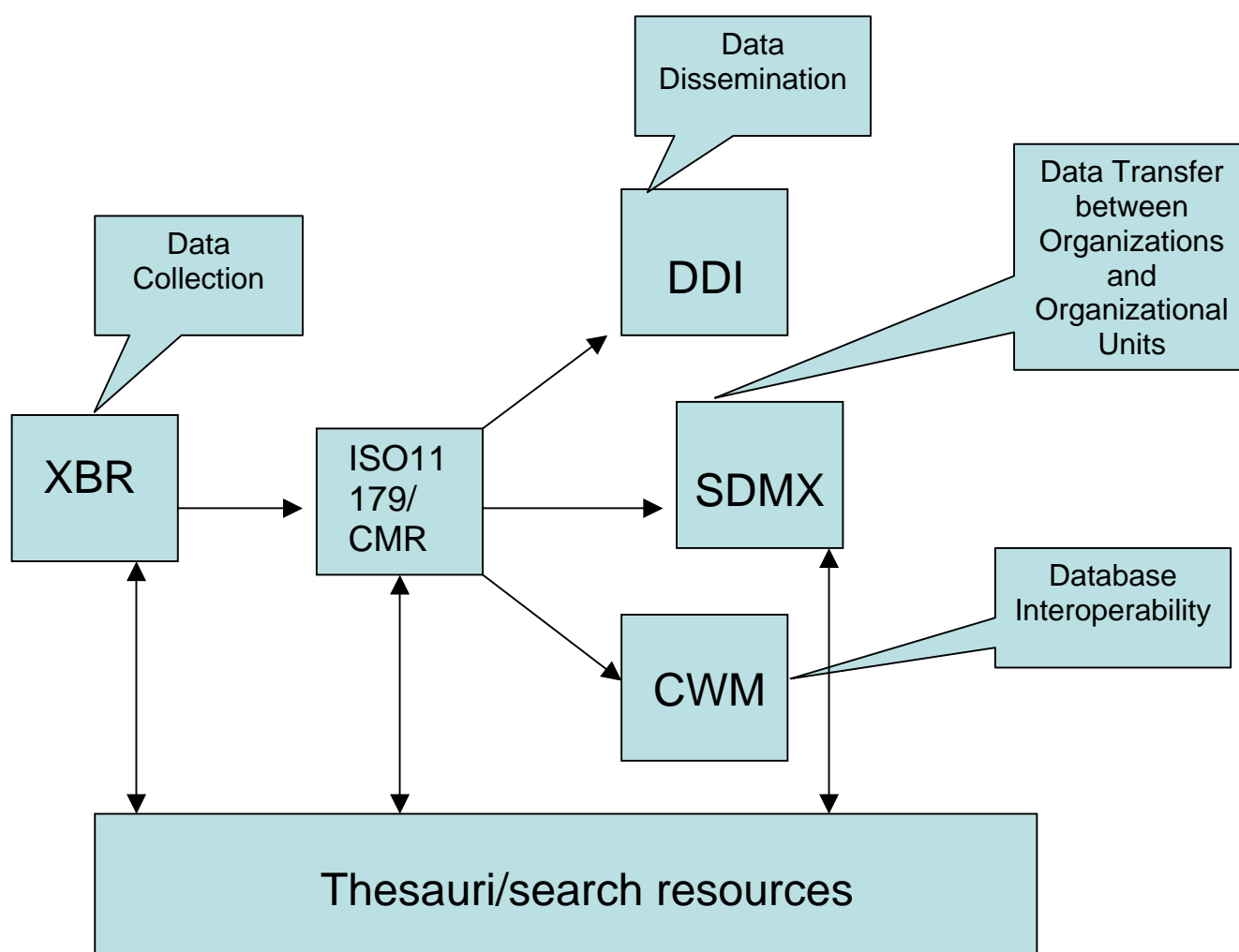
²² OMG. (2000). *The Common Warehouse Metamodel*. Object Management Group

VI. Integration

128. There are two ways to consider integrating the standards described in section V in a statistical office. They are, loosely speaking, a systems integration and a conceptual integration.

129. From the point of view of conceptual integration, the standards can be related as shown in Figure 7. The semantics, or meaning associated with data elements, should be contained in the electronic data collection instruments according to standards such as XBRL and other EDR metadata standards. It should be possible then to map the names, definitions and other metadata attributes from this collection metadata to the statistical offices central metadata repository, structured according to the ISO/IEC 11179/CMR model, or some other comparable standard. From this repository, metadata sets should be generated automatically according to output standards such as DDI, in the case of microdata sets, SDMX in the case of transfers and exchanges of aggregated data sets between statistical offices, and CWM when statistical data are made available for analysis in data warehousing environments. It should also be possible to automatically generate other metadata sets structured according to agency specific needs, such as the SDDS standard of the IMF, or a statistical agency's own data quality statement standard.

Figure 7: Conceptual integration of Metadata standards for Statistical Office



130. A deeper integration can be conceptualized if systems integration is also taken into account. The CWM is the main hub for the systems integration, and the CMR (including ISO/IEC 11179 and the Neuchâtel models) for the conceptual part. This division is not absolute, as the CWM has a semantic component and the CMR may be useful for driving systems. Here, we use each to their strengths. See Figure 8 for a schematic diagram representing these ideas.

131. As described in section V.A, the CWM is used to integrate disparate tools by providing a common metadata framework for describing the metadata model for each tool. The CWM provides the functionality to map between metadata models, and in this way, the tools may "talk" to each other. By this we mean that metadata from one tool is used by another.

132. The situation of a set of tools talking to each other throughout the survey life-cycle was described by Kent, et al (1998). There, they described an automatic survey processing paradigm driven completely by metadata. The need for each tool in the survey life-cycle to understand the metadata model of the tool before it is of primary importance. The CWM provides a solution to this problem - see Figure 8.

133. On the other hand, there is still the need for understanding everything within the survey-life cycle. By this we mean the needs of humans to comprehend the systems they build, maintain, integrate, and upgrade. This does not necessarily mean all the systems are computerized, however, the ultimate aim is to automate as much as possible.

134. The CMR, including ISO/IEC 11179, and the more detailed models for statistical agencies developed by the Neuchâtel Group deal primarily with semantics. As described in section III, the ISO/IEC 11179 standard is a realization of the terminological approach to data. This approach, as described, is concept based. The Neuchâtel Group model for statistical data follows the ISO/IEC 11179 approach but extends the concepts to also describe data sets, cubes, and tables. The classification model from the Neuchâtel Group thoroughly describes classifications, mappings between them, and their management. Along with the extensions of the CMR to describe questionnaires and samples, a complete conceptual description of a survey possible.

135. Figure 8 in this section pictorially represents what we have discussed here. There are six parts to the survey life-cycle depicted, and they are linked serially from the conception to dissemination. The links represent the transfer of data or metadata as needed between the parts. Each part includes the systems and tools in place to complete that part of the life-cycle. There is no requirement that these be automated. In any case, there is metadata in use for each part of the life-cycle and it is represented by the database symbol. Then, there is a link between the local metadata model and the CWM, which describes and maps the models.

136. In addition, there are three other standards mentioned, for input and output of statistical data: XBRL for reporting from respondents; DDI for transferring metadata to users of statistical data; and SDMX for transfer of data and metadata to other statistical offices. These standards have links back to the CWM because they are metadata models themselves. Finally, there are links to the CMR and Concepts & Terms databases because they too are metadata models. This way the CWM contains a map for all the metadata models the office uses.

137. The CMR is used for describing statistical data and the survey life-cycle. The purpose is to provide descriptions at the conceptual level, creating a concept system for the statistical office; although attributes for driving systems are contained there, too. The cloud emanating from the CMR database icon in Figure 8 is meant to convey this conceptual purpose.

138. The special terms and concepts used in a survey form a special language, and these concepts are the ones that appear as part of the descriptions in CMR instances. As described earlier, the object classes, properties, conceptual domains, value meanings, and data element concepts used to describe data are concepts in these special languages. Survey concepts, such as universes, populations, characteristics, and those described in questions (in a questionnaire) are also concepts in survey special languages. Therefore, we need to manage concepts and the terms designating them.

139. So, the CMR takes the concept system of the statistical office and gives it purpose. The purpose is the description of surveys and their data.

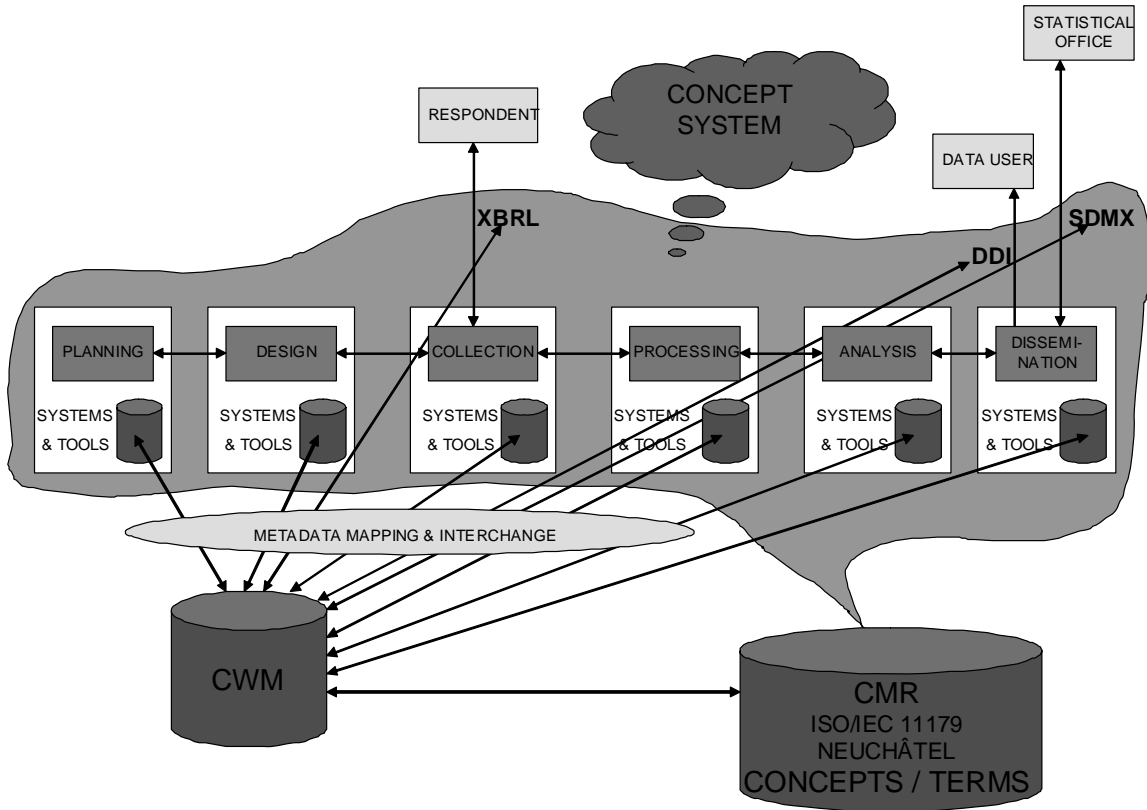


Figure 8: Metadata Architecture for Statistical Office

VII. CONCLUSION

140. Over the last few years, metadata systems and approaches for statistical offices have stabilized and matured. There is now a general recognition of what we mean by metadata and its form and function in statistical offices. A certain ecology of sometimes competing, sometimes complementary standards has also emerged and, to a certain extent, the outline of a process of natural selection can now be distinguished. This paper attempts to present the most prominent current members of this family of metadata standards and how they can be articulated to serve the needs of statistical offices. It should now become apparent that the goal is not to develop a single overarching standard that covers everything. Rather, a limited set of well developed standards, appropriately adapted to specific statistical junctions, can and

should become the target for world adoption and interoperability. A number of specific and achievable convergence targets have been identified in this paper and more could be developed by the METIS Task Force, in the context of an integrated framework such as the one presented in this paper. Given this advanced stage of development, continued sharing of experiences and expertise, but focused on consolidation and harmonization of approaches, should represent the next phase of the international collaboration undertaken under the METIS work program.

VIII. REFERENCES

- Bradley, W.J., Colquhoun, G. and Ryssevik, J., *Integrating ISO/IEC 11179 and the DDI in a Single Application*;
- Bradley, W.J., Gillman, D.W., Johanis, P. and Ryssevik, J. *Is your Agency ISO Compliant? Standardizing Metadata for Improved Knowledge Delivery in National Information System*, ISI Conference 2004, Berlin
- CEN. (1995). *Medical Informatics - Categorical Structures of Systems of Concepts*. Draft. Brussels: European Committee for Standardization.
- Date, C. (2003). *An Introduction to Database Systems (8th ed)*. Addison Wesley.
- Dunstan, T., Humphrey, C., *Discovering Microdata Variables: Comparing DDI Compliant Documentation to an ISO/IEC 11179 Metadata Registry*, Symposium 2005, October 2005, Statistics Canada
- Farance, F. & Gillman, D. (2006). Working Paper #12 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.
- Froeschl, K., Grossmann, W., & Del Vecchio, V. (2003). *The Concept of Statistical Metadata*. Deliverable #5 for MetaNet Project. Retrieved July 2004 from http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc.
- Gillman, D. (2006) Theory and Management of Data Semantics. In D. Schwartz (ed.) *Encyclopedia of Knowledge Management*. Hershey, PA, USA: Idea Group.
- Gregory, A., Nelson, C., *SDMX, ISO/IEC 11179 and the CMR*, METIS 2006, Geneva
- ICPSR (Inter-University Consortium for Political and Social Research). (n.d.). *Data Documentation Initiative*. Retrieved July 2004 from <http://www.icpsr.umich.edu/ddi>.
- ISO. (1999). *ISO 704: Principles and methods of terminology*. Geneva: International Organization for Standardization.
- ISO. (2000). *ISO 1087-1: Terminology – Part 1: Vocabulary*. Geneva: International Organization for Standardization.
- ISO. (2003). *ISO/IEC TR 20943-3: Procedures for achieving metadata registry content consistency, Part 3: Value domains*
- ISO. (2005). *ISO/IEC 11179 - Metadata registries (All Parts)*. Geneva: International Organization for Standardization and International Electrotechnical Commission.
- Johanis, P. (2000). *Statistics Canada's Integrated Metadatabase: Our Experience To Date*. Invited Paper #3 presented at the UNECE Workshop on Statistical Metadata. Washington, DC.
- Johanis, P., Brooks B., Dunstan T., and Lévesque, J-S. (2003), *Statistics Canada's Implementation of the Data Element Model*, Open Forum, Santa Fe

- Kent, J.-P. (1998). *Take Care of the Meta, and the Meta Will Take Care of the Data*. Working Paper #6 presented at the UNECE Workshop on Statistical Metadata. Geneva, Switzerland.
- Lakoff, G. (2002). *Women, Fire, and Dangerous Things* (Reprint edition). University of Chicago Press.
- Mechanda, K., Johanis, P., and Webber M., (2003) *Conceptual Model for the Definitional Metadata of a Statistical Agency*, Open Forum, Santa Fe
- Oakley, G. (2004, February). *Using ISO/IEC 11179 to help with metadata management problems*. Invited Paper #9 presented at the Joint UNECE, Eurostat, OECD Workshop on Statistical Metadata,. Geneva.
- OASIS (2002). *OASIS/ebXML Registry Information Model, v2.0*. Organization for the Advancement of Structured Information Standards
- Ogden, C. and Richard, I. (1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt.
- OMG (2000). *The Common Warehouse Metamodel*. Object Management Group
- Poole, J., et al (2002). *The Common Warehouse Metamodel*. New York: John Wiley and Sons.
- Ryssevik, J., Glover T. and Colquhoun, G., *Relationship to ISO/IEC 11179*
- Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.
- Statistics Canada, *Policy on Standards*,
<http://www.statcan.ca/english/about/policy/standards.htm>
- Technology Working Group of XBRL Spain, *White Paper on XBRL Concepts and Recommendations*
- W3C (2004). *Extensible Markup Language*. XML 1.1 reference specification
- Wedberg, A. (1982). *A History of Philosophy - Vol 1: Antiquity and the Middle Ages*. Oxford: Clarendon Press

**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

SDMX, ISO 11179 AND THE CMR

Invited Paper

Submitted by SDMX Standards Team¹

¹ Prepared by Arofan Gregory and Chris Nelson.

SDMX, ISO 11179 and the CMR

Arofan Gregory

Chris Nelson

Metadata Technology Ltd. April 2006

Contents

1	Purpose and Scope of the Paper	3
1.1	Purpose	3
1.2	Scope.....	3
2	Scope of the Standards.....	4
2.1	ISO 11179.....	4
2.2	CMR.....	4
2.3	SDMX	4
3	Areas of Commonality	6
3.1	Concepts, Data Elements, and Value Domains	6
3.2	Registry.....	6
4	Overview of the ISO 11179 Conceptual Domain	6
5	Describing CMR Object Classes.....	7
6	Describing SDMX Object Classes.....	8
6.1	Introduction.....	8
6.2	Metadata Set Object Classes	8
6.3	Data Set Object Classes	11
7	Registry	15
7.1	ISO 11179.....	15
7.2	SDMX Registry	16
8	Summary and Conclusions	17
8.1	Summary	17
8.2	Conclusion.....	18

1 Purpose and Scope of the Paper

1.1 Purpose

ISO/IEC 11179 is concerned with the semantics of data, whilst SDMX is concerned with the structure of data. In order to achieve their respective goals, the two standards have some similar constructs and, whilst these could be compared, this is not a very productive exercise. The intent of the standards is different and the way each use these constructs is different.

Therefore it is not the intent of this paper to explore whether ISO 11179 semantics can be described in SDMX, or whether SDMX structures can be described by ISO 11179, as the answer is “perhaps, but this is not how the standards should be used”. The intent is to explore whether and how ISO 11179 Data Elements can be constructed from SDMX data and metadata structures, or instances of these structures, so that organisations can map ISO 11179 repository items with SDMX repository items.

It is also the intent of this paper to describe briefly the way the Corporate Metadata Repository model (CMR) uses ISO 11179 to capture the underlying semantic of key structures in the CMR, and to describe the scope of the registry models in ISO 11179 and SDMX, to show how these can interact.

Note that, unless referenced in a specific context, ISO/IEC 11179 is referred to hereafter as ISO 11179.

1.2 Scope

The scope of this paper is summarised in the diagram below.

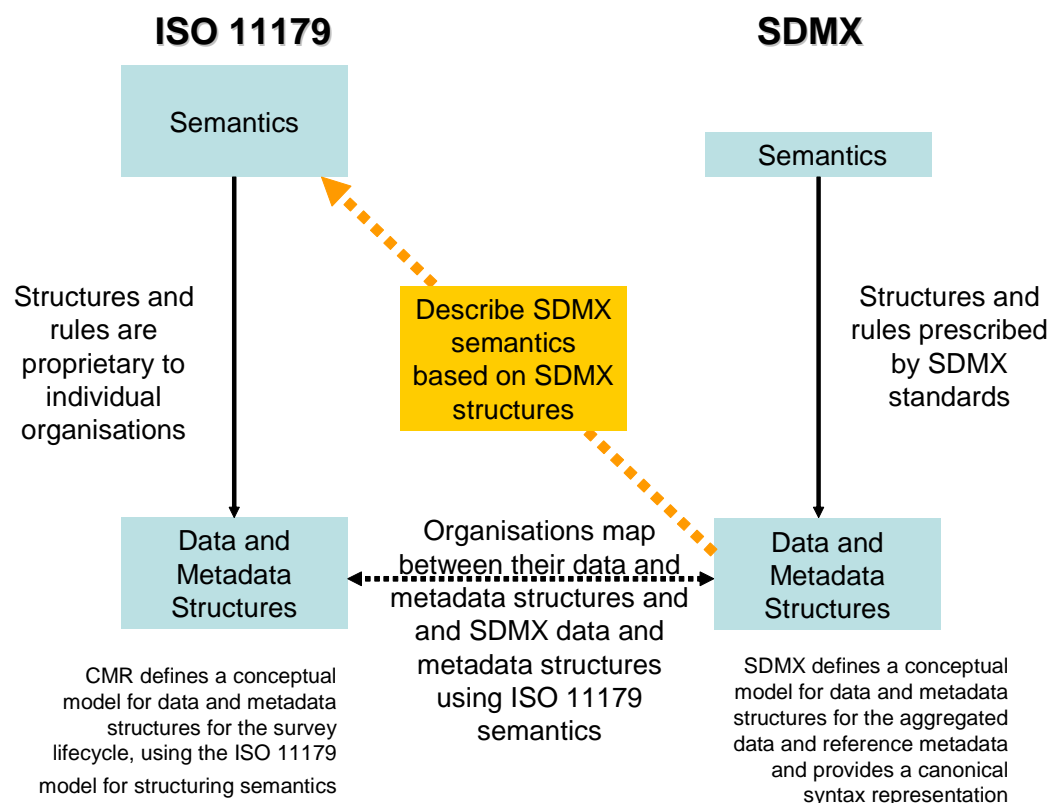


Figure 1: Schematic of the possible scope of using ISO 11179 to describe the semantics of SDMX constructs

The scope of this paper is show how data elements semantics can be derived from SDMX structural definitions, and how and why this mapping differs from the way this is achieved in CMR.

This paper takes as its base Part 3 (Registry metamodel and basic attributes) and Part 5 (Naming and identification principles for data elements), of ISO/IEC 11179 and provides a mechanism for naming data elements based on the SDMX Data and Metadata Structure Definitions.

2 Scope of the Standards

2.1 ISO 11179

“Some standards describe content and not format, metadata standards describe the data necessary to describe other data or processes, and data semantics refers to the meaning of data. ISO/IEC 11179 is a metadata content standard focused on the semantics of data”.

(Daniel Gillman “CORPORATE METADATA REPOSITORY (CMR) MODEL” EU Metanet R&D Project Stockholm 2002)

ISO 11179 is not concerned with the structure of data: the metamodel is concerned with the artefacts required to describe the semantics of data. There is no canonical representation of ISO 11179 in syntactic form such as an XML schema.

2.2 CMR

The CMR extends the ISO/IEC 11179 (Part 3) metamodel in the field of statistics. Specifically the CMR model is designed to support:

- metadata necessary to describe the survey life cycle
- linkages between similar designs and processes used across surveys
- use of metadata to drive systems in support of the survey life cycle

The CMR model is a conceptual model and like ISO 11179 there is no canonical representation in syntactic form.

2.3 SDMX

The SDMX metamodel is concerned very much with the structure of data and metadata and with the semantics required to understand the meaning of the data and metadata carried in these structures. It is not concerned per se with modelling structures for defining the semantics of data elements.

Therefore, it should be possible to define an SDMX object (i.e. an instance of an SDMX class such as a Data Set or Observation) in terms of the semantics defined in the ISO 11179 metadamodel. How this can be achieved is detailed below.

SDMX is also concerned with syntax specifications, and harmonising content standards such as the names of “Concepts”.

The SDMX specifications encompass:

- Information Model
- Syntax implementations based on the Information model (XML schemas and UN/EDIFACT)
- Content

- Statistical subject matter domain scheme
- Cross-domain concepts
- Metadata terminology (MCV – Metadata Common Vocabulary)

The diagram below shows the way these three key deliverables interact:

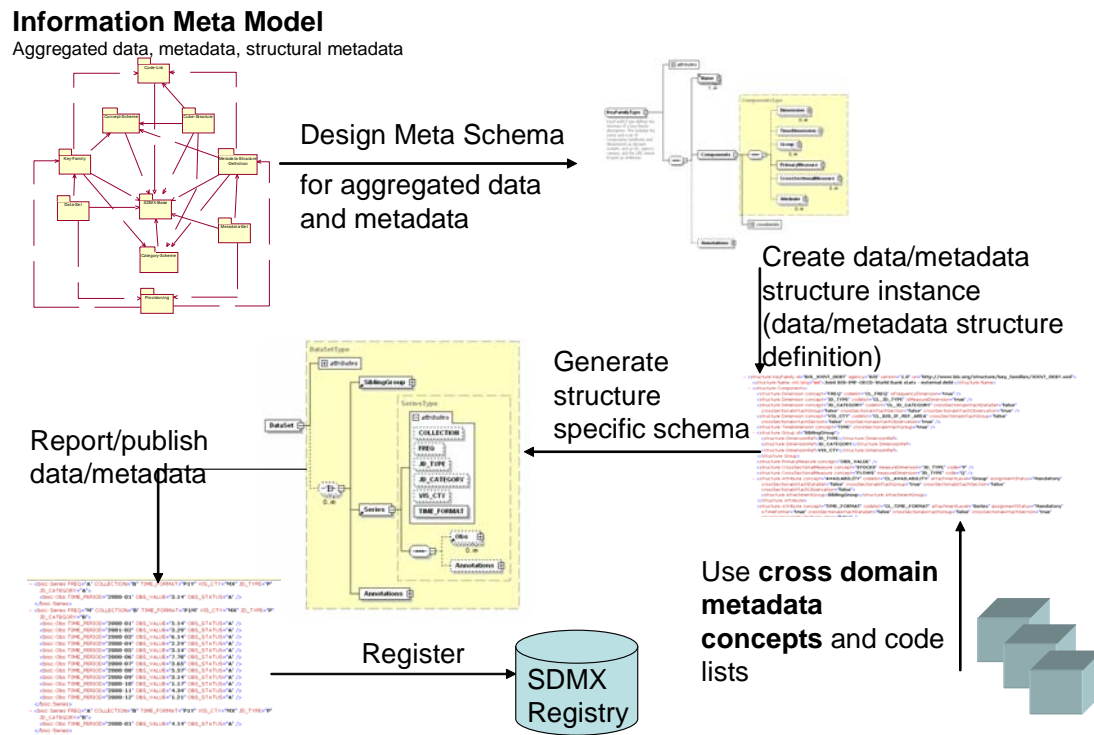


Figure 2: Interaction of key SDMX deliverables to create data and metadata schemas

The Information Model defines the artefacts necessary to support data and metadata publishing and exchange, and the way these artefacts are associated with each other. The model is used as the source of the syntax specifications – there are specifications for both XML representations (SDMX-ML) and UN/EDIFACT representation (SDMX-EDI) of the model, though the SDMX-EDI is limited to supporting just the data side and the necessary structural metadata to describe this. The diagram above depicts the XML scenario. The meta schema specifies the way to define a Data Structure Definition (also known as a Key Family), or a Metadata Structure Definition. A specific Data or Metadata Structure Definition (e.g. for External Debt, or Demography statistics) is defined in an instance of this meta schema. In order to achieve interoperability at the semantic level as well as the syntactic level, SDMX has worked on content standards and in particular on the definition of cross-domain concepts. Data and metadata structure designers are encouraged to use these concepts wherever possible, including the “core” representations (i.e. allowed format and code lists) defined for them.

The individual Data and Metadata Structure Definition is then used as the basis for generating a specific XML schema that supports the publishing or exchange of a data set or metadata set for that specific type of data or metadata. Therefore, a schema for reporting Balance of Payments data would be different from the schema generated to report Labor statistics. However, the difference is only in the naming of the XML tags, which reflect the names of the concepts used in the structure definition. SDMX defines a standard way of generating these

schemas from the structure definition and these schemas can (and should) be generated programmatically. Indeed, there are now freely available tools that do this.

Note that SDMX-EDI has just one syntactical form for the data set – there is no equivalent to a data structure specific SDMX-EDI representation. SDMX-ML also has an equivalent “one schema fits all” for the data (and also one for the reference metadata).

As all of these syntactic forms are created from the same Information Model, it is possible and practical to transform a data set or metadata set from one form to the other. Freely available tools also exist for these transformations between equivalent XML expressions, and between SDMX-EDI and SDMX-ML.

SDMX also has the concept of a registry, which is described below. Data and Metadata Sets can be registered in the registry and can be “discovered” by applications querying the registry.

3 Areas of Commonality

3.1 Concepts, Data Elements, and Value Domains

Both standards support the definition of concepts (Data Element Concept in ISO 11179). ISO 11179 associates these Concepts to actual Data Elements but it is not the role of ISO 11179 to build these Data Elements into structures, such as Data Sets and Questionnaires. This is left to other models which are domain specific, such as the CMR and SDMX.

Whilst the CMR is directly linked to the ISO 11179 model to give a conceptual structure to support the semantics of some CMR artefacts, the SDMX model is not linked in this direct way. It will be seen later that the SDMX model is actually a metamodel for defining data and metadata structures, and it is the instances of this metamodel (i.e. specific data or metadata structure definitions) that need to be mapped to ISO 11179. Therefore, in order to achieve this mapping there must be a set of rules based on the metamodel but applied to the instance. This will be explained more fully below.

3.2 Registry

Both standards support the concept of a registry

- ISO 11179 specifies a registry metamodel
- SDMX does not specify a registry metamodel
- ISO 11179 does not specify registry interfaces based on the registry model
- SDMX specifies registry interfaces based on the SDMX model

Implementation of a registry in SDMX and mapping the SDMX Information Model to a registry model is left to the implementor. Therefore, a registry developer can use an ISO 11179 registry implementation, ebXML registry implementation, or bespoke registry implementation.

We will now look at these areas of commonality in more detail.

4 Overview of the ISO 11179 Conceptual Domain

The diagram below is high level overview of the part of ISO 11179 concerned with semantics.

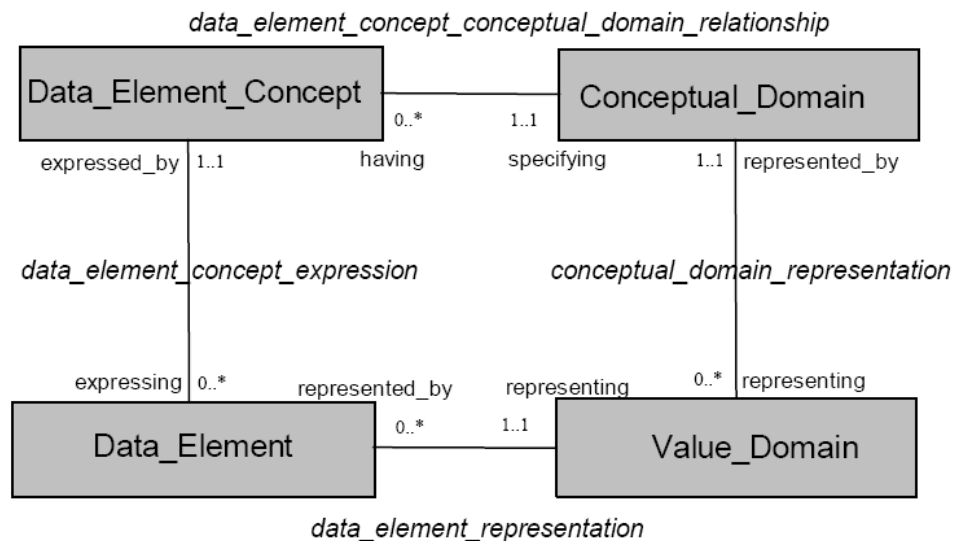


Figure 3: High level overview of the structures to specify data semantics ISO 11179

The **Data_Element_Concept** is a combination of concepts and properties that describe something of interest in the real world. The **Data_Element_Concept** is expressed independent of any internal or external representation. The metadata objects in this region are Object Classes (encompassing Concepts and Concept Relationships) and Properties, which may be combined to form Data Element Concepts.

The realm of possible values that the **Data_Element_Concept** can take when it is used as a **Data_Element** is defined in the **Conceptual_Domain**.

Data_Elements are derived from **Data_Element_Concepts** and are linked to exactly one **Value_Domain**.

In order to give an ISO 11179 semantic to SDMX objects, it is important to name those parts of the SDMX model (the Object Classes in the SDMX model) in ISO 11179 terms.

5 Describing CMR Object Classes

The CMR is a conceptual model that supports the survey lifecycle. "For the CMR model, the survey life cycle is comprised of the following stages (LaPlant, et al, 1996): Content, Planning, Design, Collection, Processing, Analysis, and Dissemination. Metadata to support and describe all the stages is accounted for in the CMR model" (Daniel Gillman "CORPORATE METADATA REPOSITORY (CMR) MODEL" EU Metanet R&D Project Stockholm 2002).

Element and Concept semantics are provided by an association to ISO 11179 artefacts as exemplified in the diagram below.

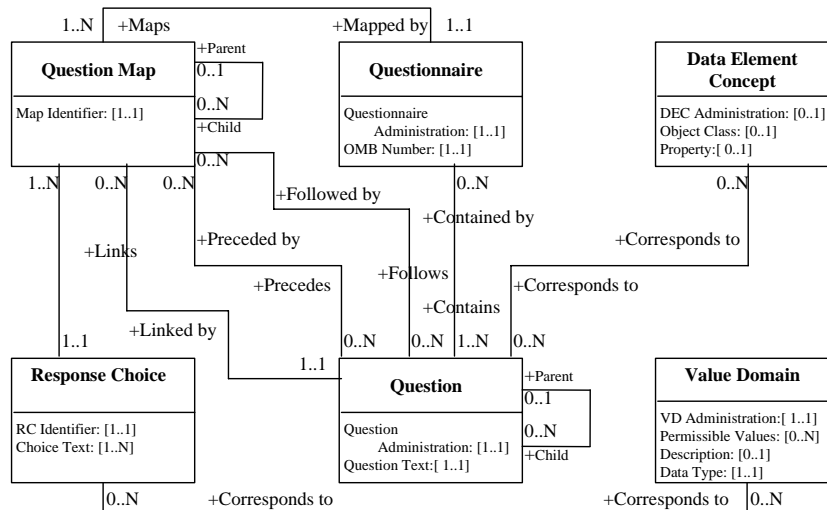


Figure 4: Example of integration of some parts of the CMR with ISO 11179 constructs

Here one can see:

- the Question is associated to Data_Element_Concept
- the Response_Domain is associated to Value_Domain

The Data Set in the CMR contains the output from a survey. Data Set components (called Data Details in CMR) are associated directly to Data_Elements in the ISO 11179 model. As the CMR Data Set is concerned with Data_Elements the classes in the ISO 11179 can be used directly in this way to give the semantics required.

6 Describing SDMX Object Classes

6.1 Introduction

The core of the SDMX Information Model is concerned with the structure of data and metadata sets. The Information Model has a far wider scope than this (mapping, process, registry), but for the purpose of this section of the paper the focus is on this core of data and metadata sets.

For the data side of SDMX, there is a very tight constraint on the Object Classes for which metadata or data can be reported, whereas for the metadata side of SDMX, metadata can be reported for any Object Class in the model, provided it can be identified (i.e. it must be an IdentifiableArtefact in SDMX terminology). It is useful, therefore, to start with the metadata set as ISO 11179 equivalent Data_Elements can be created in a semantically meaningful way.

6.2 Metadata Set Object Classes

The structure of a metadata set in SDMX is described in terms of:

- The Object Class to which metadata is to be attached, and the components that comprise the unique identifier of the Object Class
- The Concepts and the relevant Value Domains that are used to describe the metadata that can be attached to an Object of that Object Class.

A schematic of this part of the SDMX metamodel is shown below.

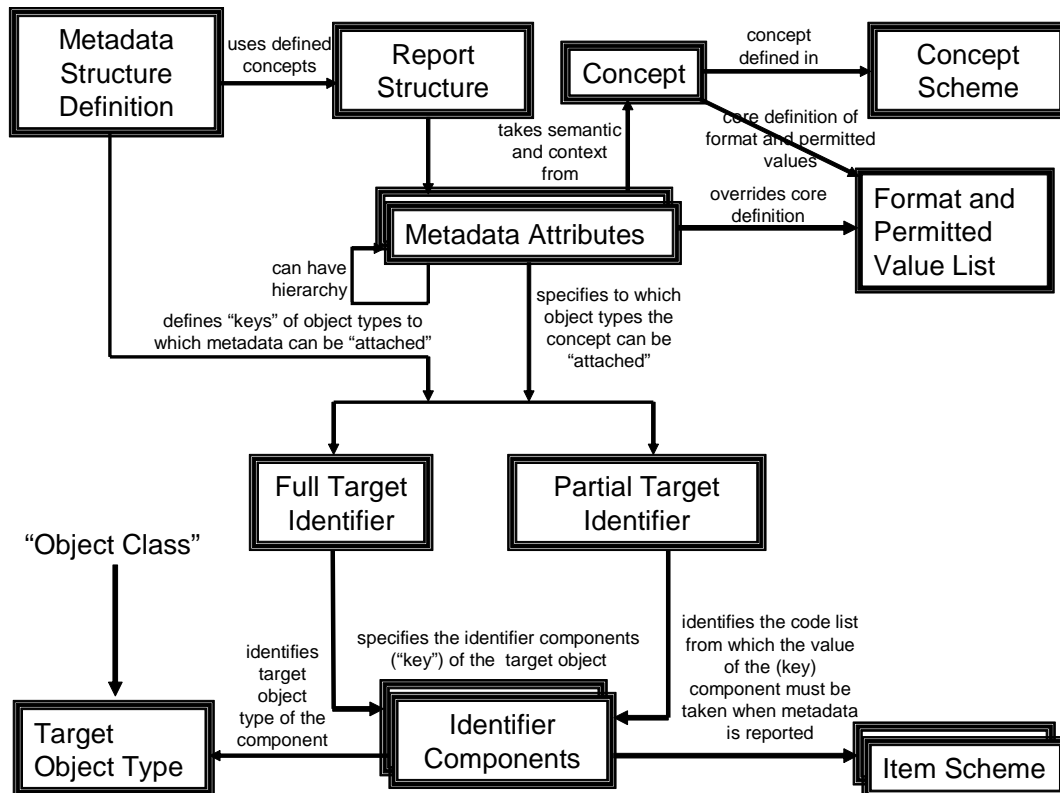


Figure 5: Schematic of the SDMX metadata structure definition constructs

In theory the metadata structure definition there is no restriction on the type of Object Classes for which metadata can be reported – in reality metadata could be reported for objects corresponding to any Object Class in any model, though in SDMX this is normally confined to the Object Classes in the SDMX model. The Object Class is identified by the sum of its Target Object Types.

Therefore in ISO 11179 terms an instance of the structure in terms of the Object Class and the Metadata Attributes can be defined as a number of Data_Elements. For instance, an organisation may wish to collect metadata from countries about the way data are collected for different domain categories such as Balance of Payments, National Accounts, Labor (e.g. the IMF SDDS). A structure definition for this type of metadata would have, as its target object, the union of a Domain Category and a Data Provider (Country). This identifies the object.

A schematic of this part of the model is shown below.

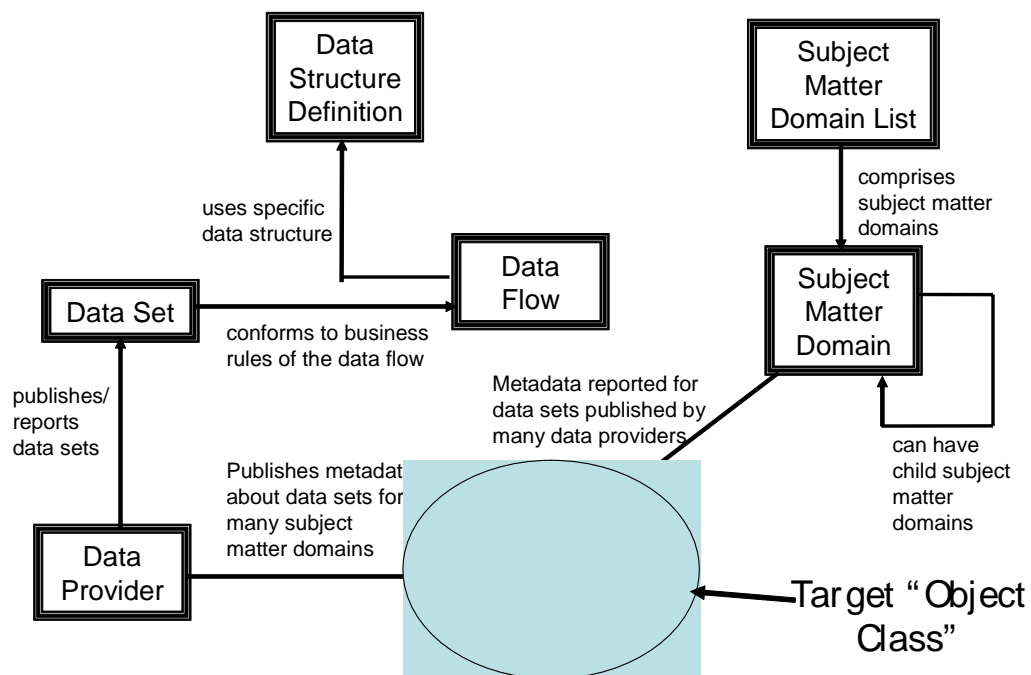


Figure 6: Schematic of part of the SDMX metamodel for data and metadata reporting

The actual metadata is collected for a variety of Concepts, each of which has a specific type of representation. Example of such Concepts are:

- Confidentiality
- Survey Source
- Simultaneous Release

The metadata structure definition would contain the following specification

Target Object Type	Metadata Attributes
DATA_PROVIDER_DOMAIN Identified by components:	CONFIDENTIALITY SUREVEY_SOURCE SIMULATANEOUS_RELEASE
DATA_PROVIDER SUBJECT_MATTER_DOMAIN	Each of these has a specific representation(Value Domain) defined in the metadata structure definition

It is possible to generate ISO 11179 Data_Elements from the instance of the metadata structure definition. Example element names would be:

Data_Provider_Domain.Confidentiality.Code

Data_Provider_Domain.Survey_Source.Text

Data_Provider_Domain Simultaneous_Release.Text

If one analyses the structure for these Data Element names in terms of the ISO 11179 model each one is structured as follows:

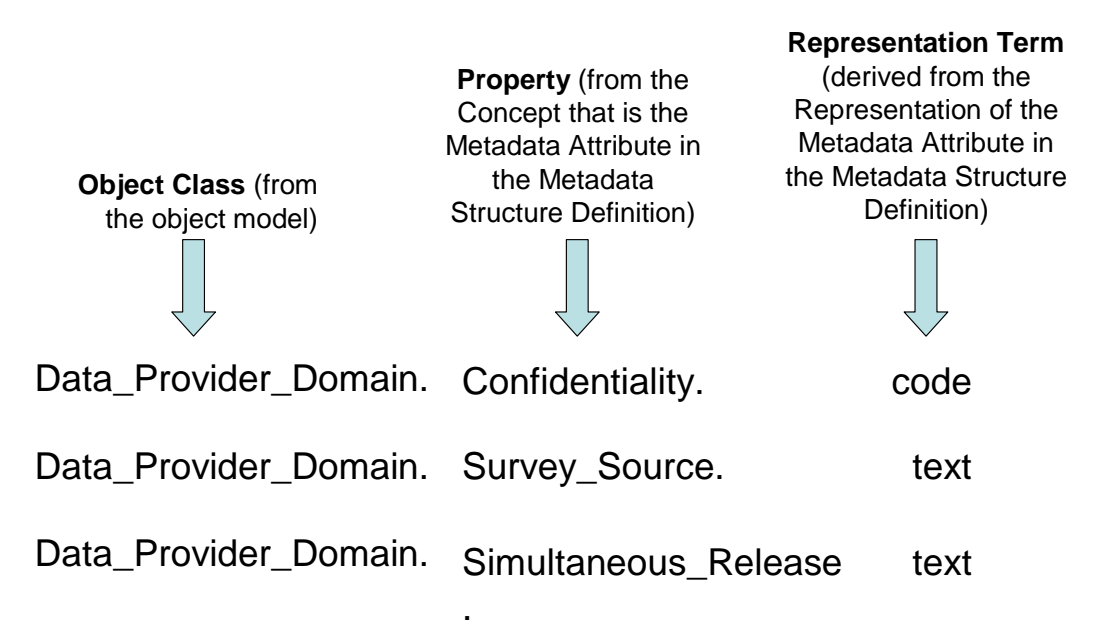


Figure 7: Derivation of data elements from an SDMX metadata structure definition

6.3 Data Set Object Classes

For the data side of SDMX the schematic of the model is:

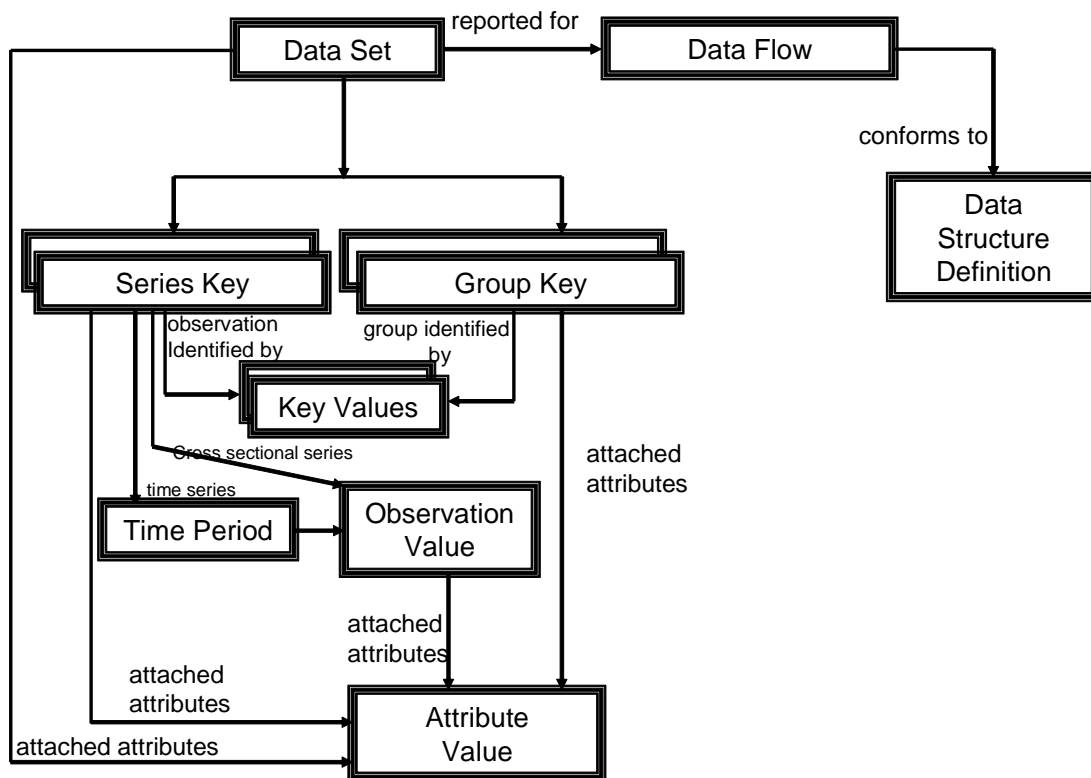


Figure 8: Schematic of the SDMX data set constructs

The object classes to which metadata (Attributes) can be attached are restricted in the model to:

- Series
- Group (qualified by the Id of the Group)
- Observation
- Data Set

The only Object Class to which a data value can be attached is the Series, qualified by the Time Period for a timeseries. The Object Class for this is the Observation Value. Therefore, the Observation Value can have both a data value and attributes.

Example ISO 11179 equivalent Data_Elements that could be generated from an instance of a Data Structure Definition would be:

Observation.Confidentiality.code

Observation.Value.number

Series.Availability.code

Group.Title.text

Data_Set.Title.text

If one analyses the structure for these Data Element names in terms of the ISO 11179 model each one is structured as follows:

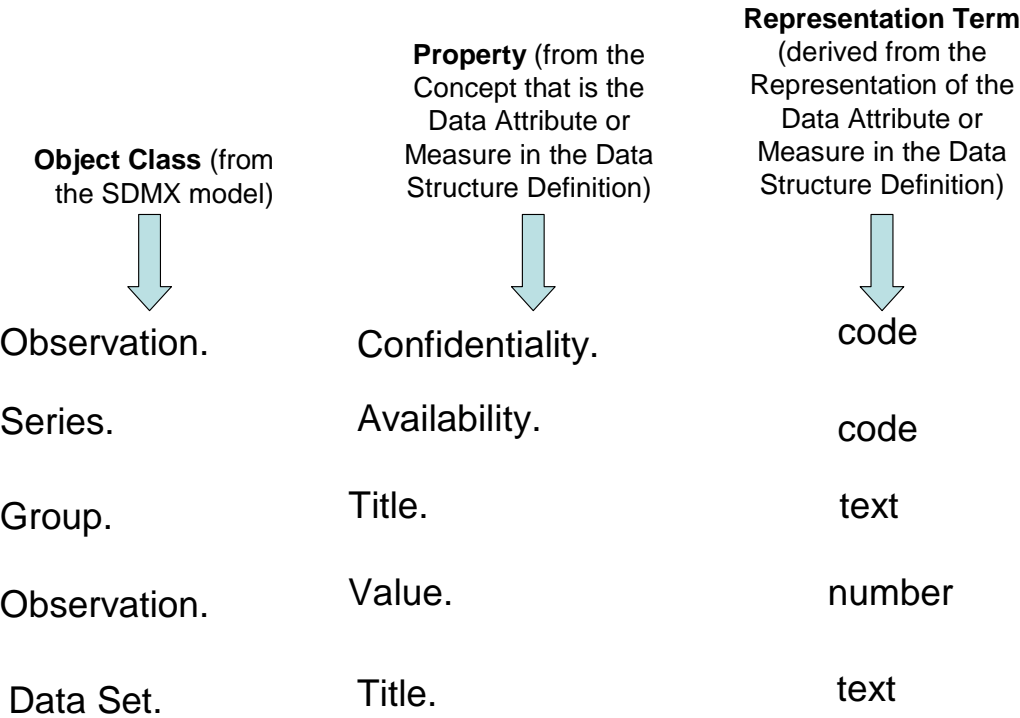


Figure 9: Derivation of data elements from an SDMX data set

This works well for the Data Attributes for the observation and data set (e.g. Confidentiality, Title, Availability) but it does not reveal the true semantic of the Observation Value, or Data Attributes for the Series or Group (a subset of the full key for the series).. The true semantic of this value is different for each Series and Group. Consider the following data set.

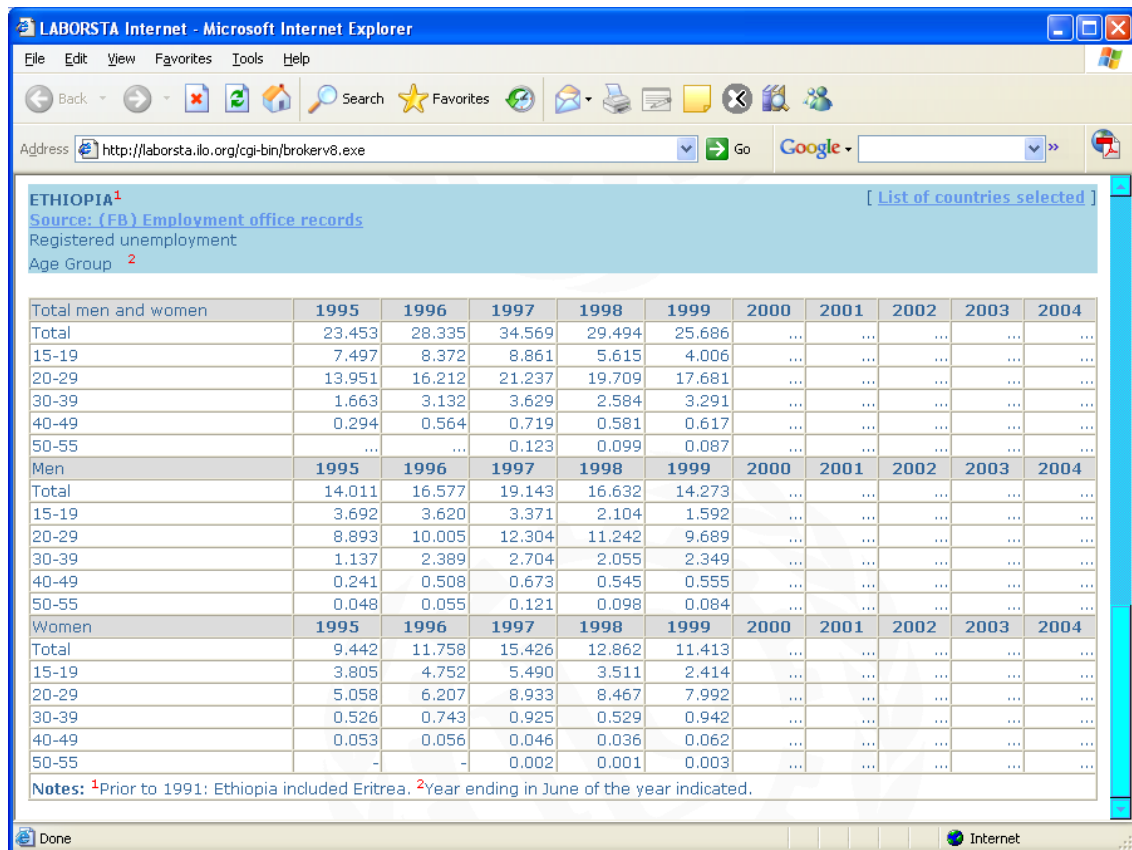


Figure 10: Example data set

This has the following (simplified) structure.

Target Object Type	Properties (Data Attribute) or (Measure)
Series Identified by components FREQUENCY COUNTRY GENDER AGE_GROUP	[No series level attributes in this example] TIME.NUMBER_OF_UNEMPLOYED
Country Group Identified by components COUNTRY	GEOGRAPHICAL_COVERAGE TEMPORAL_COVERAGE
Observation Value Identified by component: NUMBER_OF_UNEMPLOYED	MEASURE_UNIT UNIT_MULTIPLIER STATUS CONFIDENTIALITY

In the example data set the actual semantic of the observation value is the value of each component of the Key plus the qualifier of the Time Period as shown in the examples below.

Observations

```
Annual.Ethiopia.Male.15-19.1995.Number_of_Unemployed.number  
Annual.Ethiopia.Male.15-19.1996.Number_of_Unemployed.number  
Annual.Ethiopia.Male.20-29.1995.Number_of_Unemployed.number  
Annual.Ethiopia.Male.20-29.1996.Number_of_Unemployed.number  
Annual.Ethiopia.Female.15-19.1995.Number_of_Unemployed.number  
Annual.Ethiopia.Female.15-19.1996.Number_of_Unemployed.number  
Annual.Ethiopia.Female.20-29.1995.Number_of_Unemployed.number  
Annual.Ethiopia.Female.20-29.1996.Number_of_Unemployed.number
```

In other words, in ISO 11179 terms each observation value is a Data_Element. Whilst statistical systems do not use this notation explicitly for aggregated data, they do nevertheless store and manipulate data with this semantic, and could, if it is useful, generate such element names.

The other Data_Elements that are derived from this specific data set example are:

Country Group Attributes

```
Ethiopia.Geographical_Coverage.text  
Ethiopia.Temporal_Coverage.text
```

Observation Attributes

```
Number_of_Unemployed.Measure_Unit.code  
Number_of_Unemployed.Unit_Multiplier.code  
Number_of_Unemployed.Status.code  
Number_of_Unemployed.Confidentiality.code
```

7 Registry

7.1 ISO 11179

ISO 11179 has an explicit registry metamodel as part of its model. Indeed the Part 3 of the standard is called “Registry metamodel and basic attributes”. A part of the registry metamodel is shown below. ISO 11179 specifies which artefacts from the model are an “Administered_Item” and can therefore be registered in an ISO 11179 registry (e.g. Data_Element, Data_Element_Concept are both an Administered_Item).

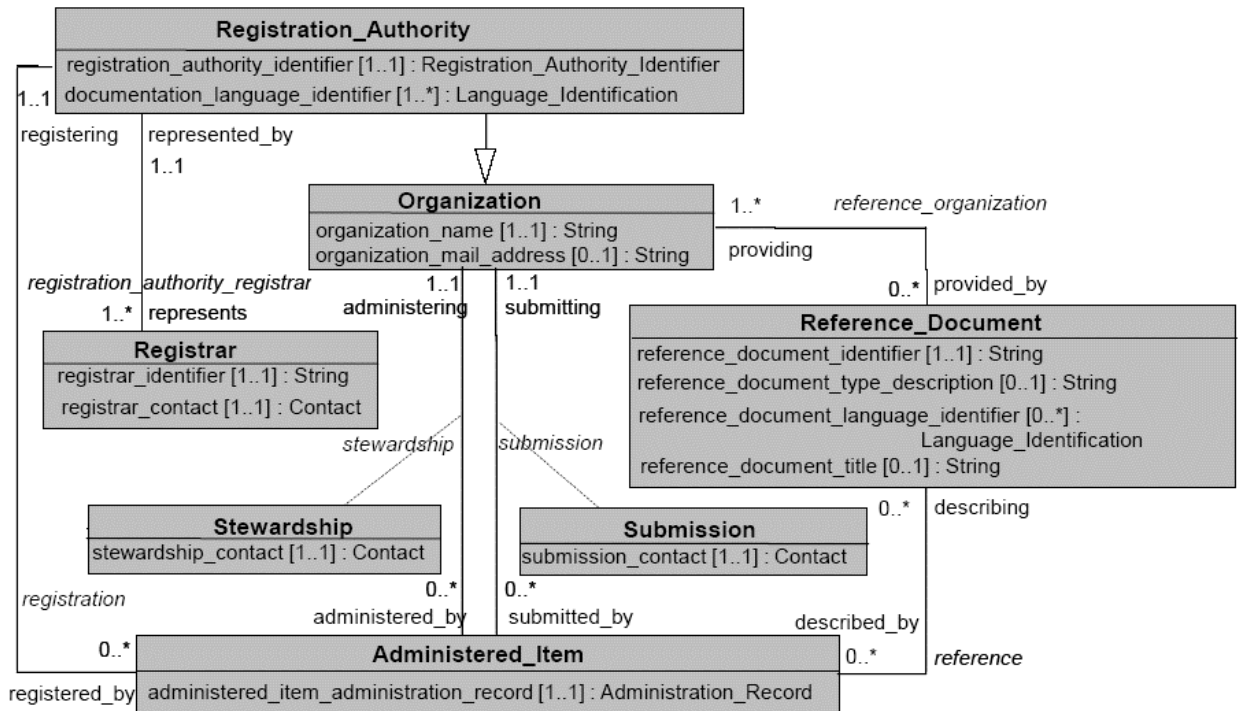


Figure 11: Class diagram of a part of the registry area of the ISO 11179 model

However, as yet there is no specified interface for interacting with an ISO 11179 registry and no registry service specification.

7.2 SDMX Registry

On the other hand, SDMX has a registry interface specification, but does not specify a registry metamodel. The SDMX registry interfaces are based on the semantics and use cases of the SDMX Information model.

SDMX Standards – Registry

Information Meta Model

data/metadata provisioning and registry
interfaces

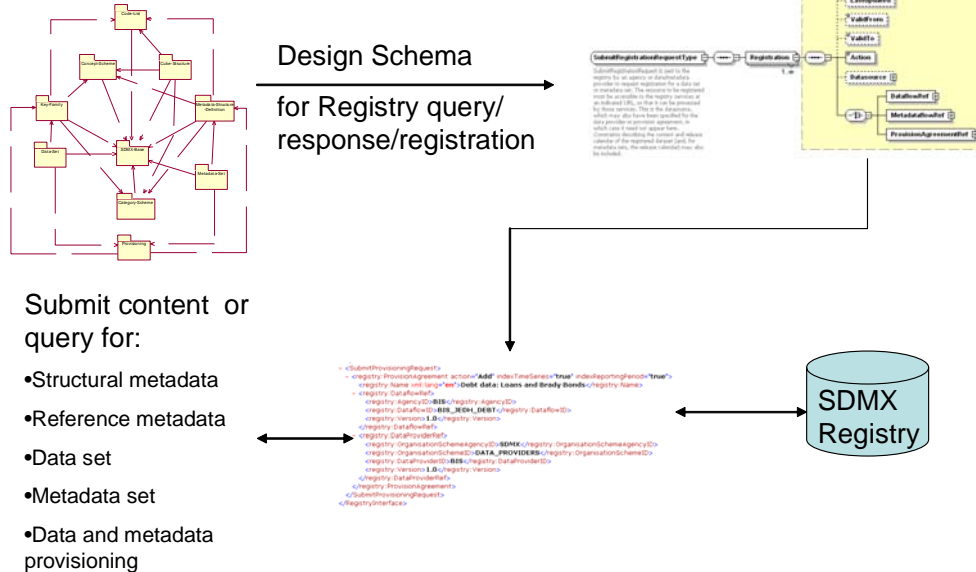


Figure 12: Deriving SDMX registry interfaces from the SDMX model

The interfaces (e.g. `Submit_Registration_Request`) are specified in the SDMX registry services schema. A registry provider can implement these interfaces in whatever registry product he chooses to use. The SDMX pilot project registry implementation uses a registry that complies to the ebXML registry specification (this is now an ISO standard – ISO 15000 parts 3 and 4). Of interest is that the ISO 11179 model was one of the inputs to the ebXML RIM (registry information model) and so has much functional equivalence to the “registry” region of the ISO 11179 conceptual model.

8 Summary and Conclusions

8.1 Summary

The ISO 11179 model supports the semantic specification of data. This model can be used as the basis for organising the semantics of data in an organisation. The organisation may wish to use instances of these semantic constructs to populate structural metadata about the objects that are of interest to it. For survey metadata one such structural model is the CMR and relevant constructs in this model can be associated with the semantic constructs in ISO 11179 in order to provide support for the semantic meaning. For aggregated data and reference metadata there is the structural metamodel of SDMX. SDMX is a metamodel and as such the semantics of the objects are embedded in the instances of the data or metadata structure metamodel. In other words, the semantics of SDMX objects can be derived from a specific data or metadata structure.

The reference metadata metamodel supports the definition of properties (called metadata attributes in SDMX) for any identified Object Class (called Target Object Type in SDMX). This maps well to the scope of ISO 11179 and so these objects and properties can be named and stored in an ISO 11179 compliant registry where other semantic aspects can be added in accordance with the ISO 11179 model.

On the other hand, it is more difficult to realise the true semantic of the objects described in an instance (specific data structure definition) of the data metamodel. This is because the equivalent “Target Object Type” or Object Class, is a multi-faceted artefact – it is a multi-

dimensional key. Therefore, whilst the constructs of a data structure definition can realise ISO 11179 data elements, this may not be very useful to organisations. It may be that a more useful semantic can be realised by the instance of the data structure definition i.e. the data set itself. In other words, the true semantic of an observation value in an aggregated data set is realised by the sum of the actual values of the individual key components.

8.2 Conclusion

In this paper we have shown the broad scope of the three models of ISO 11179, CMR, and SDMX, in the areas in which they interact. We conclude that ISO 11179 can act as the pivotal model for mapping the semantics of CMR structures with the semantics of SDMX structures. We have done this from a conceptual point of view based on the structures of the models and the way these structures are used.

We also conclude that ISO 11179 was not built to support the definition of data elements in a metamodel because the definition of these semantics is not useful. Rather, we conclude that any definition must be derived from an instance of the metamodel. For a specific metadata structure definition this results in meaningful data elements that can be given further semantic in an ISO 11179 centric world.

However, for a specific data structure definition this may not result in meaningful data elements, and it may be necessary to base data element definitions on the content of the data set itself. We do not seek and, indeed are not qualified, to prove that any such mapping is useful for specific applications - the utility of this is for others to judge. We have established that a clear semantic mapping is possible, however, should this prove useful in an application context.

**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

CONCEPTUAL MODELLING OF STATISTICAL METADATA AND METADATA DATA MODEL IN CoSSI

Supporting Paper

Submitted by Statistics Finland¹

I. INTRODUCTION

1. Statistics Finland has been starting to implement a statistical metadata concept based on Statistical Information Model called CoSSI (Common Structure of Statistical Information)².
2. In modelling of statistical information the methodological starting point of the definition of metadata is that in the conceptualisation of the contentual description of statistical information use is made as far as possible of the concepts characteristic of statistical information, the concepts and concept structures it contains and the logic that allows sufficiently multifaceted and complex concept structures for an exhaustive description of the information content³.
3. As the used description method of CoSSI allows implementation of complicated structure descriptions, the procedure does not have essentially any factors that would per se somehow force to contract or limit the contentual description.
4. Results obtained when defining statistical information by setting out from the above points of departure are described in the adjacent figure (Figure 1). On the one hand, statistical information has been defined by using a conceptual analysis, the results from which have been depicted as conceptual models of statistical information and, on the other, an analysis has been made of different forms of organising statistical data and

¹ Prepared by Heikki Rouhuvirta, Statistical Methodology R&D, Statistics Finland; heikki.rouhuvirta@stat.fi.

² Technical description of CoSSI is contained in the definition, see Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.91, Statistics Finland 2003. Also available on the Internet at:

http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf

³ The foundations and points of departure for the structuring of statistical information, as well as the requirements set on the system for describing it are discussed in more detail in Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at: http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

presenting statistical information, which has been used to specify basic models for presenting statistical data. Structural models of data and related data models have been produced for concept models and different forms of organising data, and definitions for these have been implemented in the CoSSI model as multi-level hierarchical (so-called tree-structured) data models⁴. The data models have been documented as XML DTD definitions. The basic method used in the implementation was the "From model to markup" approach.

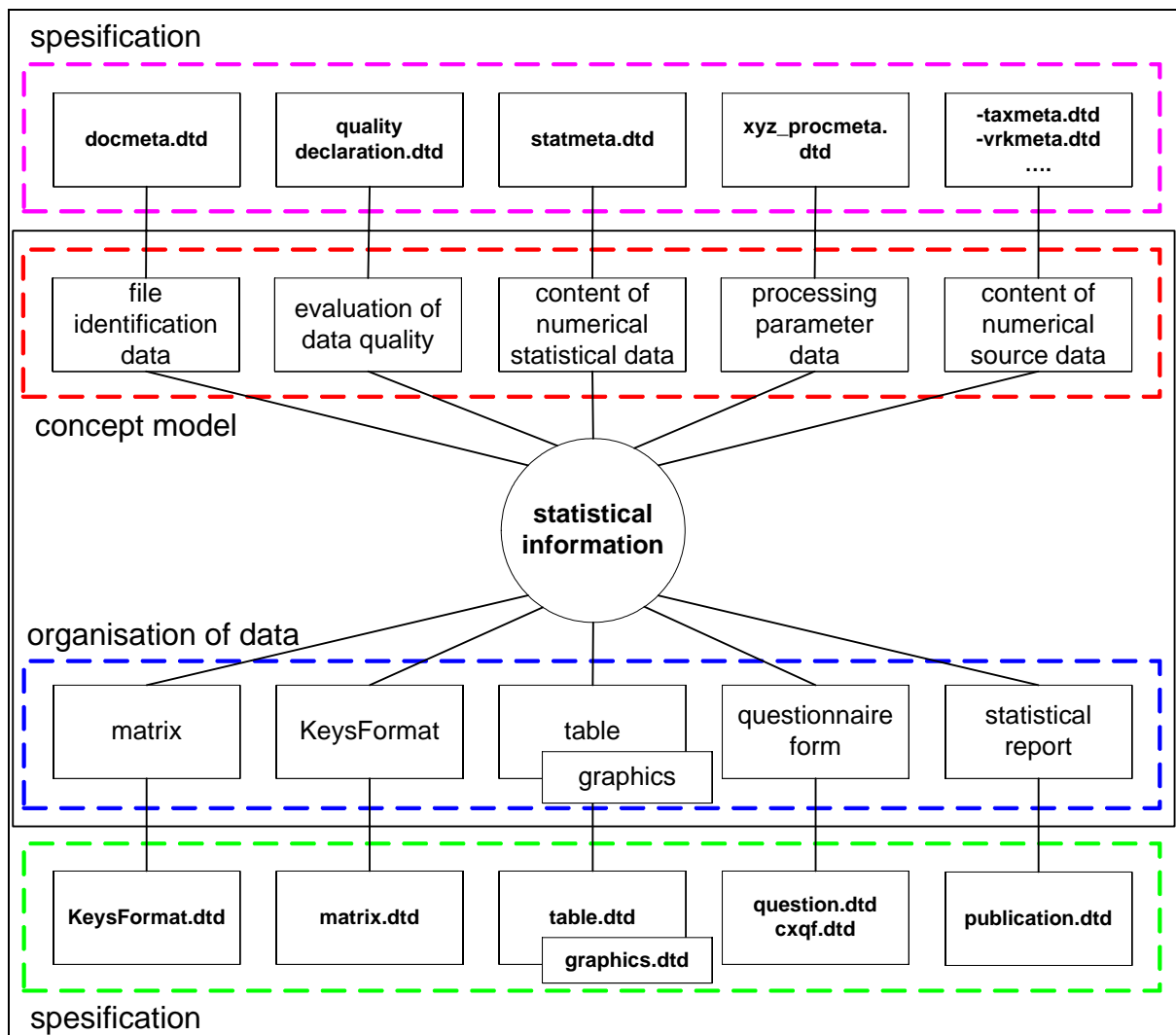


Figure 1. Common Structure of Statistical Information (CoSSI) – parts and entity

5. Basic models for organising statistical data (statistical data files, tables, etc.) are characterised by the fact that the definitions of different forms of organising statistical data allow presentation of the same information content irrespective of the form. Thus, the scope of the data content is not a criterion on the choice of organisation of data but other factors relating to the processing of the data determine the form of organisation that will serve best the production of statistics and the dissemination of statistical information in each case.

6. In CoSSI, all information describing the content, defining, etc., of produced data represent metadata. The following typology of metadata has been used as the metadata frame in CoSSI:

- (1) Statistical metadata that are content-specific and necessary for the interpretation of numerical statistical data.
- (2) Metadata relating to the identification and archiving of datafiles, which form document metadata.

⁴ On demands imposed on the hierarchy of statistical data, see Rouhuvirta, H., Structuring of statistical information and statistical metadata. Codamos 2004. Project IST-2001-38636.

(3) Metadata concerning processing, of which some belong to statistical metadata as statistical and methodological process data and some belong to the process description as technical metadata required by the used applications.

(4) Technical metadata concerning the process, which contain the technical data required by applications and the metadata used or created in the steering of the project.

7. On the one hand, data obtained from diverse sources for statistical purposes, such as descriptions of data in administrative registers, are based on the own, specific logic of each data source and, on the other, on the availability of data and on the possibility of converting the data into a form where the descriptive information can be electronically attached to the source data and thereby utilised in the production of statistics. Descriptions of source data do not as such form an independent area of their own deviating from statistical metadata, but the descriptive information of the source file is "included" in one way or another in the statistical metadata as part of the description of the content of the final statistical information.⁵

8. The metadata definitions specifying and describing the contents of statistical information have been technically gathered into the following modules in the CoSSI model:

- (1) file metadata (docmeta.dtd)
- (2) quality evaluation (qualitydeclaration.dtd)
- (3) metadata on statistical information content (statmeta.dtd)
- (4) metadata on inquiry (question.dtd)
- (5) metadata on register information (e.g. Taxmeta.dtd)
- (6) process metadata (e.g. procmeta.dtd).

9. The defining module can be used combined with each other, or as entities supplementing each other dependent of the situation and data description requirements.

II. STATISTICAL METADATA

10. In all situations, the way of processing statistical information is eventually based on the fact that, on the one hand, we have observation units, which in statistics production are also called statistical units. However, on the other hand, besides identification of the observation unit, we also have information produced with different measurement methods on the characteristic of the said unit, which we here refer to as variables for short. This structural characteristic of statistical information (data) is utilised in the CoSSI model to attach and anchor statistical metadata to a variable. Thus, the task of statistical metadata is to describe exhaustively the content and characteristics of the variable for the needs of both producers and users of statistics (see Figure 2).

11. Some of the qualitative information on statistical data describe the characteristics of a variable and some the nature of the entire statistical datafile, and the information is not overlapping or summary in all respects. The metadata relating to a datafile cannot be simplified or assigned to the quality descriptions of the variables it contains but require their own overall quality, i.e. a separate examination of the material entity formed in a certain way. Because of the information relating to quality has been divided into two components and assigned, on the one hand, to the variable insofar as it describes the quality of the variable and, on the other, to the datafile insofar as it describes its characteristics. File-specific quality evaluations are presented in quality descriptions appended to the files.

⁵ An example of how the descriptive data of an administrative register is handled in the CoSSI framework has been given in Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Final Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636. Also available on the Internet at:

http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf.

An example of how to attach the description of a register into statistical metadata has been given in Rouhuvirta, H., Conceptual Modelling Of Administrative Register Information And Xml - Taxation Metadata As An Example . UNECE Work Session on Statistical Data Editing, Ottawa 2005. Also available on the Internet at:

<http://www.unece.org/stats/documents/2005/05/sde/wp.3.e.pdf>.

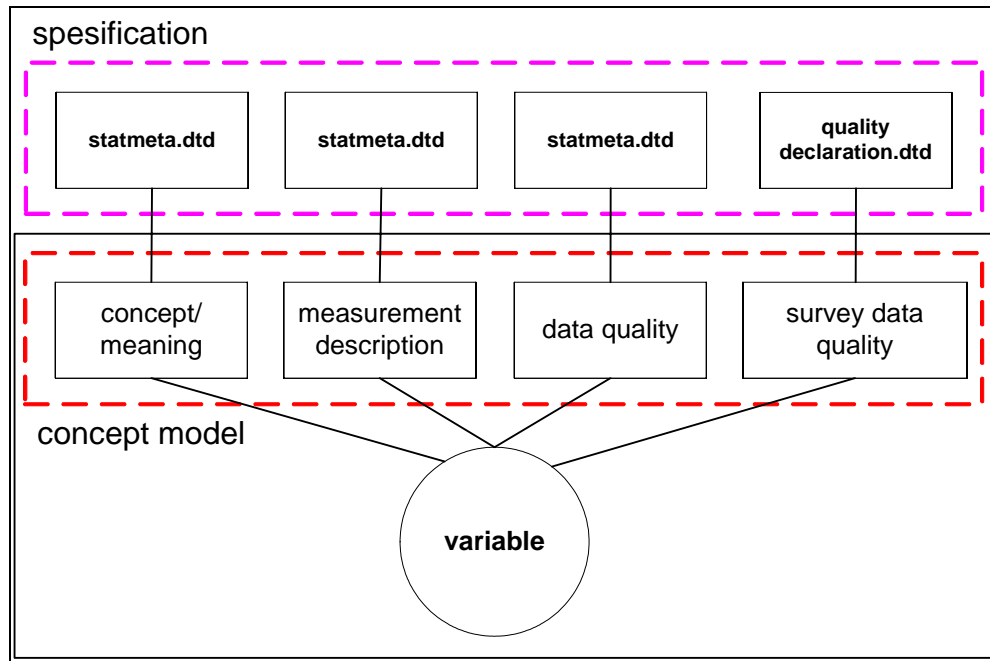


Figure 2. Components of the concepts of statistical metadata and their definitions in CoSSI

12. Variable-centredness also brings the practical benefit that the same metadata description can be used unchanged, and even in the same physical format, in different production stages and in all forms of organising statistical data. This way, many adaptations of the syntax or structure of metadata can be avoided, which might otherwise be necessary for productional reasons.

13. Variable-centredness is a foundation that ensures that metadata are transferred with data to wherever the data are transferred to during statistics production. Irrespective of how the measurement values of measured observation characteristics are handled in different stages of statistics production, the metadata remain the same provided the statistical data themselves are not manipulated in a way that affects their interpretation. Variable-centredness is also a basis whereby descriptions of the contents of administrative and other similar files that are used as sources of statistical data can be combined with the statistical information formed from them in production and in certain cases also with the final description of the statistical information in its dissemination.

14. The description of statistical information at the unit level is comprised of the documentation describing the data of the statistics, which contains statistical metadata by variable and a quality description that contains general methodological descriptions and quality evaluations relating to the data. The variable-specific descriptions of statistical metadata can be supplemented with application-specific process metadata descriptions, in which the technical information required by the application, such as length of data record or its number or character format, can be attached to the metadata descriptions.

III. DATA MODELS OF METADATA

15. The basic conceptual model of statistical metadata is described as a logical data model in the adjacent figure (see Figure 3). The basic, main concepts of statistical metadata relate to the conceptual defining of the content of a variable and to the defining of the measured characteristics. The meaning of a variable is described in a conceptual definition and the matters relating to the measurement in the operational definition of the variable. If the variable is a summary one or one formed in some other manner, the formula used in its formation can be attached to the description of the metadata.

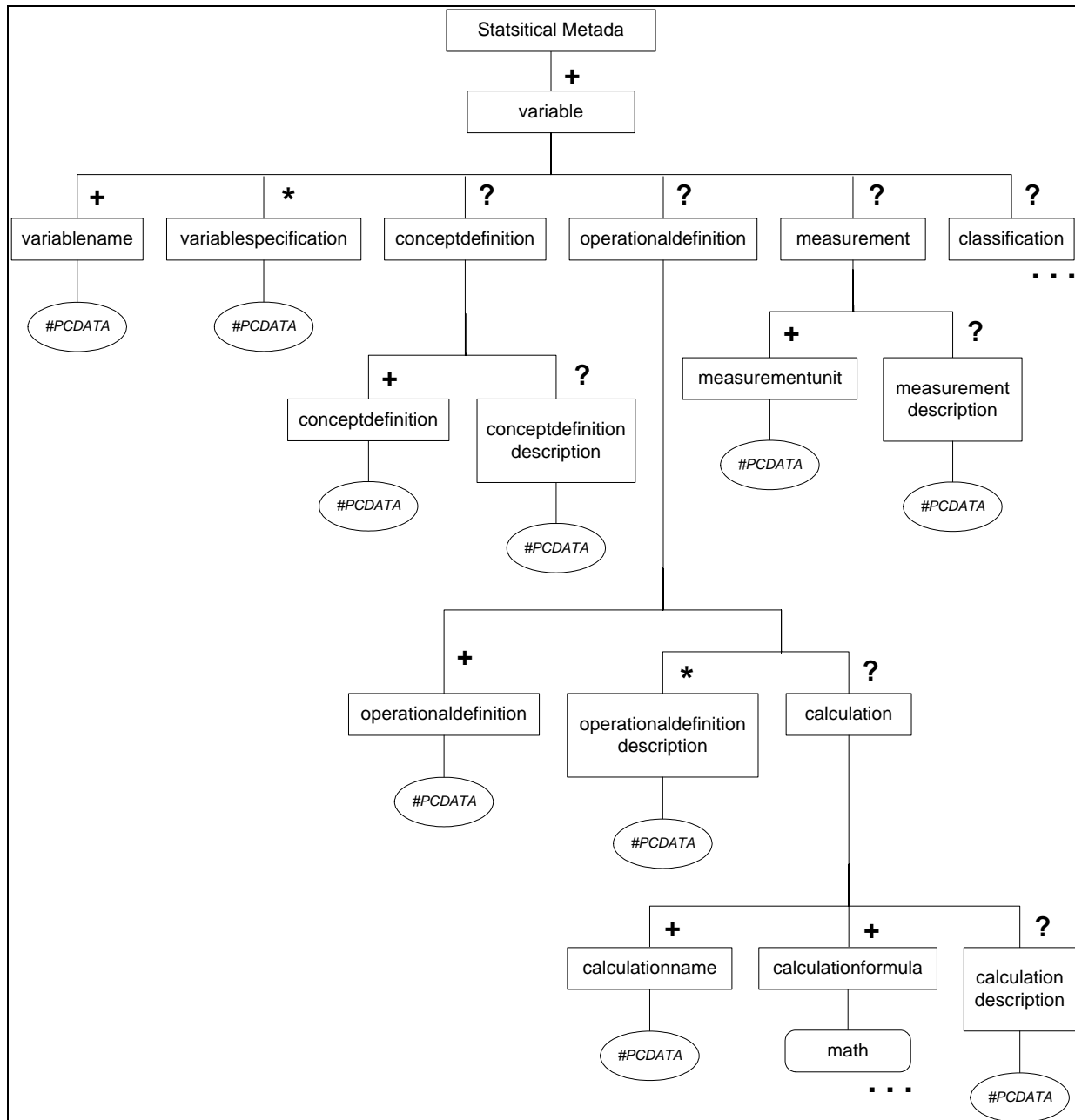


Figure 3. Logical data model of statistical metadata⁶

16. The main elements of statistical metadata descriptions and their purposes of use are presented in Table 1.

17. After initial implementation of the definition of the basic elements of statistical metadata it has become clear that the set of concepts relating to statistical metadata must be enlarged by both conceptualisation of the qualitative data on individual variables and the data that are necessary to steer the processing of statistical data. The steering data assist the communication of the professionals working in production and the realisation of division of responsibilities.

⁶ The logical models presented in this context are indicative and detailed, normative data models have been described in the CoSSI definition, see Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.91, Statistics Finland 2003. Also available on the Internet at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_en.html/cossi_definition_descriptions_v_09_2003.pdf.

Table 1. Basic elements of the concept model of statistical metadata and their purposes of use

Item	Purpose of use
variablename	The name of a variable. Variable name element is used for conceptual naming of variables in natural language. Variable name is not meant to be a code or an abbreviation.
variablespecification	Variable specification is used when the naming of a variable is not enough to describe it. Variable specification element gives a more specific description of the variable.
conceptdefinition	Conceptual definition element contains the conceptual definition of a variable.
conceptdefinition description	Conceptual definition description is used when the information in the conceptual definition element is not enough to clarify the conceptual aspects of a variable.
operationaldefinition	Operational definition element contains a written operational definition of a variable.
operdefinition description	Description element of an operational definition includes a written description of the operational definition. The description is given in natural language. Description of an operational definition is used when the information in an operational definition element is not enough to clarify the operational aspects of a variable.
calculationname	Name of a calculation. If a calculation is given it must be named. It is possible to give the name of the used method without giving the actual calculation formula. The name can be a generic or a case-specified name of the method.
calculationformula	The actual calculation formula is given here in MathML format.
calculation description	Describes a calculation formula. Calculation description is used when the information in the calculation name and calculation formula is not enough to clarify the composing of a variable.
measurementunit	This element names the measurement unit of a variable. The measurement unit is given as a standardised Finnish abbreviation (at Statistics Finland).
measurement description	The description is used to clarify the measurement if the measurement unit is not clear enough.

18. Extensions made at the first stage to the logical data model of statistical data to serve the said purposes are presented in the adjacent figure (see Figure 4)⁷.

19. The new elements attached to a variable for production purposes are:

-field name

Technical identifier: Technical identifier enables the use of short names of variables in different information technology environments where the use of long, natural names is not always possible because of technical reasons. In addition to, and separate from this, a variable has a universal identifier (ID).

-survey id

Survey identifier: Statistics departments collect data for more than one statistical survey simultaneously and/or from different "versions" of a variable. To help identification in production, a variable can be given a survey identifier in the form of a set of characters.

-variable code

Variable code: Variable code facilitates denoting hierarchical variables so that the code can be used in data processing and output. An example would be the long list of income categories in income distribution statistics, in which income variables have been given numerical codes with which different income concepts can be summed up semi-automatically. In this respect, the numerical codes of the income categories could be compared to the sets of codes used in regional classifications. The numerical codes of the variables are included as separate elements in the model because their purpose of use is different from that of the technical or ID codes.

-derive rules

Derive rule: Derive rule is a productional element into which the compilers of statistics can record in their own way in statistical jargon the rule by which a variable is formed.

⁷ The extensions shown in this context are due to be implemented into version 2.0 of the CoSSI definition during 2006.

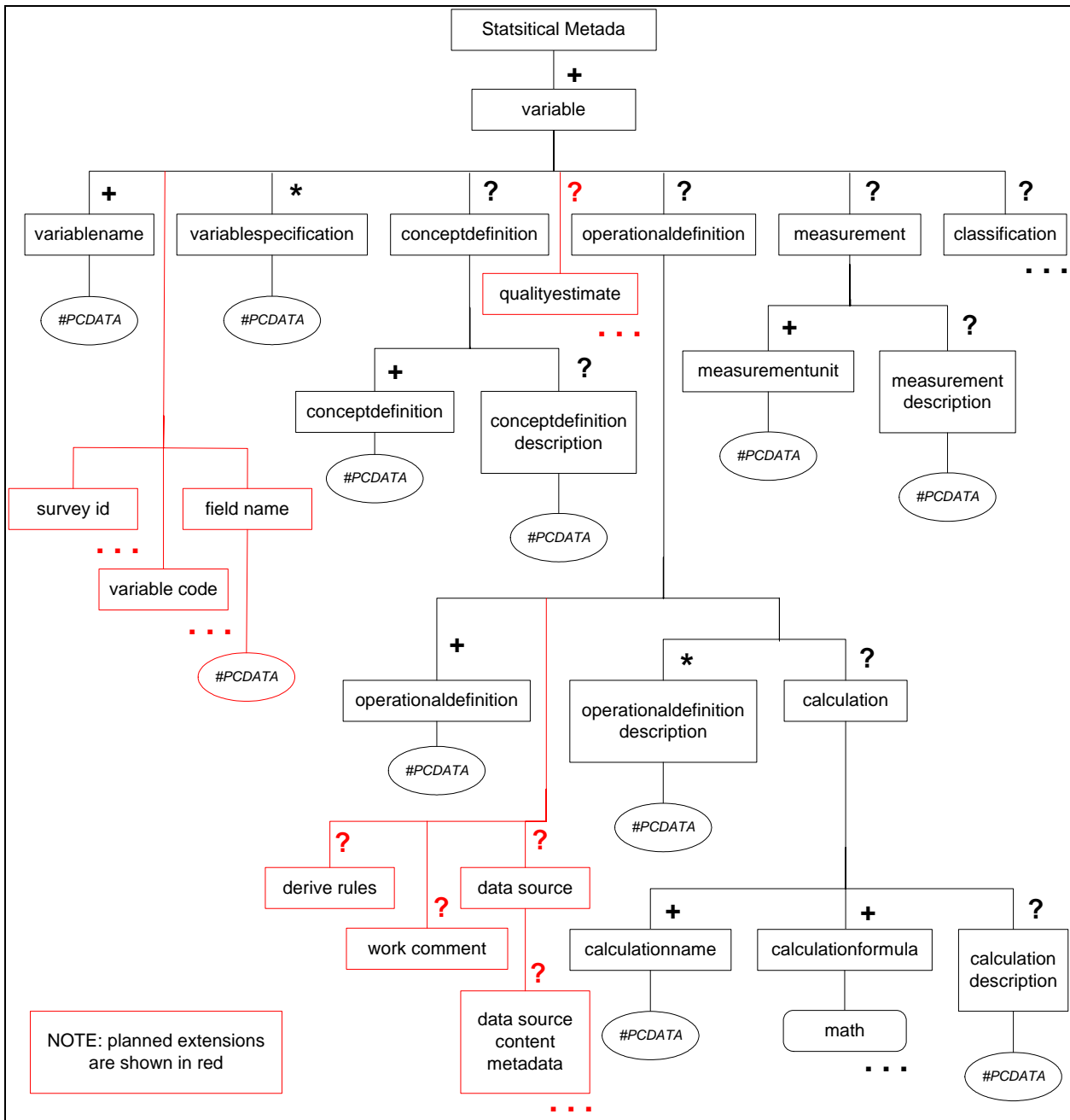


Figure 4. Extensions of the logical data model of statistical metadata

The intention is that these preliminary expressions will be used to develop an operative definition of a variable, which can be registered as an operational definition of the variable in terms of its accuracy and understandability while at the same time retaining the original derive rule expressed in statistical jargon for productional purposes. The derive rule functions at the same time as a common definition document for the compiler of statistics and for the application developer/programmer.

-work comment

Work comment: Work comment is intended to be used in the supervision of the work of compilers of statistics and as a production check list.

-data source

Source of data on variable: A link or direct reference can be given to a variable to either an external data source, such as an administrative register, or to a question in a question database of data collected in-house.

-metadata on the content of source data

Description of the content of source data: Description of the content of an external data source can be attached here, if the structural description of the data is known, as is the case with taxation data if they have been described according to taxmeta.dtd or if the description of a question relating to survey data collected in-house is in the format specified in question.dtd. A description complying with question.dtd contains the original question text and the values and descriptions of the reply alternatives to it.

20. The above-described data are primarily meant for production purposes, and it is not the intention to include them in systems for disseminating statistical information. Descriptive information on data sources, which in itself is public information but whose inclusion in the dissemination of statistical information is subject to a separate agreement with the original data producer is, of course, a borderline case. This procedure must be followed irrespective of the fact that the data themselves can be used for statistical purposes. In an ideal case, description of the input data can be included as part of statistical metadata when understandably expressed.

21. Once the data are completed the purpose of the data evaluating the quality a variable is to help the users of statistical information to assess and use the statistics correctly. A preliminary data model for data evaluating variable quality is presented in the adjacent figure (see Figure 5).

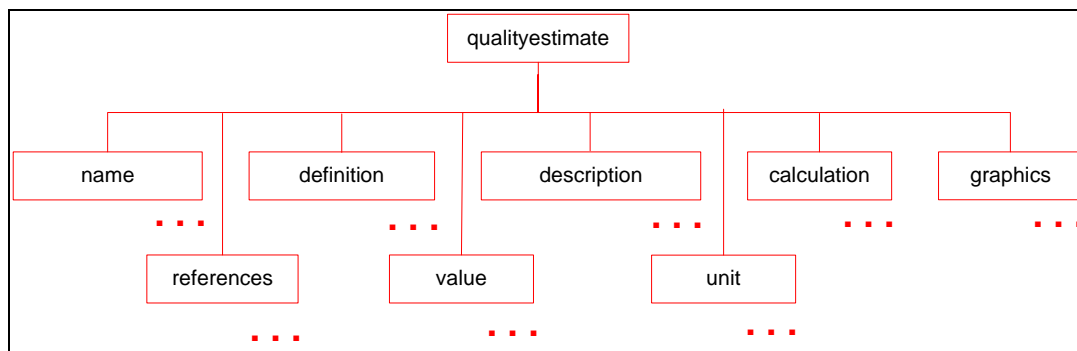


Figure 5. Extension of the logical model of statistical metadata – quality evaluation of variable

22. The quality evaluation data attached to a variable comprise the following elements:

- quality estimate/quality index/indicator [quality judgement],
- name of parameter,
- method or formula for calculating parameter,
- definition of description of parameter,
- reference to possible methodological source,
- description or interpreting instructions,
- calculation result/value/result value,
- calculation unit of parameter and
- graphic depiction or presentation of result or result value.

23. The central component of statistical metadata is description of the classification of the variables. In the concept model, matters relating to used classifications have been described in two ways. On the one hand, the used classification standard can be identified or, alternatively, the used category values and their importance can also be described. The entire element concerning classification in the concept model is presented in the adjacent figure (see Figure 6).

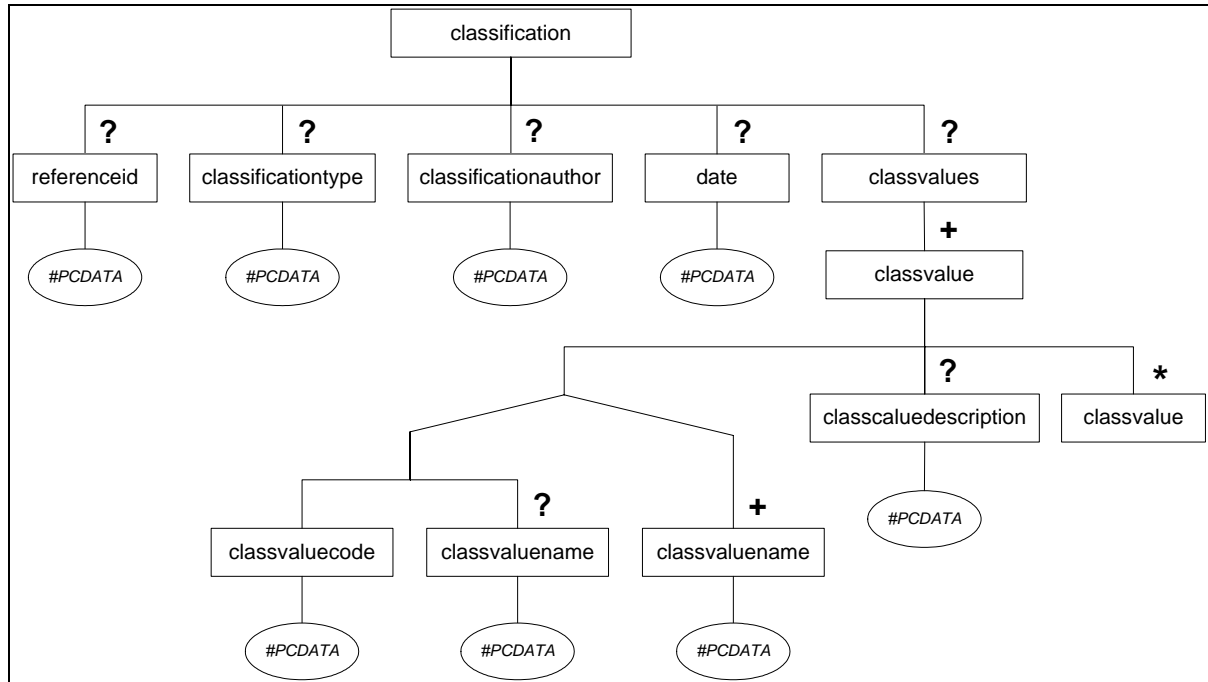


Figure 6. Logical data model of statistical metadata - classifications

24. The purposes of use for the metadata elements describing statistical classification data are presented in Table 2.

Table 2. Basic elements concerning statistical classifications in the concept model of statistical metadata and their purposes of use

Item	Purpose of use
classification	The classification element of statistical metadata can be used in two ways. First, when a standard classification is applied the element can be used to give the identification data (codes, names, etc.) of the classification. Second, the element can be used to describe an own classification for which general identifiers need not be given
referenceid	Reference identifier of the used standard classification. Classification identifier can be used as a direct link to a classification.
classificationtype	Classification type is used for descriptive purposes of standard classifications.
classificationauthor	Author of the classification.
date	Date when classification was published.
classvalue	Collecting element for all information of the class value. Hierarchy level of the class value is described with a hierarchy level attribute. Description of this element's content model: The obligatory information that must be given on the class value is either its code or its name. Both can also be used simultaneously. A separate description can be attached to the class value. The class value may also contain subclasses that may contain further subclasses. At the moment the hierarchy is limited to six levels
hierarchylevel	Attribute describes at which level in the classification hierarchy this item is.
classvaluecode	Code of a class value
classvaluenam	Element contains the name information of a class values in textual format.
classvalue description	Class value description is used when the information in a class value name element is not enough to clarify the conceptual issue of the class value.

25. After the first implementation of the definitions of the basic concepts of classification data listed above it has become quite clear that the classification description must be extended to allow both work comments recorded for production purposes and descriptions of the formal data related to the formation of classifications. The planned additions are presented in Figure 7.

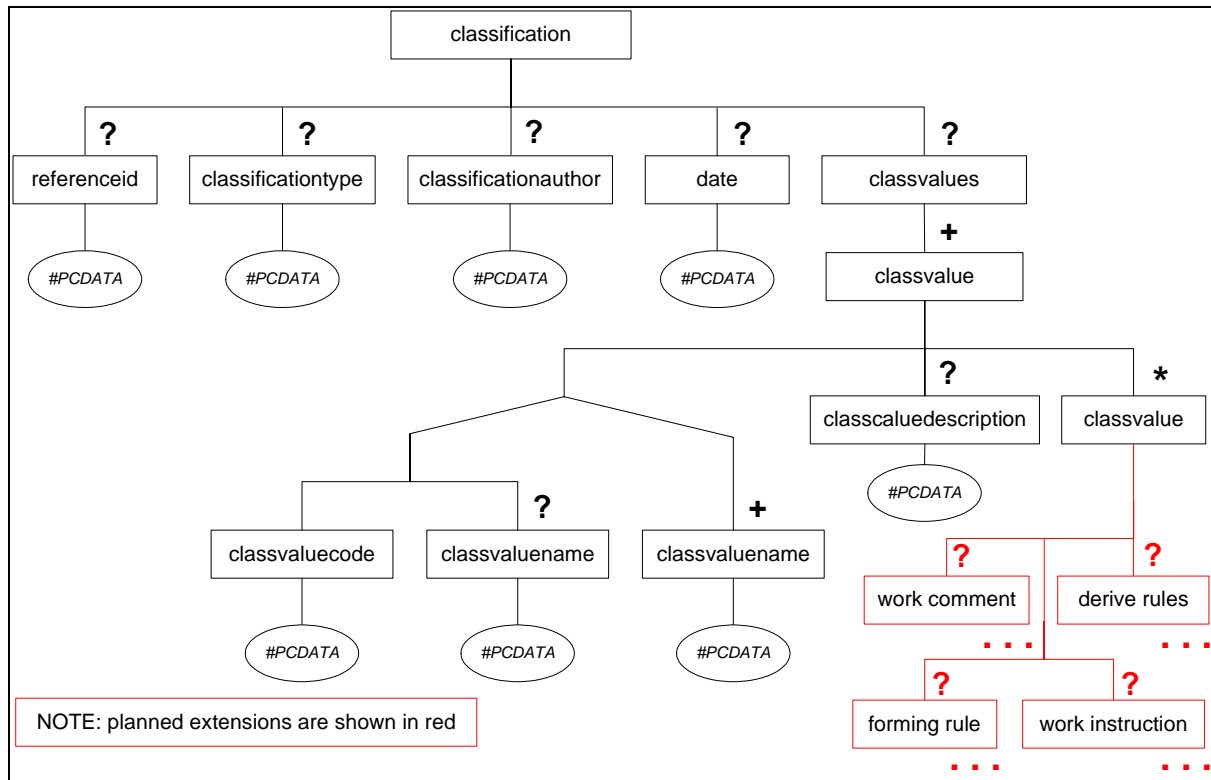


Figure 7. Extensions concerning classification data in the logical model of statistical metadata

26. The elements to be added to classification descriptions are:

-derive rules

Category value formation rule: The formation rule can contain as data the formation rule expressed as a formula together with its description.

-forming rule

Inference rule: Inference rule is a formation rule formulated for production purposes, which is often expressed in statistical jargon, but forms a basic common document for the compiler of statistics and for the application developer/programmer. Final category description is formed from the inference rule, but inclusion of description of the inference rule as a separate element has been retained because this way it can also be documented and stored separately.

-work comment

Work comment: Work comments are intended to be used in supervising the work of statistics compilers and as a check list in production.

-work instruction

Work instruction: Work instructions are intended for the documentation and storing of formal algorithms (at Statistics Finland these are referred to as ADP rules on the forming of a category value or similar).

27. The quality description of a datafile follows the approved manner for producing quality descriptions at Statistics Finland. The concept model of the planned quality description is presented in Figure 8.

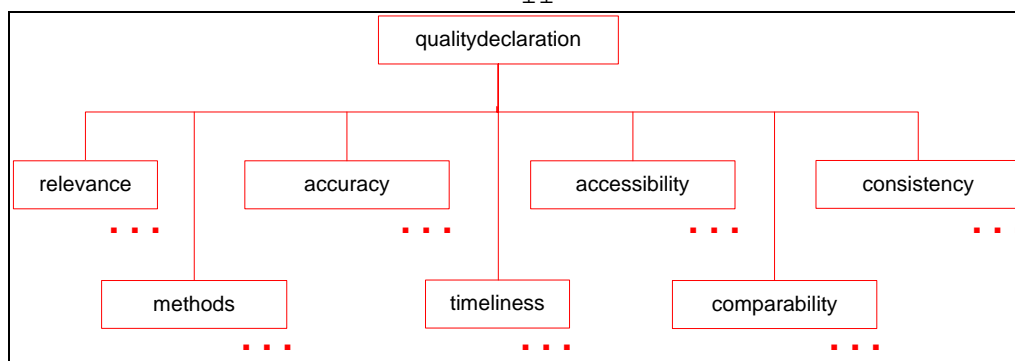


Figure 8. Logical data model of quality description

28. Purposes of use for the central concept elements of a quality description are presented in Table 3.

Table 3. Basic elements of the concept model of a quality description and their purposes of use

	Item	Purpose of use
1.	relevance	Relevance of statistical data
1.1.		Produce a detailed summary of the product's information content and end use. Identify the phenomenon that this set of statistics is designed to describe and explain its history.
1.2.		Introduce concepts that are important to understanding the statistics, classifications used or object of study, to identifying the data collector and informants.
1.3.		Describe the any acts, decrees and recommendations upon which the statistics are based.
1.4.		Assess the relevance of the statistical information produced in relation to customer needs, and how any changes in the phenomenon concerned have been taken into account in compiling the statistics.
2.	methods	Description of the methods used in statistical surveys
2.1.		Describe methods precisely, e.g.the methods applied, i.e. the population of the statistics, the materials used, the survey design (census survey or sample survey), the sampling design, data collection method, editing, imputation, the use of weighting coefficients in sample surveys and estimation methods required by the final results
2.2.		Justify the methods used and any changes made (including an assessment of the impacts of those changes upon time series).
2.3.		Methods descriptions identify the data sources used in statistics production (also for auxiliary information).
2.4.		Review the whole process of statistical survey.
3.	accuracy	Accuracy of information
3.1.		Demonstrate that the statistics measure what they are supposed to measure.
3.2.		Report on all facts that may have a bearing on the reliability of the statistics. Also mention key uncertainty factors, i.e. possible sampling and non-sampling errors.
3.3.		Estimate the correspondence between the target population and the population of interest and the quality of the sampling frame used.
3.4.		Describe the main uncertainty factors, i.e. possible sources of error, and assess their impacts on the estimates published: <ul style="list-style-type: none"> – Sampling errors, – Non-sampling errors: <ul style="list-style-type: none"> – Coverage error, – Measurement error, – Processing error, – Non-response error.
3.5.		Using the main classifications employed in the statistics, tabulate statistical parameters for the estimates, such as standard deviations that take the sampling design into account, mean square errors (MSE) and parameters

		estimating the efficiency of the sampling design (deff)
3.6.		Interpret tables produced in 3.5.
4.	timeliness	Timeliness and promptness of the information published
4.1.		Indicate the point of time or period that the statistics describe.
4.2.		Indicate whether the information is preliminary or final.
4.3.		Where necessary examine how time series data have changed over time (e.g. on account of seasonal adjustment).
5.	accessibility	Accessibility and clarity of information
5.1.		For statistics where the data constitute comparable time series, indicate the length of the time series available.
6.	comparability	Comparability of statistics
6.1.		Describe the comparability of the statistics over time and with other materials.
6.2.		Examine changes that have affected comparability and their significance, e. g. in the statistics production process, survey design concepts and classifications.
7.	consistency	Consistency
7.1.		Assess the consistency of the statistics in comparison with other statistics on the same subject. For example, examine differences in their concepts and data collection processes and assess their impacts.

29. From the point of dissemination of statistical information and exploitation of statistical datafiles, essential metadata are also contained in the metadata concerning individual files. The metadata defining a file contain information on its producer, subject, identifiers, etc. (see Figure 9.).

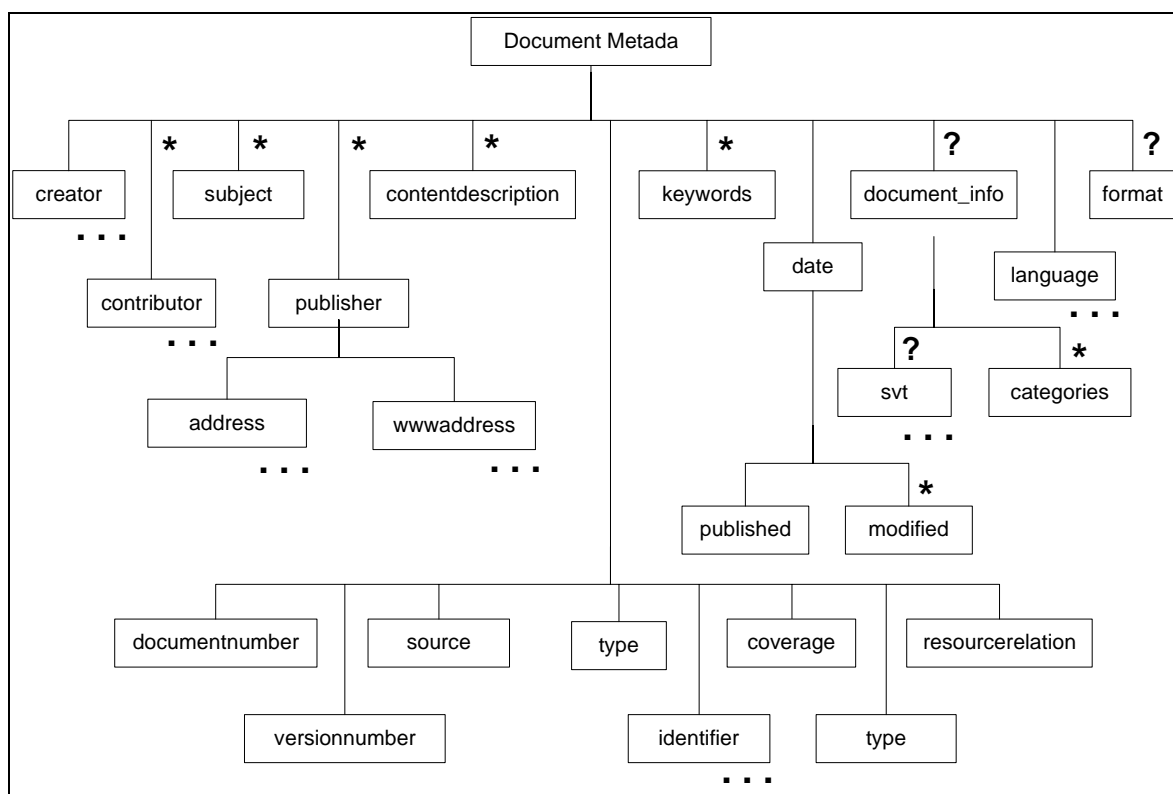


Figure 9. Logical data model of document metadata

30. Purposes of use for the central concept elements of the document metadata are presented in Table 4.

Table 4. Basic elements of the concept model of document metadata and their purposes of use

creator	Equivalent to Dublin Core element Creator: An entity primarily responsible for making the content of the resource. Examples of Creator include a person, an organisation or a service. Typically, the name of a Creator should be used to indicate the entity. At Statistics Finland Creator is a logical element and the name of the creator is given in the child element person. In statistical publications the creator is the person who can give more information about the concerned topic. In other publications (i.e. research reports) the creator is the person responsible for the text. There can be only one creator for any one document. If other persons have contributed to the making of the document, they are mentioned in the contributor element. When information about the creator is copied to the bibliographical information system using Dublin Core, and the document constitutes Official Statistics, then Statistics Finland uses its organisational name as the creator.
subject	Equivalent to Dublin Core element Title: A name given to the resource. Typically, Title will be a name by which the resource is formally known. At Statistics Finland Subject is the title text of a document. It means that the subject element in this module repeats the title element of the parent DTD (e.g. publication.dtd or table.dtd). For example, if a document is a table (table.dtd), then subject is the content of the title element of the table document. When subject information is copied to the bibliographical information system using Dublin Core, then the subject is the same as the title.
keywords	Equivalent to Dublin Core element Subject and keywords: Topic of the content of the resource. Typically, Subject will be expressed as keywords, key phrases or classification codes that describe the topic of the resource. Recommended best practice is to select a value from a controlled vocabulary or formal classification scheme. At Statistics Finland the first keyword in an Official Statistics publication is the category of the document, so it is repeated here. Otherwise (e.g. in research reports), keywords can describe the publication's frame of reference.
contentdescription	Equivalent to Dublin Core element Description: An account of the content of the resource. Examples of Description include, but is not limited to: an abstract, table of contents, reference to a graphical representation of content or a free-text account of the content. Not currently used by Statistics Finland's system architecture. This is an optional element for possible future needs.
publisher	Equivalent to Dublin Core element Publisher: An entity responsible for making the resource available. Examples of Publisher include a person, an organization, or a service. Typically, the name of a Publisher should be used to indicate the entity. Use at Statistics Finland Publisher is a logical element for connecting Statistics Finland's organisational information to the Dublin Core publisher information. The name of the publisher, in different languages, is always Statistics Finland. With statistics produced by other statistical organisations, the publisher is the organisation concerned.
contributor	Equivalent to Dublin Core element Contributor: An entity responsible for making contributions to the content of the resource. Examples of Contributor include a person, an organization, or a service. Typically, the name of a Contributor should be used to indicate the entity. Use at Statistics Finland Contributor is a logical element and the information about a contributor is given in the child element person. Other persons besides the creator who have contributed to the making of the document. In Official Statistics this means the secondary source of additional information.
date	Equivalent to Dublin Core element Date: A date of an event in the lifecycle of the resource. Typically, Date will be associated with the creation or availability of the resource. Recommended best practice for encoding the date value is defined in a profile of ISO 8601. At Statistics Finland Date is written in the form DD.MM.YYYY according to the Finnish standard.
type	Equivalent to Dublin Core element Type: The nature or genre of the content of the resource. Type includes terms describing general categories, functions, genres, or aggregation levels for content. Recommended best practice is to select a value from a controlled vocabulary (for example, the DCMI Type Vocabulary). To describe the physical or digital manifestation of the resource, use the FORMAT element. At Statistics Finland Types of document are: table, publication (Official Statistics), statistical news, figure, article, research report, handbook or statistical metadata document. If the type of document is research report, then the identification number of the document is given in the element documentnumber.
format	Equivalent to Dublin Core element Format: The physical or digital manifestation of the resource. Typically, Format may include the media-type or dimensions of the resource. Format may be used to identify the software, hardware, or other equipment needed to display or operate the resource. Examples of dimensions include size and duration. Recommended best practice is to select a value from a controlled vocabulary (for example, the list of Internet Media Types defining computer media formats). In Statistics Finland's systems the format is either XML or XHTML. Format is currently only used in electronic publications.
language	Equivalent to Dublin Core element Language: A language of the intellectual content of the resource. Recommended best practice is to use RFC 3066 which, in conjunction with ISO639, defines two- and three primary language tags with optional subtags. Examples include "en" or "eng" for English, "akk" for Akkadian, and "en-GB" for English used in the United Kingdom. At Statistics Finland this element is a logical element for the main language and other language elements.

document_info	Document_info is a logical element that contains Statistics Finland-specific metadata information. Series and category of Official Statistics are covered here.
identifier	Equivalent to Dublin Core element Identifier: An unambiguous reference to the resource within a given context. Recommended best practice is to identify the resource by means of a string or number conforming to a formal identification system. Formal identification systems include but are not limited to the Uniform Resource Identifier (URI) (including the Uniform Resource Locator (URL)), the Digital Object Identifier (DOI) and the International Standard Book Number (ISBN). This is a logical element for the different identifiers of a document. (URN URL ISBN ISSN DOI documentnumber)
URN	Uniform Resource Names. Currently not in use at Statistics Finland.
URL	Uniform Resource Locator. Fixed URLs facilitating source references should be used at Statistics Finland.
ISBN	International Standard Book Numbers. Statistics Finland uses its own ISBN series.
ISSN	International Standard Serial Number. Statistics Finland uses its own ISSN series.
DOI	Digital Object Identifier. Currently not in use at Statistics Finland.
documentnumber	Number issued for a publication to identify it in a certain series. If a publication B32constitutes Official Statistics, then it uses its own series number, which is described in the element number. Statistical news have an identifier in the form of "year:number" (e.g. 2002:157) and research reports belonging to a given series have their specific number sequence.
rights	Equivalent to Dublin Core element Rights:Information about rights held in and over the resource. Typically, Rights will contain a rights management statement for the resource, or reference a service providing such information. Rights information often encompasses Intellectual Property Rights (IPR), Copyright, and various Property Rights. If the Rights element is absent, no assumptions may be made about any rights held in or over the resource. Use at Statistics FinlandThe name of the organisation that has the rights to the content of the document.
coverage	Equivalent to Dublin Core element Coverage:The extent or scope of the content of the resource. Typically, Coverage will include spatial location (a place name or geographic coordinates), temporal period (a period label, date, or date range) or jurisdiction (such as a named administrative entity). Recommended best practice is to select a value from a controlled vocabulary (for example, the Thesaurus of Geographic Names) and to use, where appropriate, named places or time periods in preference to numeric identifiers such as sets of coordinates or date ranges. Currently not in use at Statistics Finland.
resourcerelation	Equivalent to Dublin Core element Relation:A reference to a related resource. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system. Currently not in use at Statistics Finland.
source	Equivalent to Dublin Core element Source:A Reference to a resource from which the present resource is derived. The present resource may be derived from the Source resource in whole or in part. Recommended best practice is to identify the referenced resource by means of a string or number conforming to a formal identification system. Use at Statistics FinlandSource is reserved for tables. It identifies the statistical data matrix from which the table is produced. Not in use for publications.
person	Person is a logical element that includes a personal information set describing the creator of and contributors to the document.
organisation	Organisation is a logical element containing an information set in one language about the organisation that has produced the document.
published	Logical element for the official publishing date and time of a document. The date contained by the element will not be changed even if the document is edited further or errors in it are corrected and the document is republished. In such cases the modified date is changed.
modified	Logical element for the dates and times of modifications made to a document. When a document is edited further or corrected for errors and the document is republished, the dates of republishing are recorded in the modified element. The chronologically latest element gives the date of the most recent amendment or update.
day	Date is written in the form DD.MM.YYYY according to the Finnish standard.timeTime is written in the form HH:MM:SS
main_language	Defines the main language of a document, which is also the default language of an electronic document. Main language can be any language. The practice is to use RFC 3066 which, in conjunction with ISO639, defines two and three primary language tags with optional subtags.fi, en, se
other_language	Defines one or more secondary languages of a document. The practice is to use RFC 3066 which, in conjunction with ISO639, defines two and three primary language tags with optional subtags.fi, en, se
svt	Svt is a logical element containing serial information on Official Statistics publications (SVT). SVT is used to identify publications published in the Official Statistics of Finland (OSF) series (name and number of the series, year, version number) and to identify Official Statistics tables.

categories	Categories element describes the topic of the statistical information. The attribute gives the topic (category) as a three-letter abbreviation. The whole name of the category and its different language versions are given as entity files.
type	Type is the three-letter abbreviation describing the category of the statistical information.
first_name	Person's first name.
surname	Person's surname.
position	Person's official title at Statistics Finland.
email	Person's (or organisation's) e-mail address. Use at Statistics FinlandE-mail address is given in Finnish language in the form "firstname.lastname@tilastokeskus.fi" and in foreign languages "firstname.lastname@stat.fi". Language attribute (xml:lang) must be given according to the document languages.
phonenummer	Person's (or organisation's) phone number.
fax	Person's (or organisations) fax number.
orgname	Name of the official statistical organisation in one language. Language is defined in the parent element organisation. Tilastokeskus, Statistics Finland or Statistikcentralen
division	Name of the official statistical organisation's sub-unit (publishing sub-unit) in one language. Usually means the responsible unit that has published the statistics. Language is defined in the parent element organisation.
address	Street address of the organisation.
wwwaddress	One www-address of the organisation in one language.
seriename	The identifier part is comprised of three elements: series name, year and number. Seriename identifier is used to identify the tables, figures and publications that come within the scope of Official Statistics of Finland. At Statistics Finland, the series division of Official Statistics currently differs from statistical topics. The seriename element depicts the topic of a publication in Official Statistics serie. The possible category values are: Housing; Living conditions; Energy; Prices and costs; Government finance; National accounts; Trade; Education; Culture and the media; Transport and tourism; Agriculture, forestry and fishing; Justice; Wages and salaries; Financing; Construction; Manufacturing; Health; Science, technology and research; Income and consumption; Labour market; Foreign trade; Elections; Population; Population census; Environment and natural resources; Enterprises.
year	Year of publishing of Official Statistical information.
number	Annual sequential number within a series by document type. Formed series-specifically for all publications, tables, etc., of the series concerned. Tarkoitus käyttää ensisijaisesti taulukoiden yhteydessä. Sisältö joudutaan mahdollisesti määrittelemään uudelleen, kun painetut julkaisut häviävät.
versionnumber	This element is the version number of a document. Tarkoitettu käytettäväksi lähinnä taulukoiden ja kuvioiden kanssa.

IV. SOME CONCLUSIONS

31. According to our initial experiences, application of a structured information model to the conceptualisation of statistical metadata has opened new possibilities for exploiting the new technologies developed for easy handling of information in text format. XML can be regarded as representing such technologies. More efficient processing of text-format information facilitates management of richer statistical metadata. This aspect can be exploited equally well in the production of statistics and in the dissemination of statistical information.

32. The practical benefits brought by the application of structuring to the management of metadata include the ease with which the data model can be expanded and the extensions can be technically implemented. The editability can be utilised to develop as consistent and all-embracing content specification as possible for statistical metadata.

33. Besides the scope of its content and its flexibility, as a frame of reference for metadata the CoSSI model examined here differs quite essentially from the metadata systems conventionally used at Statistics Finland in that
- it makes it possible to change from decentralised to centralised management of metadata in which the producer of statistics can control the correctness of the metadata concerning the data material, and which can also be used in the dissemination of statistical information, and
 - when receiving numerical statistical data, in the same connection the users also receive the metadata that are essential in their interpretation, and instead of untargeted metadata in separate reference volumes or other similar sources, metadata can be presented immediately adjacent to numerical statistical data.
34. Examined from the perspective of production, the point of departure in the modelling and organisation of metadata could have been its attachment to the numerical value of data. However, the now implemented attachment of metadata to the variable instead of the data value facilitated the use of a simpler and more informative data model relative to the scope of the data content, and simplified and rationalised the management of data. Attachment of statistical metadata to the structure of statistical data through a variable is technically considerably simpler to implement than linking of metadata to individual data values in an information system, whatever the available production technology.
35. In practice, the structured model of statistical information represents a real alternative as a frame of reference for statistical metadata, both in respect of its approach and concept defining, and at the moment there appears to be no specific need to change the developed basic solution. Indeed, the needs for further development concern primarily extensions of the data content along the lines described above. Moreover, we are endeavouring to improve the functionality of the technical solutions of statistics production so as to make the use of statistical metadata effortless and easy in different stages of production. These kinds of solutions serving the production of statistics include creation of user interfaces with statistical metadata into such tabulation applications as SAS and SuperStar.

References

Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636. Available on the web at:

http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

Rouhuvirta, H., Conceptual Modelling Of Administrative Register Information And Xml - Taxation Metadata As An Example . UNECE Work Session on Statistical Data Editing, Ottawa 2005. Also available on the Internet at: <http://www.unece.org/stats/documents/2005/05/sde/wp.3.e.pdf>.

Rouhuvirta, H., Structuring of statistical information and statistical metadata. Codacmos 2004. Project IST-2001-38636.

Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003. Available on the web at: http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf

Rouhuvirta, H., Lehtinen, H., Karevaara, S., Laavola, A., Harlas, S., Demonstration Report on Taxation Metadata in Secondary Data Collection - How to connect the metadata of taxation to numeric taxation data and use them at the same time. Codacmos 2004. Project IST-2001-38636. Available on the web at:

http://www.stat.fi/org/tut/dthemes/papers/demoreport_on_taxation_metadata_codacmos_2004.pdf

**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 Slovenia¹

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 - for 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 through 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 find 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 through 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 of 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.

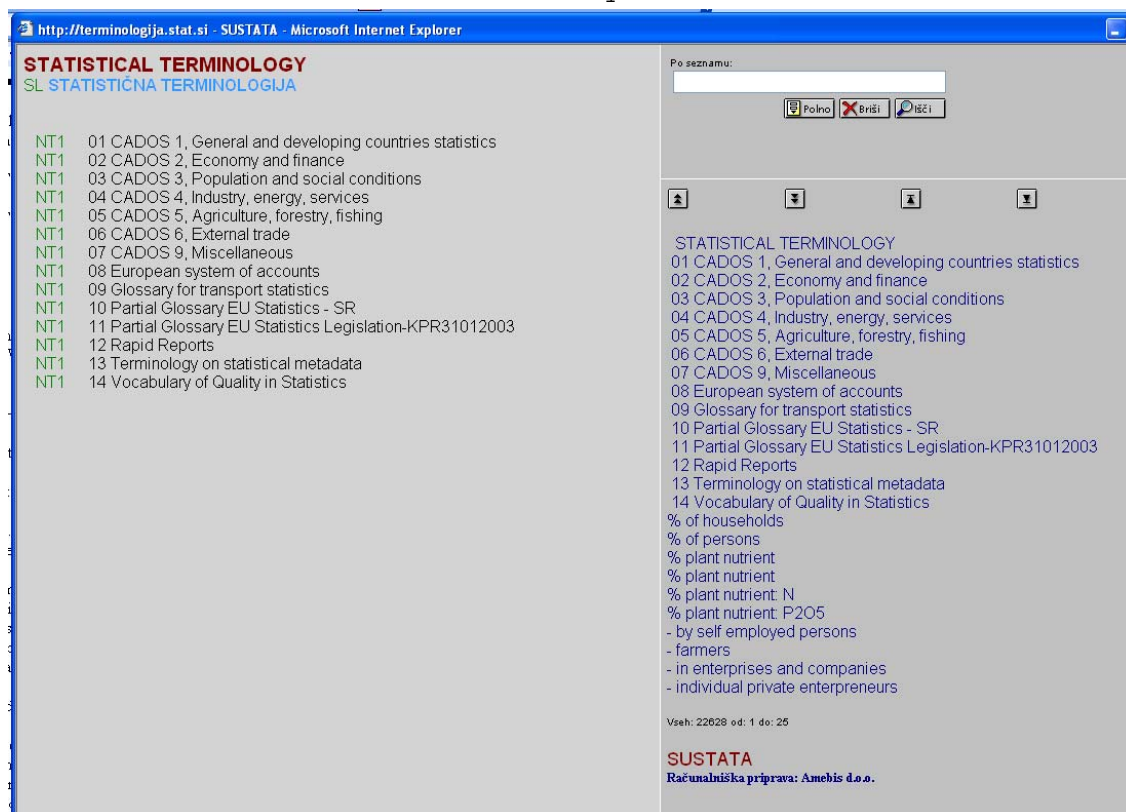


Figure 2: Statistical terminology as seen from the website

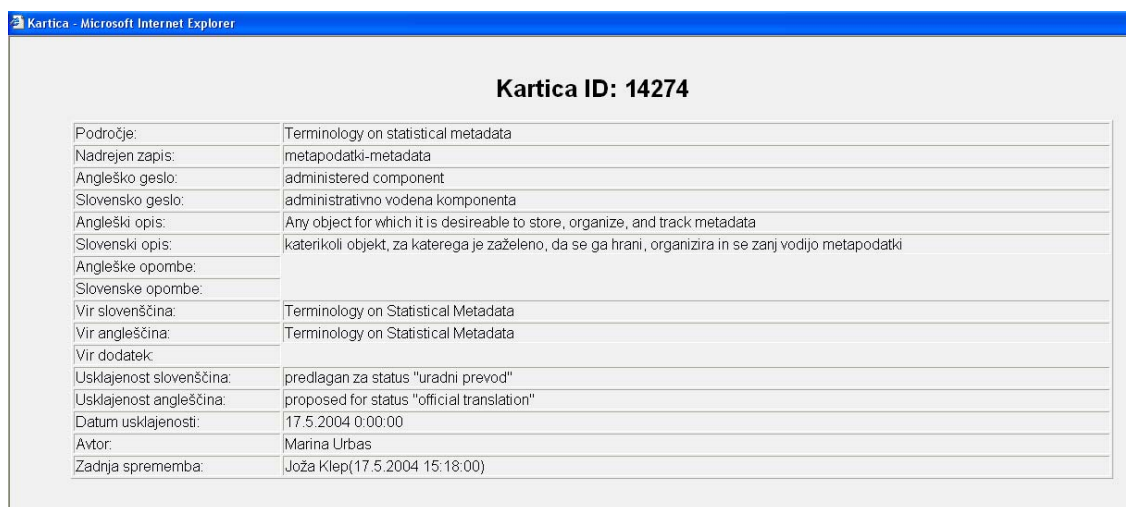


Figure 3: "Card- view" with the set of attributes for "administered component"

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).

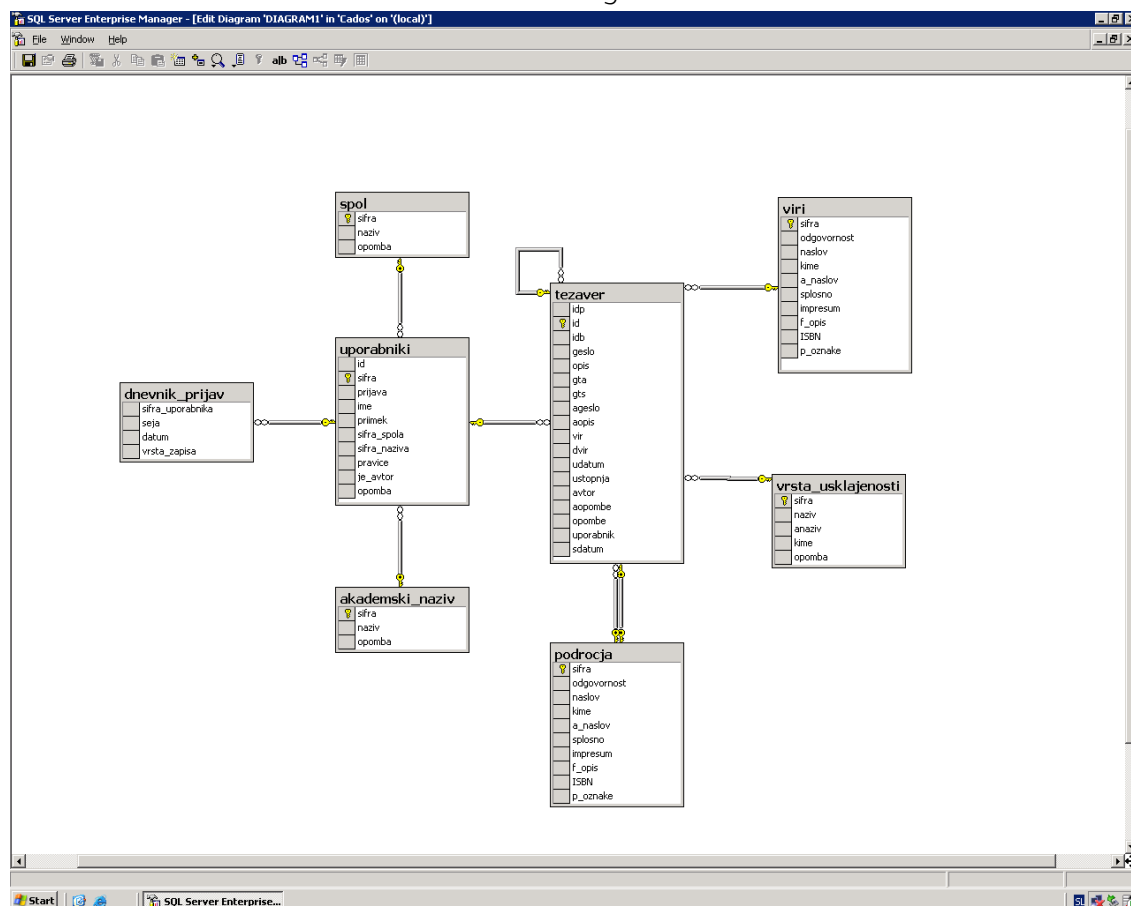


Figure 4: Database schema

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>

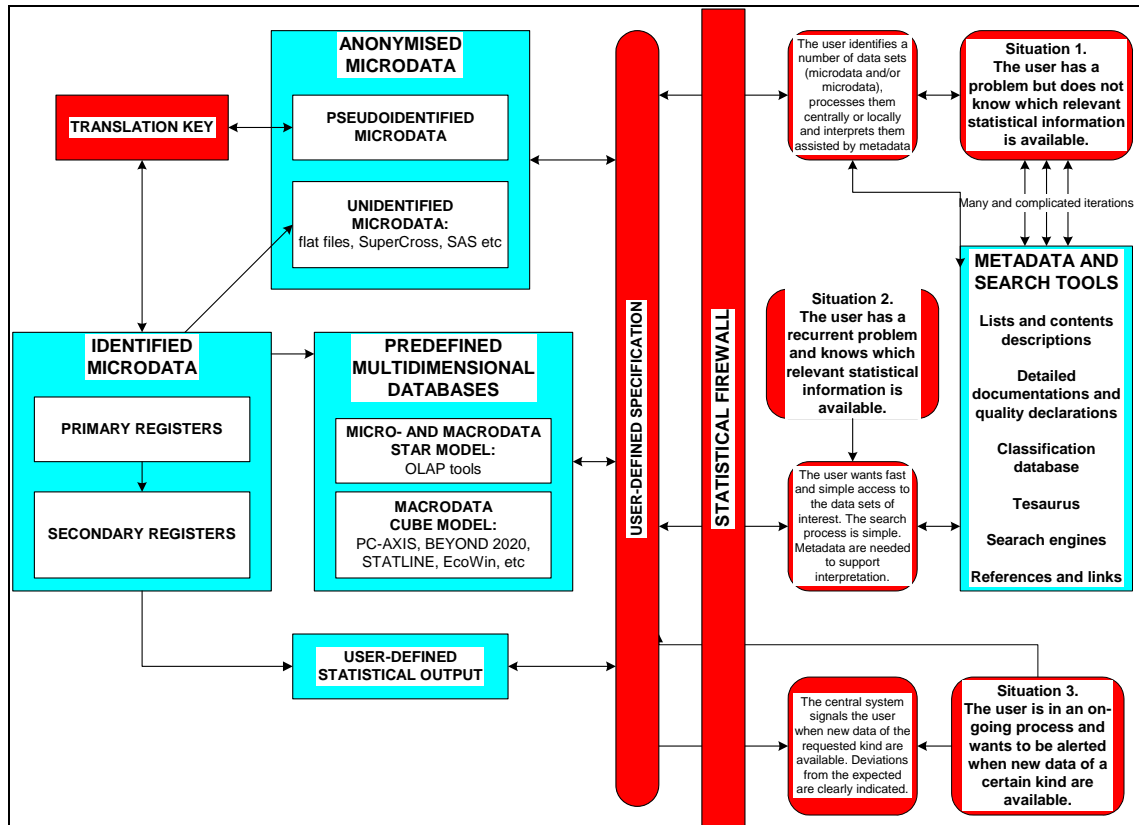


Figure 5. An example of user access to statistical data (Sundgren 2003)

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

**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

**STATISTICAL METADATA MODEL DEVELOPED IN SPAIN:
CURRENT AND FUTURE USE AND APPLICATIONS**

Supporting Paper

Submitted by Instituto Nacional de Estadística (INE), Spain¹

I. INTRODUCTION

1. This paper covers two main objectives. In a first place, the article describes the metadata model functional description and the current maintenance status. In addition to that, the core of the paper is devoted to explain the metadata model role in the Spanish National Statistics Institute.

2. The main objective of the metadata project is to build a tool in order to facilitate the integration and co-ordination of the whole information requested by INE to data providers. Generally speaking the project will allow INE to produce information more harmonised, rationalised and, specially, more comparable. Also data users will get ready a tool to get more information about every statistical operation performed by INE. Moreover, in the future this project will favour the co-ordination among the different actors of the Spanish Statistical System (INE, Ministerial Departments, Regional Statistical Offices, Spanish Central Bank...).

¹ Prepared by: Mar Blanco Frías (mblafri@ine.es) and Ana Isabel Sánchez-Luengo (anaisan@ine.es)

II. METADATA MODEL DESCRIPTION AND STATUS

3. Any model can be seen from quite different points of view. In order to give a complete view of the model, the following paragraphs explain it from the functional, management and user points of view.

A. Logical Architecture and management of the System

FIRST PHASE

4. The first phase of the project has as input the Statistical Operations performed in the Institute. To understand this phase the following definitions are needed:

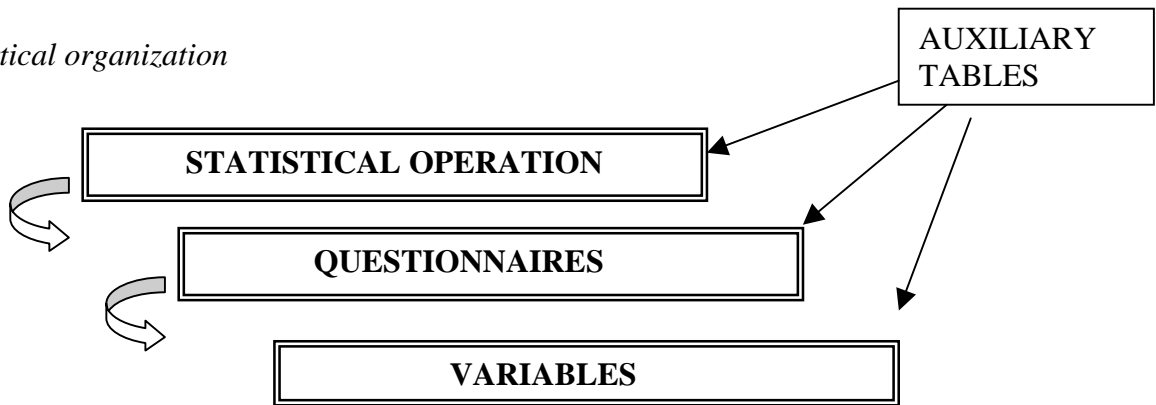
- a. *Statistical Operation*: set of activities leading to obtain statistical results about a specific sector or issue, from individually collected data using surveys or public sources.
- b. *Variable*: information about the unit being studied. From the practical point of view, and in most cases these variables use to correspond to a unique question in the questionnaire.
- c. *Classifications*: Set of classes or categories, defined using codes, which fix the values that a variable can get. When talking about an international regulation, we talk about standard classifications.
- d. *Standard Classifications*: A standardized classification. There are international standard classifications, which derive from international recommendations and Regulations, and standard national classifications, which statistical use is recommended or mandatory due to a national regulation.
- e. *Thematic classification of classifications and variables*: Internal classification at 3 levels that allows to organize information so that to make possible doing comparisons between different statistical domains.

5. The objective is to have a common repository in which all the surveys are included but not independently but establishing relations between variables and classification in different surveys by means of thematic classification, so that to be able to compare and therefore harmonise the information. The final objective is twofold:

- a. In one hand, to achieve a high degree of both internal and international harmonization that allows to perform better comparisons.
- b. On the other hand, to provide the users with a tool which allows them to achieve a better understanding of the information.

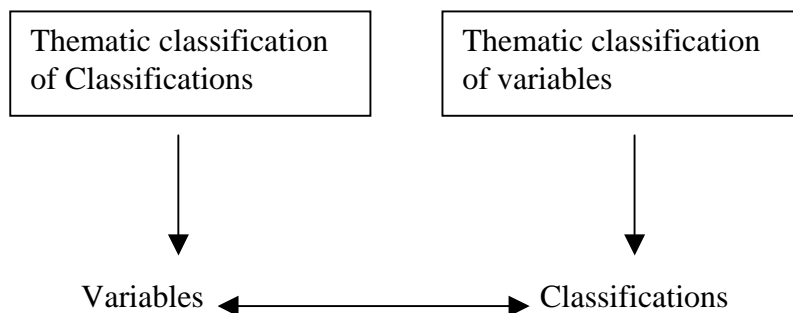
6. Functionally speaking the model has been built as a relational database supported by Oracle. This Data Base is composed by interrelated tables which follows a hierarchical structure. This hierarchical structure responds to a two types of organization of the information:

a. Vertical organization

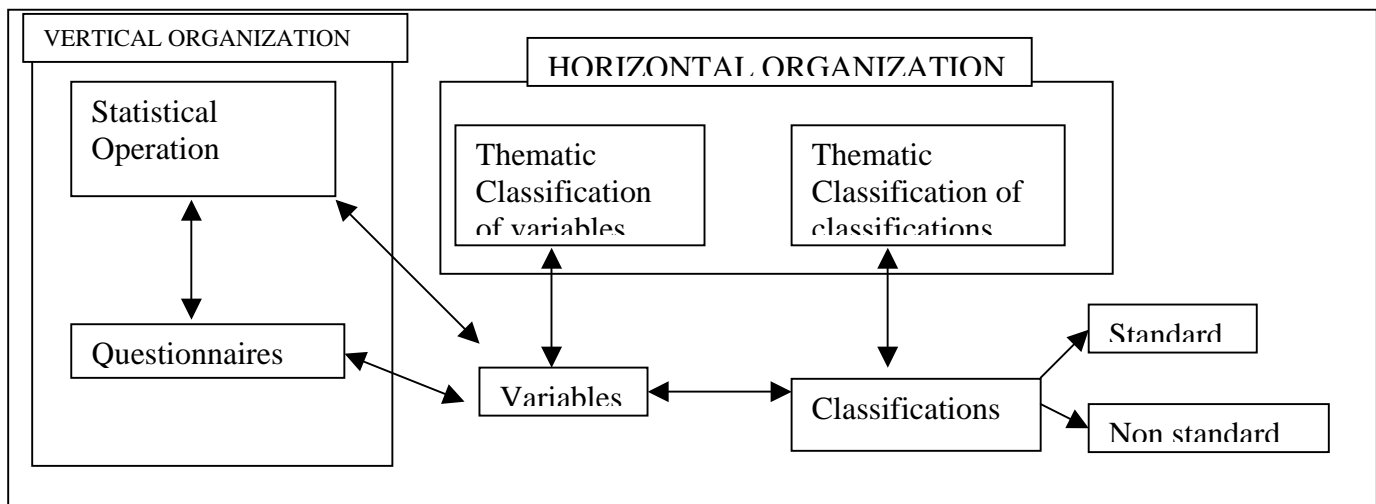


b. Horizontal Organization

7. Variables are linked to classifications. Both variables and classifications are classified thematically using 3 levels of detail. This will allow thematic searches used to compare the information in the different Statistical Operations, so that to be able to harmonize the way information is requested by the different Units. In addition to thematic searches, there exists also the possibility to search information about variables, Statistical Operations and classification by literal and by global character chains, as we will see in the next paragraph.



8. The following chart summarises the main idea of this first phase:

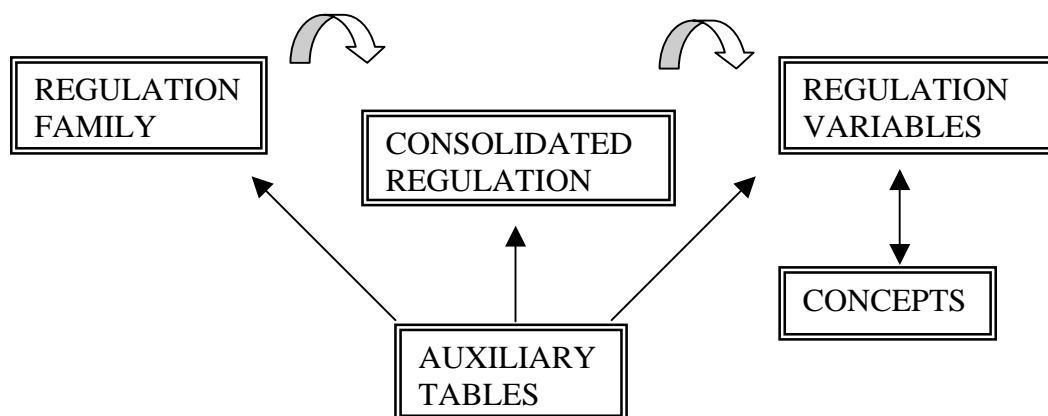


9. The maintenance of the system is carried out using a metadata management system instead of working directly on the database tables. This prevents the operator from making mistakes when introducing new information or when any modification is needed. See annex (FIGURE 2) for a picture of the main window of the management system (currently the development is available only in Spanish). The management system allows to introduce and modify information about all the elements in the system: Statistical operations, variables, classifications, and also to manage the thematic classification in the classification and variables, in a easy and friendly way.

SECOND PHASE

10. The second phase of the project is devoted to international Regulations and standardised concepts. The objective in this case is to bear a special attention to those variables included in a Regulation, among all variables. To fulfil this objective, a second relational database has been constructed, based also in Oracle, which contains in a similar hierarchical model the Regulations for each Statistical Operation and the variables contained on them. In addition to this, the concept for all the variables in the Regulations are also included in the data base, if exists. For the time being, the information is included in Spanish and sometimes in English.

11. This data base allows:
- To ensure that all variables requested in the Regulation are indeed included in the questionnaires, and in the contrary side, which variables included in the questionnaires are not requested in the Regulation.
 - To have a repository of definitions which allows the Official Statistic both national and international harmonization in concepts.
12. The following chart summarised the idea:



13. The databases constructed in the first and second phases are closely interrelated by means of a transition table, which contains the Regulation variables that are also in the Statistic Operations. The management of this database is performed directly in the tables instead of having a management system like in the first case, due to the fact that it is not expected to find frequent changes in the Regulations as they are quite consolidated. The same applies for the concepts.

B. Search Engine

14. A very important actor in the metadata project as a whole, especially from the user point of view, is the so-called "Enquire System". It is not useful to have great amounts of information stored in tables if you are not able to efficiently exploit this information as much as possible; currently the amount of information contained in the Data Base is:

System element	Number
Statistical Operations	102
Survey Questionnaires	253
Variables	8.187
Non-standardised classifications	1.434

15. Concerning the thematic classification, the following number of themes are available at each level for variables and classifications:

	Thematic classification of Variables	Thematic Classification of Classifications
# First level	28	17
# Second level	181	133
# Third level	445	535

16. The Enquire System allows the user to easily find different types of information using different criteria. It is available in the Intranet.

17. The exploitation tools have been developed following the main necessities of the INE internal users. However, it is an alive system in the sense that new tools can be developed so as to respond to new necessities.

18. The tools that compose this System are broad and varied. It allows to:

- a. Perform Thematic searches: The user is able to perform a thematic search on the classification and also on the variables. By selecting a theme (at 3 digit level), the tool shows all the classification (resp. variables) classified in that theme, coming from all the statistical operations included in the Data Base.

The user is able to select the classification(resp. variable) he/she is interested in and retrieve all the information about it:

- a "technical file" of the classification (resp. variable) containing the statistical operation and questionnaire in where is located the classification (resp. variable), if there exists a equivalent classification (resp. variable) in other statistical operations, the type of variable...
- in case of variables, the tools allows to see the classification associated with the variable (if any) and also if there is more variables associated with the same classification.

These thematic searches allow the comparison of how the different questionnaires request information about the same theme.

It is also possible to search classification or variables by character chain instead of theme.

- b. Perform Vertical searches: due to the hierarchical structure of the Database, it is possible to search starting from Statistical operations → Questionnaire → Variable. The user sees all the variables and all the classifications in the selection.

It is also possible to download the selection (the physical questionnaire as the responsible Unit spreads it).

- c. Compare Statistical Operations: By selecting two statistical operations the tool allows to study, by means of an algorithm which analyse similarity between variables and classifications:
- if they have common variables (both qualitative or quantitative)
 - if there are variables or classification classified with the same thematic classification
 - If there are common or similar classification
 - And shows all the variables in both Statistical Operations classified by themes (at 3 digit level), so that the user can see, even though there are no common variables, if there are related variables, in the sense that they are classified in the same way.
- d. Perform searches in the Regulations: Many of the variables included in the Statistical Operations performed at INE are in accordance to requests from the European Union or an international organization. This application allows to consult the variables included in the Regulations and its attributes and the relationship with the variables included in the Statistical Operations. There are two types of searches:
- By statistical Operation: it allows to see all the variables in the statistical operation and which of them are related with a variable in the associated Regulation
 - By Regulation: it allows to see all the variables in the Regulation and which of these variables are related with a variable in the corresponding Statistical Operations
- e. Perform searches in the concepts: it is possible to search a concept by literal or by theme. The source of the concept is a Regulation and also CODED (the Eurostat concepts and definitions data base).
- f. Questionnaire simulation: By selecting a specific questionnaire, it is possible to "reproduce" the questionnaire as it is stored in the data base: the user will not see the final questionnaire used by the responsible Unit, but a more technical version containing all the variables and classification and its internal attributes, internal codes, etc...

19. The following plot summarises the metadata model based on variables, concepts and Regulations:

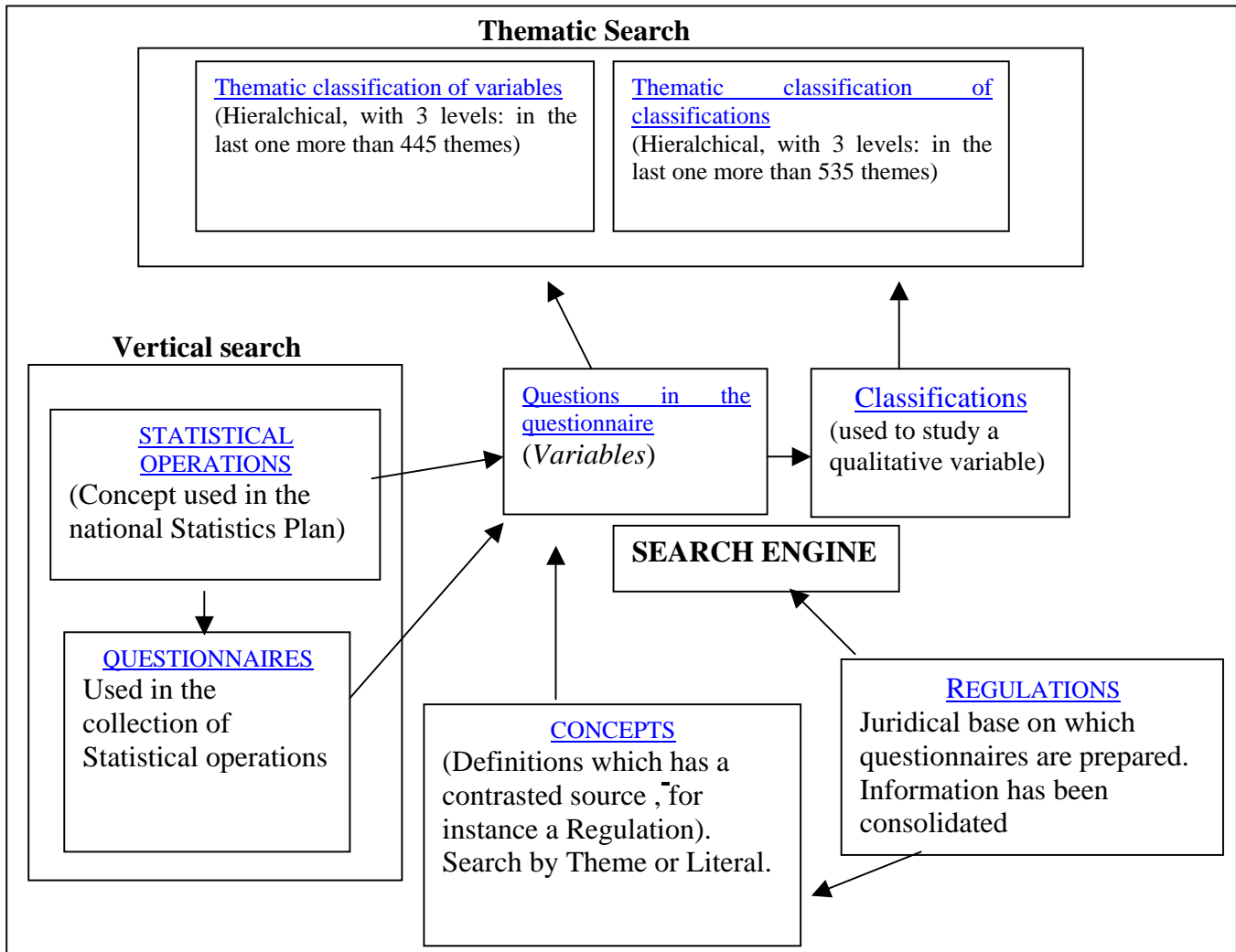


Figure 0: Metadata based on concepts, variables and Regulations

20. By means of the Enquire System, the information is organised so that its exploitation is quite complete and allows the user to have a complete picture of the whole data base and the relationships between variables, classification, Regulations...

III. ROLE OF THE METADATA INSIDE INE

21. Currently there is a growing interest in metadata inside INE, being a useful tool for all Units performing surveys and because it gives a general and complete view of the official statistics status. Nowadays, the Metadata Model is a strategy project inside INE and has much support from the Management Units.

22. Among all feasible uses we can emphasize:
- Coordination and integration of the information
 - Rationalization of the information when it comes to elaborating questionnaires
 - Integration of data into metainformation
 - To provide a tool which allows a better analysis of the information by the users
 - To provide a coordination tool between all the statistical System (INE, Ministries, and Government Units)
 - Use in International projects

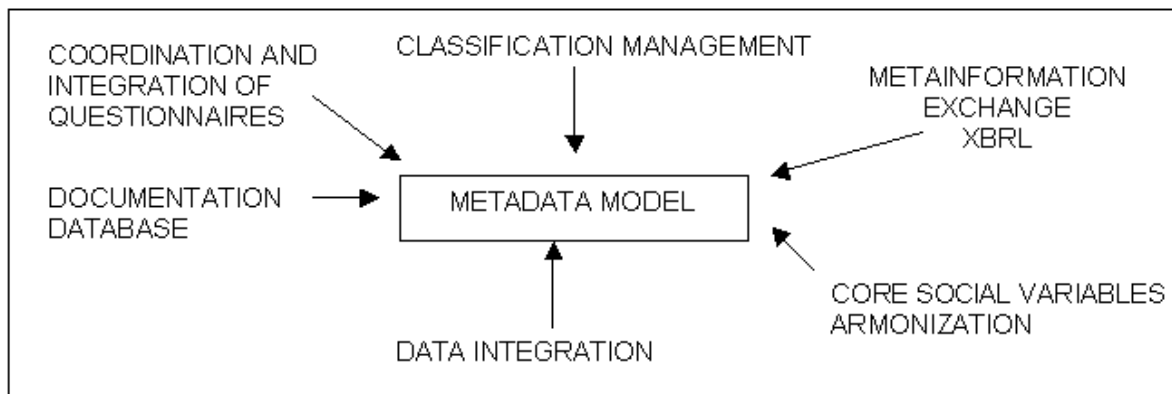


Figure 1: Metadata into INE

23. Our future works are aimed at two different goals:
- (1) On one hand we want to include not also Statistical Operations under responsibility of INE, but also those carried out by other Ministries and Government units, as well as the public registers.
 - (2) On the other hand, our intention is to link the metadata to the data (micro and macro). We are facing a main technical problem: the great amount of data we need to manage. We are talking about annual statistical operations, but also of bi-annual and monthly ones. Subsequently, the problem is quantitative rather than qualitative. A possible solution is to keep the data under control of the responsible Unit, but connecting them with the variable in the metadata model using a unique identifier.

IV. ANNEX

24. Some of the interface developed for the management and exploitation of the information is included hereafter as an example.

- **Management System main window:**

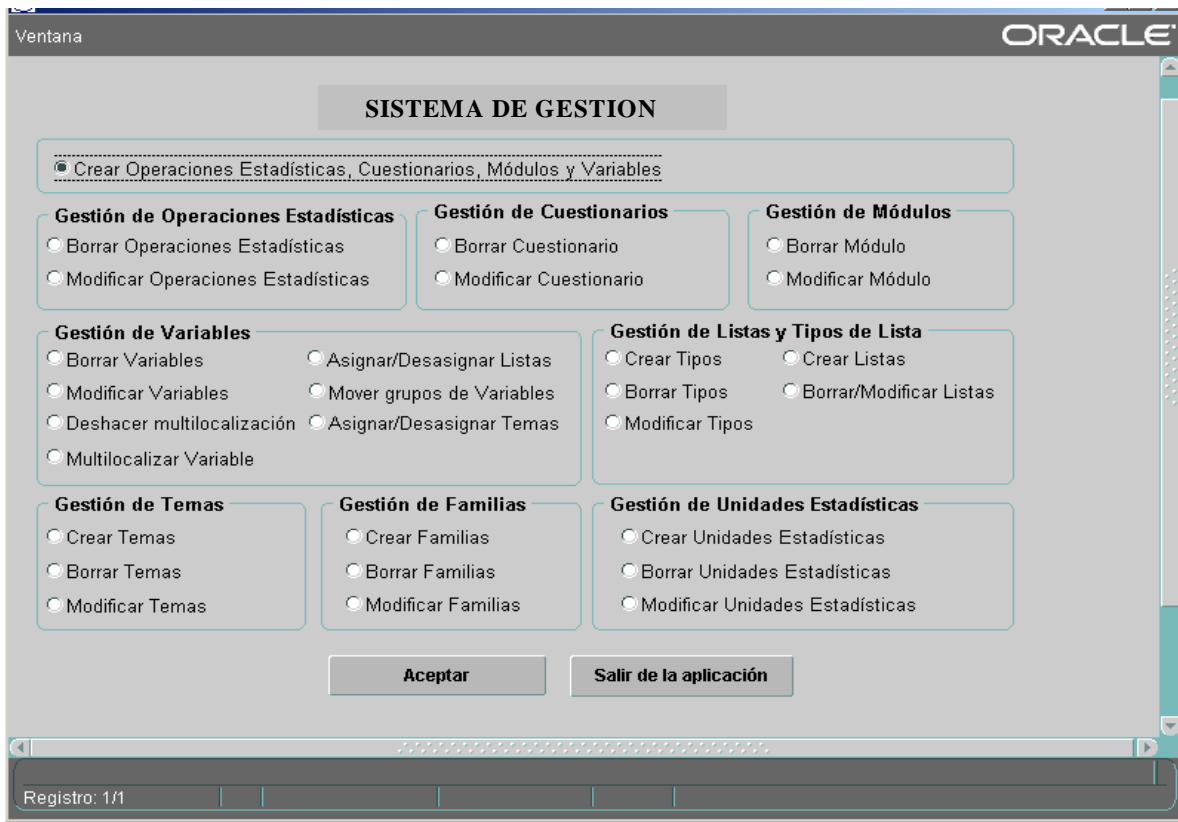


Figure 2: Management System Main window

- **Enquire System:**

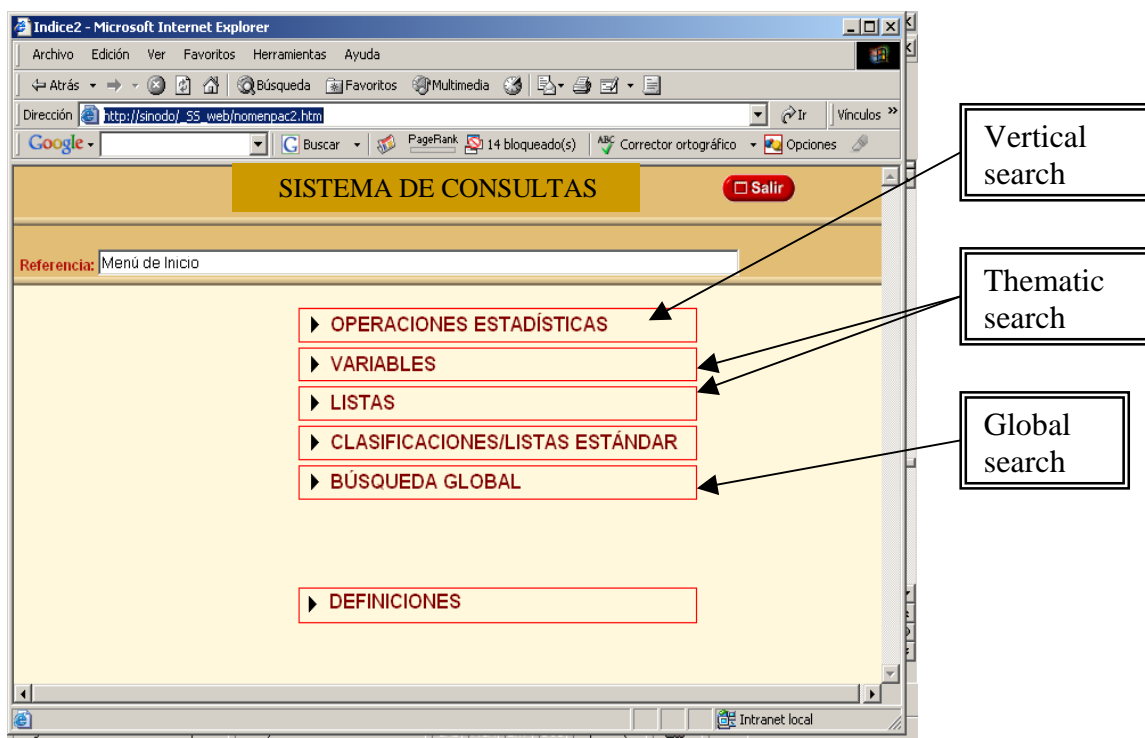


Figure 3: One of the Enquire System windows

The screenshot shows a web browser window with the address bar displaying 'http://sinodo/pls/nomen4/reglamento1'. The page title is 'SELECCIÓN DEL REGLAMENTO A ESTUDIAR'. The page has a yellow background with a red header bar containing the 'IN e' logo. The main content area is divided into two sections: 'Familia:' and 'Reglamento:'. The 'Familia:' section has a dropdown menu with the selected value 'Estadísticas estructurales de las empresas'. The 'Reglamento:' section has a dropdown menu with the selected value '32002R1614 Texto consolidado del Reglamento (CE, Euratom) n° 58/97 del Consejo relativo a las estadís ...'. Below these sections are three red buttons: 'Reglamento', 'Atributos de las variables de reglamento', and 'Variables de OE'. A 'Salir' button is located at the bottom left. Two annotations with arrows point to the page: one from a box labeled 'Selected Regulation' pointing to the 'Reglamento:' dropdown, and another from a box labeled 'Variables from Statistical Operation included in the selected Regulation' pointing to the 'Variables de OE' button.

Dirección <http://sinodo/pls/nomen4/reglamento1>

IN e SELECCIÓN DEL REGLAMENTO A ESTUDIAR

Familia:

Estadísticas estructurales de las empresas

Reglamento:

32002R1614 Texto consolidado del Reglamento (CE, Euratom) n° 58/97 del Consejo relativo a las estadís ...

Reglamento Atributos de las variables de reglamento Variables de OE

Salir

Selected Regulation

Variables from Statistical Operation included in the selected Regulation

Figure 4: Performing searches in Regulations

**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 interchange

THE NATURE OF DATA ¹

Contributed Paper

Submitted by the Bureau of Labor Statistics², United States

I. INTRODUCTION

1. The term data is used everywhere in the statistical and computer science literature. Almost never do you see the concept defined. The apparent reason for this is that data as a term is thought to have a well-defined concept behind it. And this is not unique, as many terms are so commonly used, they cease requiring any kind of definition or justification to their use. The writer assumes everyone knows what the terms mean. Data, function, interface, object, and service are just a few common examples in the computer science field.

2. Occasionally, we run into definitions of commonly used terms such as data. We will look at two of them in this paper. The fundamental question is whether these definitions describe the essence of the concept. What we want to know from any definition is the following: What is the thing we are trying to describe? These prior definitions of data do not answer this question, as we shall see. Once you start digging below the surface, the lack of good definitions leads to confusion. This paper is an attempt to shed light on these and related issues.

¹ The opinions expressed in this paper are those of the authors and do not necessarily reflect the official policies of the Bureau of Labor Statistics or Farance, Inc..

² Prepared by Daniel W. Gillman at the US Bureau of Labor Statistics and Frank Farance at Farance, Inc, New York, NY USA. Contacts gillman.daniel@bls.gov and frank@farance.com.

3. Terminology theory is the study of concepts and their representations in special languages. It focuses on the essential characteristics of concepts, therefore on what a concept is. Applying the theory, we define data in a new way, by defining what data is, by investigating its essential characteristics. This, in turn, provides a way to distinguish between data and information usefully. The role of metadata is clearly defined. Some other related topics are discussed as well.
4. The method for uncovering the essential characteristics of a concept is through concept analysis. See Frankfurt (2005) for a humorous but effective example. Here, we try to uncover the essential characteristics of the concept under review. Of course, a concept is described by a definition, and one that describes the essential characteristics of the concept is an intensional definition. So, our goal is to arrive at intensional definitions.
5. The paper is divided into several sections. First, we describe the essential ideas in the theory of terminology. Then, through a concept analysis, including a critique of the two definitions of data, we arrive at a new definition. From there, new definitions of information, metadata, and data element are derived. Along the way, some surprises are uncovered.

II. THEORY OF TERMINOLOGY

6. Terminology is the study of concepts and their representations in special language. It is multidisciplinary, drawing support from many areas including logic, epistemology, philosophy of science, cognitive science, information science, and linguistics. Work in the area dates all the way back to the ancient Greek philosophers.
7. To begin, we describe some useful constructs from the theory of terminology. These come from several sources (Sager, 1990; ISO, 1999; ISO, 2000). The constructs and their definitions follow below:
 - **characteristic** - abstraction of a *property* of a set of *objects*
 - **concept** - unit of knowledge created by a unique combination of *characteristics*
 - **concept system** - set of *concepts* structured according to the relations among them
 - **definition** - expression of a *concept* through natural language, which specifies a unique *intension* and *extension*
 - **delimiting characteristic** - *essential characteristic* used for distinguishing a *concept* from related *concepts*
 - **designation** - representation of a *concept* by a *sign*, which denotes it
 - **essential characteristic** - *characteristic* which is indispensable to understanding a *concept*
 - **extension** - set of *objects* to which a *concept* refers
 - **general concept** - *concept* with two or more *objects* that correspond to it (e.g., planet, tower)
 - **generic concept** - *concept* in *generic relation* to another that has the narrower *intension*
 - **generic relation** - relation between two *concepts* where the *intension* of one of the *concepts* includes that of the other *concept* and at least one additional *delimiting characteristic*
 - **individual concept** - *concept* with one *object* that corresponds to it (e.g., Saturn, Eiffel Tower)
 - **intension** - sum of *characteristics* that constitute a *concept*
 - **object** - something conceivable or perceivable

- **property** - observation, used to describe or distinguish an *object* (e.g., "Dan has blue-gray eyes" means "blue-gray eyes" is the property of Dan. It is abstracted to a characteristic, color of eyes, of people - see *characteristic*.)
- **sign** - *general concept* whose *extension* contains only perceivable *objects*
- **specific concept** - *concept* in *generic relation* to another that has the broader *intension*
- **subject field** - field of special knowledge

8. Designations come in three types: A term is a verbal designation of a general concept; an appellation is a verbal designation of an individual concept; and a symbol is any other designation. Signs, through which designations are represented, are left undefined, but a sign is what a person perceives and interprets as designating some concept. Basically, however, a sign is a concept whose extension is a set of perceptible objects. Examples of signs are each of the lines and dots on this page we interpret as words, letters, and punctuation. So, what we see and interpret is not really a sign, but an object in the extension of the sign. The objects **F** and **F** are in the extension of the same sign.

9. Characteristics are used in concept formation. They are abstracted from properties of objects and are used to form the intension of concepts. The objects whose properties are abstracted into the characteristics that form the intension of some concept make up its extension. Characteristics may be concepts in their own right, too. They are used in concept analysis, concept modeling, formulation of definitions, and even term formation.

10. The term *specialization* is often used to denote the creation of a specific concept in generic relation to a given, generic, one. Examples are a dog is a specialization of a mammal and a triangle is a specialization of a polygon.

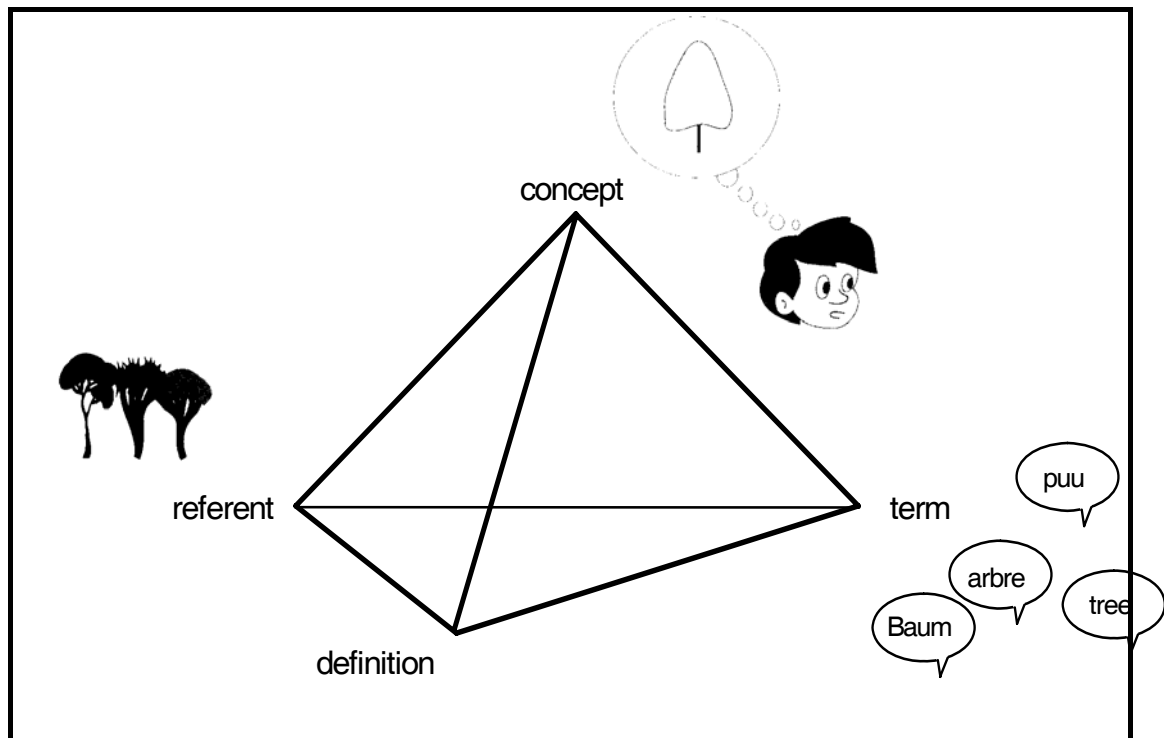


Figure 1: Relationships between referents (objects), concepts, terms (more generally designations), and definitions.

11. The ancient Greek philosophers began the study of terminology and concept formation in language (Wedberg, 1982), and they discovered a useful relationship between designation, concept, object, and definition, that is illustrated in Figure 1 (CEN, 1995). This diagram, minus the definition part, is often referred to as Ogden's Triangle (Ogden and Richard, 1989).

12. Figure 1 shows how terms, concepts, objects, and definitions are related. From the definitions above and Figure 1, several important observations need to be made:

- Concepts, terms, and objects are not the same things
- For any concept, there may be many designations (synonyms)
- For any concept, there are one or more objects in its extension
- For any concept, there may be more than one definition (especially in multiple languages)
- For each term, more than one concept may be designated (homographs)

13. Concepts are human constructions (Lakoff, 2002). No matter how well we define a concept, a complete description is often impossible. Identifying the relevant characteristics is culturally dependent. So, some objects in the extension of a concept, called prototypes, fit the characteristics better than others (Lakoff, 2002). For example, a robin fits more of the characteristics of a bird than a penguin does.

III. Data

A. Problem Description

14. Data is a concept that seems both intuitive and obvious, yet when we try to write down a definition, we struggle to avoid the circular definitions that define "data" using the term "information" and then defining the term "information" using the term "data".

15. Let us consider a definition for data from ISO (1993). The definition reads "Reinterpretable representation of information in a formalized manner suitable for communication, interpretation, or processing." The main problem with this definition is that it tells the reader what data does, not what it is. Data can be reinterpreted; they are suitable for communication, interpretation, and processing. These are all things we want data to do. It is hard to argue with this, except that it is very difficult to tell whether any given objects are data or not. This is because we still don't know what data is from this definition. For this reason, an intensional definition is better - it describes what data are from the point of view of the essential characteristics of the concept.

16. Another approach is to use the familiar triangle that is divided into 4 layers: data (bottom), information, knowledge, and wisdom (top), see Figure 2 - the knowledge hierarchy (Bradley, 2004). We are unsure of precisely what this triangle conveys. The layering might convey composition, definition, or representation. For example, information could be composed of, defined by, or represented as data. In fact, Larry English (1999) defines each layer of the triangle in terms of a function on the thing in the layer below. So, information is a function on data (and some other things). However, the functions are not very precisely defined. Alternatively, the layering might convey something else.

17. English does provide a definition of data: Data is the representation of facts about things. This is required since there is no lower tier on the triangle. On the surface, this meets our criteria. It states what data is. However, on closer examination, there are several questions. Namely, what are meant by representation, facts, and things? These aren't defined, so the definition is not really intensional. For instance, is an idea a thing? What constitutes a

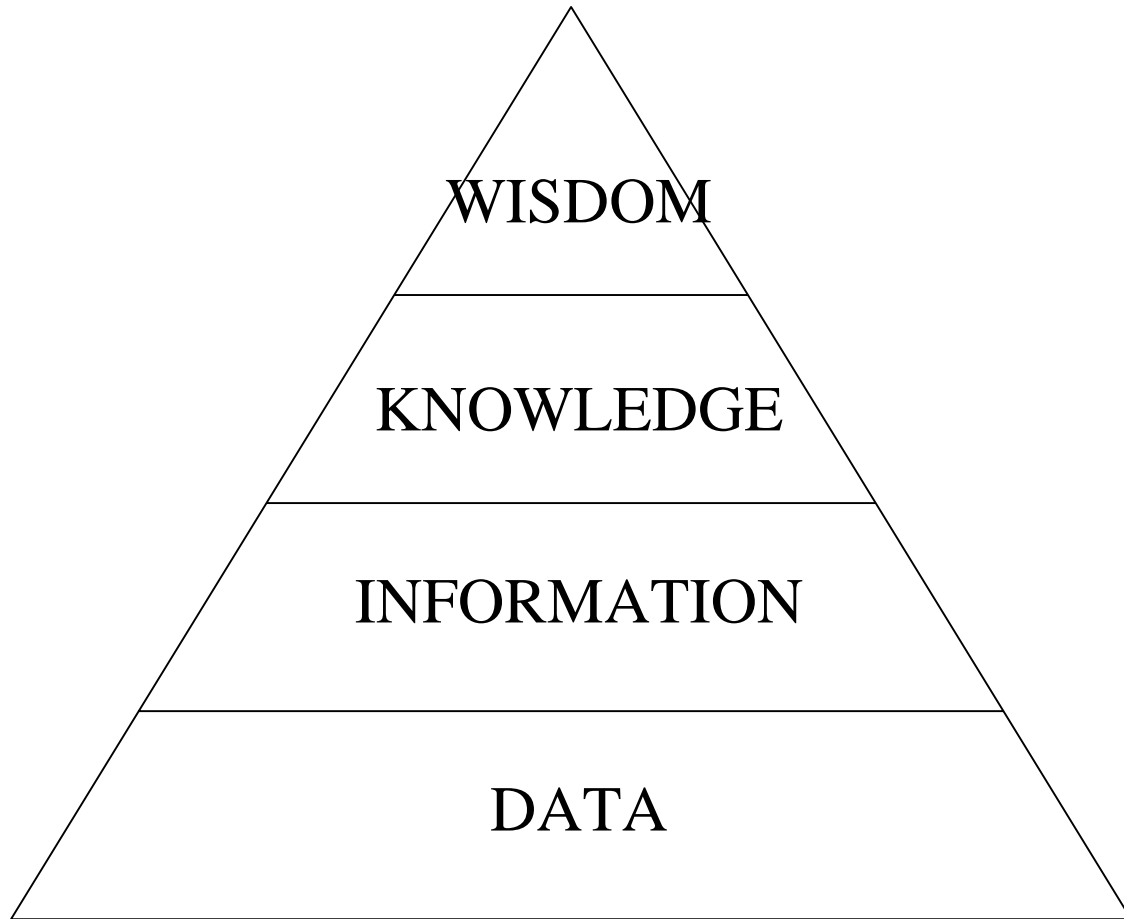


Figure 2: Knowledge Hierarchy

fact? Do data have to represent something that is true? What constitutes a representation? Do all data correspond to characters of some kind? We don't know. Therefore, we don't know the essential characteristics.

18. We believe there is a parallel relationship between data and information, i.e., neither is below the other. We aren't going to try defining knowledge and wisdom, but they might be the subject of some future paper.

19. As stated above, we believe information is a concept related to the concept of data. The topic of "information" has been explored by Langefors (1995), Shannon (1948), and others. Our presentation of data and information is consistent with Shannon's concept of information, and we include a discussion of a close association with Langefor's theory. Yet, we frame the discussion in terminology theory.

20. In this paper, rather than speak of data, we will talk about a datum and derive a definition of datum based on terminological principles. If we look at a datum in a table, say, we recognize that there are underlying concepts the datum represents. The title, column and row headers, and any footnotes in the table tell us much about our datum. These words represent concepts, and these concepts tell us how to interpret the datum. Maybe these words are not enough for a full understanding of what the datum means, however, they serve a purpose. They tell us that a datum is some designation (amount, count, code, etc.) representing a concept in a concept system.

21. An example is the US unemployment rate. If one saw the sign "4.7" in a table showing US unemployment rates for the past 12 months under the column marked "January 2006 in percent", then the reader may infer that the US unemployment rate for January 2006 was 4.7%. Knowing what "unemployment rate" means and relating it to the month of January 2006 gives the reader much information. So, again, the datum is the designation of some concept within a concept system. In order to fully explain these ideas in what follows, we will make this precise.

B. Equality on Concepts

22. A fundamental characteristic of a datum is that it can be compared with other datums³. So, there must be a notion of equality associated with them. A datum represents concepts, so equality on datums requires defining equality for concepts. Therefore,

Definition: A concept has **equality** if its extension is determined consistently

23. Many times, two instances of a concept don't have the same extension. Take the concept nearness, defined as proximity between pairs of objects. The extension contains pairs of objects, which are deemed near. Frank may decide that a table and chair in his room are near to each other. Dan, on the other hand, walks into the same room and decides differently. The table and chair are not near to each other; they are just too far apart for his liking. Frank's and Dan's extensions are not consistently determined, they differ, so their concepts of nearness are not equal.

24. Another example is a pile of salt. Does the pile contain a lot or a little salt? Let's say we open a small bag of salt obtained at a famous American fast-food restaurant and pour the contents on the table. There might be a gram or two of salt spilled out. One person looks at the pile (maybe a child) and exclaims "Look at all the salt!" Another, maybe an older sibling says "Oh, that's not such a big deal. That's a tiny pile." So, one child thinks there is a lot of salt, and the other doesn't. Their extensions for the concept of "how much salt" are not consistently determined.

25. Likewise, two concepts with the same extensions may not have the same intension. Consider two sets of integers defined as follows: $A = \{\text{the greatest integer less than } 5\}$ and $B = \{\text{the least integer greater than } 3\}$. Both these sets define concepts, but their intensions, the essential characteristics, are different. The extensions, $\{4\}$, are the same.

26. Sometimes, the extension of a concept changes over time. This is why the definition for equality above contains the phrase "consistently determined". Requiring that the extension always have exactly the same objects is too strong a condition. The concept of "valid ZIP⁴ codes" used by the US Postal Service is a case in point. ZIP codes change over time. Some new ones are added, and others are removed. At any point in time, two people should be able to determine all the current valid ZIP codes. However, if the comparison is done over a long enough time period, chances are good the lists will be different.

27. Now, it may be the case that for some concepts everyone agrees on the extension for all time. A simple example is the concept "even positive prime integers". Well, we get agreement as long as the concept (the intension) is understood. Not everyone remembers what a prime number is! But, agreement is certainly a subset of being consistently determined.

³ This is deliberate. We want to avoid the use of the word "data" for now.

⁴ ZIP stands for Zone Improvement Plan, instituted in the 1960's in the United States.

C. Partitions

28. Now, a concept with equality has a very important feature. It can be used to partition a set of objects objectively (i.e., independent of who is doing the partitioning), where some of the objects are in the extension of the concept and others are not. A simple example is the concept "even positive numbers". The extension is the well-known set $E = \{2, 4, 6, \dots\}$. The concept partitions the positive integers into even and odd numbers. The reason the concept works as a means to partition the integers is that the extension is consistently determined.

29. Take a concept without equality, such as the nearness concept described above. The extension of the concept nearness, as applied to furniture, is all instances of pairs of pieces of furniture that are near. This extension undergoes changes almost continuously, however this is not the problem. As seen above, two people will not be able to consistently determine the extension even at the same point in time. We will not be able to partition objects objectively with such a concept.

D. Values

30. The concept of five, defined by "having five elements", is a concept with equality. Given a set of elements, it is easy to determine whether it has 5 elements or not. The same is true for any other number. The letter "A", for example, requires the concept "being the letter A". The same is true for other letters, symbols, words, and alphanumeric strings.

31. In every case, the concept partitions, and when that happens we refer to these cases as values. For example, five is "the value assigned to the number of elements in this set {a, b, c, d, e}." Note, the fact that five is a concept with equality now means the value five is useful. Each time five is used as a value, we are sure the extension of the concept, the instances of five, is determined in the same way - consistently.

32. Consider another example that might be a little more illustrative: The concept male, defined by being of the male sex. We apply this to people, and note that people are either male or female. That isn't quite the case, but for practical purposes it is. Male is a concept with equality, and just as in the example of five, we can use it as a value. Thus,

Definition: **Value** is a concept with equality

33. A value partitions the set of objects under study into 2 classes: Those objects that are in the extension of the value, and those that are not. Taking several values at once, it is clear that if the values are carefully chosen, one may construct a partition in which each class is a value. An example is the classification based sex with the values male and female.

34. Since values may be associated with numbers (e.g., integers, real), then the partition might have an arbitrarily high number of classes. Mathematically, a set of numbers may be infinite, but computationally, the set must be finite. Computers and people can't process an infinite number of things. Using a set of number values as the classes of a partition is not typical. For the purposes of this paper, though, we will do it.

35. Examples of partitions based on number values include the following two cases, with one for discrete data and the other for continuous:

- $\{1, 2, 3, 4, \dots, 125\}$ is used to record the age of a person since his last birthday
- $\{\text{all real numbers between 0 and 1 with a precision of 4 digits}\}$ is used to record probabilities - note, there are 10,000 such numbers, not an infinity

E. Datum

36. To summarize, a value is a concept with equality. Since, concepts are known and communicated through their definitions, then we need to define values, too. We refer to this definition as the value meaning.

37. Typically, we designate values many ways. For instance, the value male may be designated 'M', 'male', '0', etc. One of these designations is a term (male), and that is how we refer to the value. Confusingly, the term 'value' is also used to refer to the designations of values. E.g., 'M' as called a value even though it is really the designation of one.

38. A datum is how we refer to the designation of a value. It has the characteristic that it represents the class of some partition. Each class in the partition is a value, and the datum is a stand-in for the value. Since a value is a concept, it may have a designation. Therefore,

Definition: **Datum** is the designation of a value

39. In the modern theory of terminology, many concepts exhibit what is known as prototyping (Lakoff, 2002). The idea is that some objects in the extension of the concept are a better fit for the characteristics than others. As described above, a robin is viewed as a more prototypical bird than a penguin. That isn't to say a penguin is not a bird. It just doesn't fit the concept as well.

40. This effect can even be seen with values. Not all concepts used as values in practice quite fit the definition of a value given above. Data are used for many purposes, and compromises must be made to try to measure society. An example is the categories in an industrial classification system, such as the North American Industrial Classification System (NAICS). Domino's Pizza® is a chain of pizza making and delivery services throughout the US and other countries. Each shop is involved in 2 separate business activities: making and delivering pizza. The NAICS classification is based on production (what is done rather than inputs or outputs), so either cooking or delivering pizza is applicable in this case. Domino's Pizza® shops don't correspond well to either the pizza making industry or the pizza delivery industry. An arbitrary decision was reached on how to handle this (delivery), but this is the case of an object not fitting the scheme very well.

41. The example shows that values sometimes are subject to prototyping as well as concepts in general. Prototyping is an inherent attribute of concepts in language, and it arises when determining the extension of the concept. Since a value is a concept (with equality), then it may be subject to prototyping as well.

42. There are concepts that don't exhibit prototype effects. One is the concept "even positive prime integers." It is easy to see that the extension of this concept is well-defined; it is {2}.

43. The effect of this is that it is often hard to determine the correct value within a partition for objects, and this is inherent. This difficulty is called measurement error, and it is a problem in almost all scientific and statistical data collection activities. Though, it is possible to mitigate the effect by improving definitions and the collection process, as there are undoubtedly errors in collection itself, we also need to see that the problem is inherent to data.

F. Context

44. As we defined above, the concept associated with a datum is its value. There are also a number of other concepts associated with a datum that give it more meaning. These are

described in detail in Gillman and Johannis (2006) and Gillman (2006), however, from the perspective of a survey statistician, they are some of the following:

- Population
- Characteristic
- Category, class, or unit of measure
- Datatype
- Time, space, and survey

45. This list is not complete, but these concepts along with the value form a concept system. This concept system answers the questions "who", "what", "when", "where", "how", and "why" for the datum. We refer to these as the W5H questions. Therefore,

Definition: **Context** is the concept system associated with a datum that answers, minimally, the W5H questions

46. An example is the following:

4.7% is the unemployment rate for the US in January 2006. The details, corresponding to the bulleted items in the paragraph above are

- US labor force
- Unemployment
- Rate as a percent
- Real number
- January 2006, all of the US, and Current Population Survey

These answer the context questions (at least partially!) posed above.

47. We say the context is a concept system because it contains a set of concepts, which is evident, and it contains some relations between those concepts. For instance, the characteristic is related to the population, since it is a characteristic in the terminological sense of the concept represented by the population. The rate is a measure for representing the characteristic, unemployment; the rate has a real number datatype; and the value designated by the datum is a real number. This is a partial list, but the depth of the context is apparent.

G. Data Element

48. Since we have defined datum and context, now we can supply a definition of data element. Intuitively, a data element is a collection of datums that share similar semantics. Often, the notion that each datum cannot be further decomposed is added.

49. Consider the following example of a datum described as in the previous section: A man in the US responds to the Current Population Survey in March 2006. He responds that he is a male to the questions about his sex. The details, corresponding to the bulleted items in section 3.2 are

- US adult person
- Sex
- Male
- Character
- March 2006, all of the US, and Current Population Survey

50. What does a data element containing this datum look like? As we stated above, it is a set of datums with shared semantics, i.e., sharing the same context. But if all the datums share the entire context, then they will all be the same value, i.e., all male.

51. A value is part of the context for a datum, however, it changes from datum to datum. All the rest of the context is shared. Now, a data element need not contain actual datums,

only that it could. This corresponds to a database table that is not yet populated, but the database designer knows what it means. As long as the shared context is known, then the data element can be said to exist. Therefore,

Definition: **Data element** is a set of datums that have the same context except the value⁵

H. Metadata

52. Metadata is often defined to be "data about data." This is not very precise. A better definition says metadata is data that are used to describe. This gets to the essential characteristic of metadata, they serve to describe something. The something is deliberately vague. Metadata can be used to describe anything. Here, we care about metadata describing data.

53. Our definition of datum leads us directly to a new definition for metadata. The context of a datum is the concept system. A concept system is a collection of concepts and the relations among them. The concept system associated with data can be reified, i.e. their meaning may be written out rather than kept as concepts in the heads of individuals. The reification is data, and this data describes the datum. It is metadata. Therefore,

Definition: **Metadata** for a datum are the data reified from the context of the datum

54. Now, there are obvious problems with this definition. It does not convey the essential characteristics of metadata. The definition, data that are used to describe, does this. The point is to convey a greater understanding of the definition of datum, and especially the role of the context. More will be said about this in the next section.

J⁶. Information

55. Information is a term that is widely and loosely used. Information is everywhere. The Internet is known as the Information Highway. All media are said to convey information to their audiences. So, what is information?

56. There are at least 2 well-known definitions of information, due to Langefors (1995) and Shannon (1948). Here we will look at the definition provided by Langefors. He says that information is an interpretation of data, and defines a function, the infological equation, to explain what interpretation means. The function is

$$I = i(D, S, T)$$

Where

- I is the information
- D is data
- S is pre-knowledge
- T is time
- i is the interpretation function

⁵ In fact, a full description of a data element will include all the possible values and the signs designating them.

⁶ The use of the letter I for a sub-section heading could not be distinguished from the Roman Numeral for the number one. So we skipped to the letter J.

57. In general, S is considered to be the total life-experience of the individual making the interpretation. An interesting consequence is that not everyone will interpret the same data in the same way at a given time.

58. The definition of datum given above, allows us to make a different definition of information, but one that is consistent with Langefors. First, our approach is limited to one datum at a time. Second, we don't define a process, the interpretation defined in the infological equation, but we know where the information is.

59. Each datum is a designation of a concept. Associated with that concept, the value, is a concept system, called the context. So,

Definition: The **information** represented by a datum is its context

V. CONCLUSION

60. The purpose of this paper was to provide and justify a new definition of data and some other concepts. We provided a precise definition of datum and derived definitions for context, data element, metadata, and information, as consequences. The relevant definitions are given here for easy reference:

- A concept has **equality** if its extension is determined consistently
- **Value** is a concept with equality
- **Datum** is the designation of a value
- **Context** is the concept system associated with a datum that answers, minimally, the W5H questions
- **Data element** is a set of datums that have the same context except the value
- **Metadata** for a datum are the data reified from the context of the datum
- The **information** represented by a datum is its context

61. The terminology theory is applied to define what a datum is: a designation of a concept with equality. And, each of the terms in the definition are defined as well. The concept piece for this is used to define context, data element, and information. Metadata is defined from context itself.

VI. REFERENCES

Bradley, W. (2004). *From Data to Knowledge: Standards and Systems for Sharing Data and Information Across Systems*. Working Paper #5 presented at the UNECE Workshop on Statistical Metadata,. Geneva, Switzerland.

CEN. (1995). *Medical Informatics - Categorical Structures of Systems of Concepts*. Draft. Brussels: European Committee for Standardization.

Date, C. (2003). *An Introduction to Database Systems (8th ed)*. Addison Wesley.

English, L. (1999). *Improving Data Warehouse and Business Information Quality*. New York: John Wiley and Sons.

Frankfurt, H. (2005). *On Bullshit*. Princeton: Princeton University Press

- Froeschl, K., Grossmann, W., & Del Vecchio, V. (2003). *The Concept of Statistical Metadata*. Deliverable #5 for MetaNet Project. Retrieved July 2004 from http://www.epros.ed.ac.uk/metanet/deliverables/D5/IST-1999-29093_D5.doc.
- Gillman, D. (2006) *Theory and Management of Data Semantics*. In D. Schwartz (ed.) *Encyclopaedia of Knowledge Management*. Hershey, PA, USA: Idea Group.
- Gillman, D. and Johannis, P. (2006). *Metadata Standards and Their Support of Data Management Needs*. Working Paper #7 presented at the UNECE Workshop on Statistical Metadata, Geneva, Switzerland
- ISO (1993). *ISO/IEC 2382-1: Information technology - Vocabulary, Part 1: Fundamental terms*. Geneva: International Organization for Standardization and International Electrotechnical Commission.
- ISO. (1999). *ISO 704: Principles and methods of terminology*. Geneva: International Organization for Standardization.
- ISO. (2000). *ISO 1087-1: Terminology – Part 1: Vocabulary*. Geneva: International Organization for Standardization.
- Lakoff, G. (2002). *Women, Fire, and Dangerous Things* (Reprint edition). University of Chicago Press.
- Langefors, B. (1995). *Essays on Infology*. Stockholm: Studentlitteratur
- Ogden, C. and Richards, I. (1989). *The Meaning of Meaning: A Study of the Influence of Language upon Thought and of the Science of Symbolism*. Harcourt.
- Sager, J. (1990). *A Practical Course in Terminology Processing*. Amsterdam: John Benjamins.
- Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.
- Wedberg, A. (1982). *A History of Philosophy - Vol 1: Antiquity and the Middle Ages*. Oxford: Clarendon Press

**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

**Using the MCV terminology for mapping metadata from different institutions:
the case of Eurostat and OECD**

Supporting Paper

Submitted by Eurostat and OECD¹

I. INTRODUCTION

1. This paper highlights a number of key areas where work is currently taking place, particularly within the seven SDMX sponsoring international organisations², to facilitate the standardisation of metadata concepts used by them, and for establishing stronger coordination of metadata requirements by international organisations so that comparable data and metadata can be made available by the relevant providers more easily, reducing redundancies and minimising reporting effort. These tasks entail development of metadata standards in both the area of IT standards (such as the use of common XML formats and tools) and in what the authors refer to as “metadata content guidelines”.

2. The paper illustrates this through a brief presentation of Eurostat and OECD work utilising a small set of standardised concepts envisaged by the SDMX initiative in the development of corporate metadata facilities by both organisations.

II. TERMINOLOGY IS IMPORTANT

3. The Metadata Common Vocabulary (MCV for short) is one of the key projects launched at the very beginning of the SDMX initiative, in 2001-2002, with the aim of developing a common understanding of standard metadata concepts used by statisticians in documenting the collection, processing, storage and dissemination of statistical data. The immediate objective was the development of a glossary of those standard concepts, whose definitions were consistent with existing, relevant international statistical

¹ Prepared by Marco Pellegrino (Eurostat) and Denis Ward (OECD)

² Bank for International Settlements (BIS), European Central Bank (ECB), Eurostat, International Monetary Fund (IMF), OECD, United Nations Statistical Division (UNSD) and World Bank

guidelines and recommendations, with the terminology used within international organizations, national agencies and, to the extent possible, in other related projects to develop international standards.

4. In general terms, the SDMX initiative aims at increasing the efficiency in the exchange of data and metadata within the range of activities of the seven sponsoring international organisations while minimising the reporting burden of national agencies. Any exchange of information can happen within the framework of a bilateral or multilateral exchange between parties, or with the placement of data and metadata on a location that can be accessed by all partners. In both cases, but particularly when the information is shared over the web, there is an essential need for users to understand the nature (and any limitation of use) of the statistics being exchanged. The SDMX standards intend to ensure that appropriate metadata always accompanies the data: for this reason, standards for metadata exchange are extremely important in the SDMX context. The need for a standardisation of metadata terminology is evident if one considers how many times – in actual data and metadata transmissions – the same metadata item is referred to by different names or, conversely, how many times the same name is associated with different concepts.

5. One of the assumptions made at the commencement of the development of the MCV (which flowed from initial Eurostat-OECD efforts on the joint development of statistical glossaries) was that the development and more importantly, the adoption, of a worldwide general format for metadata management by a large number of international organisations and national agencies, for all statistical domains, is not a realistic goal in the foreseeable future. Within this assumption, the MCV focused on a system of definitions for discrete metadata concepts (e.g. source, contact, periodicity, timeliness, reference period, coverage or adjustment) which can be used for any statistical domain and independently from any general model or list of metadata concepts developed by any individual organisation. The Vocabulary is only concerned with the elaboration of these building blocks, subject to ISO terminology standards, easily understandable and re-usable. Agreement on a basic common vocabulary of metadata concepts still provides each agency responsible for compiling metadata with the flexibility of deriving a variety of specific formats and models according to its specific needs. The list of terms and associated definitions simply provides a common language applicable across domains.

III. CURRENT MCV STATUS

A. The MCV structure

6. As mentioned above, the MCV builds on work already undertaken by several organisations and standardisation bodies, rather than confusing the situation by the development of a whole new set of definitions. Where standard definitions were not available or not satisfactory, suitable national definitions have been considered or new ones have been entered.

7. The MCV covers the following conceptual items: general metadata terms (mostly derived from ISO/IEC 11179 and UN or UN/ECE sources); metadata describing statistical methodologies (classifications, data collection, data editing,...), metadata for assessing quality and a selection of terms referring to data / metadata exchange and SDMX terminology (including previous GESMES/TS terms). The opportunity of having one single entry point for accessing a variety of terms, sometimes not available or hard to find on the Internet, is a clear value added. The MCV glossary is readily available on the web through extensive statistical glossary databases such as *CODED* (Eurostat concepts and definitions database) and the *OECD Glossary of Statistical Terms*. Extractions will soon be available in SDMX-compliant XML format to enable more efficient re-use and sharing of content within and outside SDMX boundaries.

8. The MCV is not intended to cover the whole range of statistical terminology, as this area is already covered by other general glossary databases or subject (classifications, data editing) / domain-specific glossaries (such as prices, national accounts, merchandise trade, etc.). The specific function of the MCV is to contain all terms which are normally used for building and understanding metadata systems. A metadata glossary is necessarily linked to a series of other subject-specific glossaries or to more universal statistical glossaries. The insertion within the MCV of some definitions derived from these other glossaries should not be seen as a redundancy, but as a means of resolving the complex and interdisciplinary nature of metadata.

9. The current MCV draft consists of 384 terms (see Annex 1 for the list of entries). It presents the following "fields": term, definition, source, context and related terms. The "context" field is used extensively throughout the glossary, sometimes to provide additional explanations, other times to highlight peculiarities in how a certain definition is applied within a certain domain or geographical context.

10. The MCV also plays an important standardisation role. The SDMX authors have sometimes chosen to present several context explanations for the same term, always quoting in detail the respective source, to show both their peculiarities and, at the same time, to highlight the possibilities for convergence. This is the case for some "quality assessment" items: users can live with different quality frameworks or different meta-models, as long as each concept is well identified, defined and known. In other words, transparency is a prerequisite for a correct interpretation (and for convergence) of any statistical framework³.

B. THE 2006 UPDATE (INTRODUCING SDMX CROSS-DOMAIN CONCEPTS)

11. The SDMX initiative recently delivered a set of draft "content-oriented guidelines" which have been posted on the web (at <http://www.sdmx.org>) for public comment. These guidelines contain recommendations for creating interoperable data and metadata sets using SDMX standards. The work, focused on the harmonisation of a limited range (i.e. 26) of high-level concepts common to a large number of statistical domains, and is aimed at encouraging the exchange of comparable statistical information both between international organisations and between national agencies and international organisations.

12. The SDMX content package also includes a newly revised version of the MCV, where high-level concepts (see Annex 2) have been taken into account and referred to the nearest – sometimes broader – term. In many instances, the "context" field of the MCV has been updated to document the use recommended in order to be SDMX-compliant.

13. The content package emphasises the identification of reference⁴ metadata concepts and subject-matter domains to test SDMX standards. In this context, the MCV plays an important role in providing the common set of terms and definitions that can be used to describe the data. But the standardisation also includes, where appropriate, the representation of concepts with code lists and the identification of the role they play within data and metadata structures.

14. Further work will be undertaken at the end of the current public consultation process to produce a consolidated version of the MCV to ensure that all definitions in the Vocabulary are in line with the agreed list of cross-domain concepts, to improve the content / wording of MCV definitions, and to make available the SDMX-ML version of the glossary.

IV. MAPPING REFERENCE METADATA TO MCV CONCEPTS

15. As outlined below, Eurostat and the OECD currently use the MCV to ensure clarity and terminological consistency within both organisations' respective metadata repositories (Eurostat free dissemination, OECD MetaStore). The use of standard definitions taken from the MCV is similarly encouraged within metadata-related projects and activities of other international organisations and national agencies.

16. The availability of a web repository of standard metadata definitions, available for all Internet users, is also a unique chance for creating a common understanding across countries, for instance across European Union or OECD Member countries.

³ The possible convergence of the quality frameworks of international organisations is currently being investigated by a task force established in 2005 by the Co-ordinating Committee for Statistical Activities (CCSA) whose work will be discussed at the Q2006 conference in Newport, Wales, on 27-28 April 2006.

⁴ In SDMX, "reference" metadata are metadata describing the contents and the quality of the statistical data, normally including "conceptual" metadata, describing the concepts used and their practical implementation; "methodological" metadata, describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes); and "quality" metadata, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy). These metadata are often stored in a separate metadata repository and they are referenced from the related data element.

A. Metadata within Eurostat

17. Eurostat has disseminated data free of charge from its web site since October 2004. Since the introduction of the new dissemination policy, actions have been taken for producing, monitoring and finally improving the quality of metadata descriptions according to a standardised template, built on the conceptual elements used by the Special Data Dissemination Standard (SDDS) format. But the original SDDS model, developed by the IMF for macro-economic and financial domains, was not sufficient for Eurostat purposes: the main reason is that IMF presents information received for each Member State dataset without being involved in further processing. Eurostat, on the other hand, processes and disseminates data and quality information for the entire EU: for this reason, users expect to receive not only national information but also an overall assessment from a Eurostat perspective. This also reflected the differences in the Eurostat and IMF quality frameworks which were taken into account when adapting the metadata elements to better serve the free data dissemination.

18. The adoption of a "corporate" standard presented for Eurostat the advantage of providing a clear target against which different national and international formats could be mapped. The template used hosts a variety of information items, common to most metadata formats. The list of items is generic (applicable to all the domains) and it has been developed to enable authors to post information at the level of detail needed for each domain and for a variety of different users; in this context, not all of the metadata items are expected to be populated by text for all statistical domains.

19. The format layout consists of a "Base Page" and a "Summary Methodology". The former covers the most basic and broad metadata - providing a short description of the domain, in terms of its coverage, periodicity, timeliness, legal basis or agreement and other general items - whereas the summary methodology has a more technical and statistical character. In the summary methodology, specific sub-elements (e.g. statistical units, reference period, adjustments and compilation of EU aggregates) were added to the original chapters in order to provide a more detailed statistical structure within the general template.

20. The information currently contained in more than 550 files - publicly available on Internet and covering the whole range of statistical data disseminated - will be managed by a new database system which is going to be operational by the end of 2006 to facilitate the creation, checking, re-use and dissemination of a detailed list of metadata items, by linking metadata to the dissemination tree. In a second stage, a "metadata handler" will offer the possibility of browsing and navigating in different metadata systems (for classifications, concepts and definitions, methodological texts,...) thereby allowing to further enrich explanatory text.

21. While working on the technical infrastructure, Eurostat is therefore currently improving the granularity of the metadata format with the aim of extending the conceptual coverage, in particular for incorporating more elements on quality assessment, according to the criteria identified by the European statistical code of practice. The list of granular concepts (described in Annex 3) is built on the current format used, with some limited extensions on quality elements which are going to be further detailed by the end of this year. The current list is going to be used for testing the possibility of disseminating a good selection of reference metadata with regard to the SDMX implementation activities with European member States⁵.

B. Metadata within the OECD

22. The current situation with respect to metadata in the OECD takes place within the context of the Organisation's decentralised statistical system wherein statistical data collection, storage and dissemination for 30 Member countries (plus a limited number of non-member states) is conducted by a number of units (Directorates) across the OECD. Only limited metadata are stored in databases actually linked to the statistics they describe. Most metadata currently reside in numerous text files that have been used in the preparation of a large number of statistical and other publications produced across the Organisation. In the

⁵ See METIS 2006, WP 27, *Using SDMX Standards for rapid dissemination of short-term indicators on the European economy*, submitted by Eurostat.

absence of a corporate metadata facility and corporate metadata model there is frequent duplication of metadata storage.

23. A key element of the new OECD corporate data environment currently being developed is the MetaStore⁶ facility which will, for the first time, enable users within the Organisation to store their metadata in a corporate environment that can be readily accessed by different in-house users and allow metadata describing common data disseminated by different Directorates to be linked to different outputs in lieu of duplicated collection and storage. MetaStore also has the capability of storing links (publicly available URLs) to metadata maintained both by other international organisations and national agencies, again in lieu of direct collection. It is also equipped with powerful text search and retrieval facilities. Finally, MetaStore is linked to other elements of the OECD corporate data environment such as the primary external data dissemination facility, OECD.stat and the *OECD Glossary of Statistical Terms*.

24. MetaStore provides sufficient flexibility appropriate to the OECD's decentralised statistical environment. In a worst case scenario it could merely be used as a corporate metadata storage facility to store existing duplicated metadata compiled and disseminated by the various Directorates within the OECD. In order to maximise the advantages and potential such a facility provides the OECD has also developed a set of governance practices, etc, regarding the insertion of new metadata in lieu of using existing metadata both within MetaStore and in the repositories of other organisations (refer footnote 5).

25. The MetaStore facility contains a set of 41 metadata items and their related definitions (refer "Child level items" in Annex 4). When developed 18 months ago these were almost completely consistent with the then existing version of the MCV. The mapping between the MetaStore items and the high level concepts contained in the recently released draft SDMX content-oriented guidelines in Annex 3 below shows that by and large this consistency remains though further adjustments / refinements will be made to the MetaStore items once the SDMX standards become more stable following public consultation. As can be seen in Annex 4, there is either a one to one relationship between the OECD and SDMX concepts or a many to one relationship where the OECD concept is more granular (e.g. coverage, statistical processing).

26. It should be emphasised that the MetaStore list of metadata items has been developed in the context of the requirements of an international organisation, specifically the OECD, where the main need is for broad metadata that describes the statistics collected and disseminated by the Organisation. In this context not all of the 41 metadata items are expected to be populated by metadata text for all statistical outputs disseminated by all Directorates across the OECD. For example, the OECD's *Main Economic Indicators* (MEI) uses a subset of around 12 of the MetaStore items. Similarly, not all the draft SDMX cross-domain concepts are used in MetaStore.

V. FACILITATING THE EXCHANGE OF METADATA THROUGH SDMX: HOW NATIONAL AGENCIES CAN BENEFIT FROM THIS

27. Recent developments in SDMX technical standards (now ISO/TS/17369) encouraged the specification of formal rules for formatting data and metadata, so that these can be exchanged, read and processed automatically. The use of standard concepts can be applied to the exchange of data and metadata sets. This involves the use of metadata concepts of generic use, common to a number of domains, such as "periodicity", "timeliness", "data source", "statistical adjustment", or "compilation", as well as other which may be specific to a statistical subject-matter domain. A web-service, using information about web locations of data and metadata, can then navigate, find and automatically process the information.

28. Through the alignment to these standards by both international organisations and national agencies, there is a concrete possibility of setting the requirements for a concept family of metadata to be exchanged and shared among countries and international organisations. Alignment does not necessarily entail the direct adoption of precisely the same concept (or concepts) by each agency. Although such adoption would

⁶ The MetaStore facility is described in more detail in the paper, *Implementation of MetaStore at the OECD* (Penlington and Thygesen), presented in Session 4 (WP25) at the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, 3-5 April 2006

facilitate the ability to exchange metadata (with the same “content”) between agencies, it would be sufficient for organisations to be able to map the granular concepts developed to meet their own circumstances and needs to the list of high-level cross-domain concepts specified in the recently released SDMX content-oriented guidelines – as envisaged in the mapping between Eurostat – SDMX concepts and OECD MetaStore metadata items - SDMX concepts in Annexes 3 and 4 below.

29. Agreement on a common set of concepts by national agencies and international organisations would represent a significant step forward. In this context, Eurostat and OECD could play a role in interconnecting European and national metadata for a wide range of indicators of common interest.

30. The mechanism for the actual exchange of metadata between organisations is beyond the scope of the current paper. However, the adoption of the common set of concepts envisaged by the SDMX content-oriented guidelines and the accessibility of metadata – based on these concepts – on national websites would facilitate direct access by international organisations in lieu of the current transmission(s) by national agencies of different metadata to different international organisations. Incidentally, this also demands a good coordination between national and international organisations, to make sure that standardised metadata covering a range of common requirements are made available through the web, while additional information is created directly at supranational level, when this is needed to document the data sets which are disseminated.

VI. CONCLUSION

31. Establishing a framework for the identification of a set of cross-domain concepts implies a dynamic update of the MCV to reflect the SDMX standards as they evolve through the current and future public consultation process. The metadata concepts and perhaps more granular items falling within the core set of terms will be revised and the MCV will need to include new terms and refine existing definitions. In general, as metadata concepts will be subject to periodic revision and supplementation, the MCV will never be considered as complete or final. On the other hand, agreement on a common terminology and on a target set of metadata items allows a simplification of metadata production, the set up of synergies and a reduction of double work for national and international organisations. This also reduces the risk of delivering inconsistent and overlapping information to the users

32. Annexes 2, 3 and 4 to the present paper provide an example of how two organisations could map their internal systems of metadata items to a set of common high-level concepts, and the list of concepts to the standard terminology embedded in the MCV. It would be interesting to know about similar exercises conducted by national institutes and other organisations. Agreement on a limited number of metadata concepts and their mapping to (perhaps more granular) sets of concepts developed by national agencies and international organisations to meet their own particular needs would provide an opportunity for setting the boundaries and contents of a more efficient exchange of metadata. This would also help achieve a reduction of effort at the national level.

LIST OF TERMS DEFINED WITHIN THE METADATA COMMON VOCABULARY (MARCH 2006)

1. Accessibility
2. Accounting basis
3. Accounting conventions
4. Accuracy
5. Adjustment
6. Adjustment Methods
7. Administered item
8. Administration record
9. Administrative data
10. Administrative data collection
11. Administrative source
12. Agency
13. Aggregation
14. Aggregation Equation
15. Analytical framework
16. Analytical unit
17. Area sampling
18. Attachment level
19. Attribute
20. Availability
21. Base period
22. Base weight
23. Base year
24. Basic attribute
25. Benchmark
26. Benchmarking
27. Bias
28. Bilateral exchange
29. Break
30. Category
31. Category Scheme
32. Census
33. Chain index
34. Characteristic
35. Clarity
36. Class
37. Classification
38. Classification changes
39. Classification scheme
40. Classification unit
41. Co-ordination of samples
42. Code
43. Code list
44. Coding
45. Coding error
46. Coherence
47. Collection
48. Comparability
49. Compilation
50. Compilation practices
51. Compiling Agency
52. Completeness
53. Computation of lowest level indices
54. Computer Assisted Interviewing, CAI
55. Concept
56. Concept Scheme
57. Conceptual data model
58. Conceptual domain
59. Confidential data
60. Confidentiality
61. Consistency
62. Consolidation
63. Constraint
64. Contact
65. Context
66. Country identifier
67. Coverage
68. Coverage errors
69. Coverage ratio
70. Creation date
71. Cross-domain Concepts
72. Cut-off survey
73. Cut-off threshold
74. Data
75. Data analysis
76. Data attribute
77. Data capture
78. Data checking
79. Data collection
80. Data confrontation
81. Data consumer
82. Data dissemination
83. Data dissemination standards
84. Data editing
85. Data element
86. Data element concept
87. Data element derivation
88. Data exchange
89. Data exchange context
90. Data flow definition
91. Data identifier
92. Data interchange
93. Data item
94. Data model
95. Data presentation
96. Data processing
97. Data provider
98. Data provider series key

99. Data reconciliation
100. Data security
101. Data set
102. Data set identifier
103. Data sharing exchange
104. Data source
105. Data status upon release
106. Data structure definition
107. Datatype
108. Date
109. Date of last change
110. Definition
111. Derivation input
112. Derivation output
113. Derivation rule
114. Derived data element
115. Derived statistic
116. Dimension
117. Dimensionality
118. Disaggregation
119. Disclosure analysis
120. Dissemination format
121. Documentation
122. Domain groups
123. Dublin Core
124. EDIFACT
125. Electronic data interchange (EDI)
126. Entity
127. Error of estimation
128. Error of observation
129. Estimate
130. Estimation
131. Estimator
132. Expected value
133. Expression Node
134. Flag
135. Flow data series
136. Follow-up
137. Footnote
138. Frame
139. Frame error
140. Frequency
141. Gateway
142. Gateway exchange
143. General Data Dissemination System (GDDS)
144. Geographical coverage
145. GESMES
146. GESMES/CB
147. GESMES/TS
148. GESMES/TS data model
149. Glossary
150. Graphical data editing
151. Grossing/Netting
152. Guidelines
153. Hierarchy
154. Identifier
155. Imputation
156. Index number
157. Information
158. Information system
159. Inlier
160. Institutional framework
161. Institutional sector
162. Institutional unit
163. Integrity
164. Internal access
165. International code designator
166. International statistical standard
167. Interpolation
168. Interviewer error
169. ISO/IEC 11179
170. Item response rate
171. Key (time series or sibling group)
172. Key family
173. Key structure
174. Keyword
175. Language
176. Level
177. Levels of data
178. Longitudinal data
179. Macro editing
180. Maintenance Agency
181. Measure
182. Measurement error
183. Metadata
184. Metadata Attribute
185. Metadata dimension
186. Metadataflow definition
187. Metadata item
188. Metadata layer
189. Metadata object
190. Metadata registry
191. Metadata set
192. Metadata Structure Definition
193. Metamodel
194. Methodological soundness
195. Methodology
196. Micro editing
197. Ministerial commentary
198. Misclassification
199. Missing data
200. Model assumption error
201. Multilateral exchange
202. Name
203. Nature of the basic data
204. Nomenclature
205. Non-probability sample
206. Non-response
207. Non-response bias
208. Non-response error
209. Non-response rate
210. Non-sampling error
211. Not seasonally adjusted series

- 212. Number raised estimation
- 213. Object
- 214. Object class
- 215. Objectives
- 216. Observation
- 217. Observation confidentiality
- 218. Observation unit
- 219. Observation value
- 220. Ontology
- 221. Organisation
- 222. Organisation identifier
- 223. Organisation Role
- 224. Origin
- 225. Out-of-scope units
- 226. Outliers
- 227. Over-coverage
- 228. Period
- 229. Periodicity
- 230. Permissible value
- 231. Permitted value
- 232. Pre-break observation
- 233. Pre-Break Value
- 234. Precision
- 235. Preferred definition
- 236. Prerequisites of quality
- 237. Primary data
- 238. Primary source of statistical data
- 239. Probability sample
- 240. Processing error
- 241. Product
- 242. Professionalism
- 243. Property
- 244. Provider load
- 245. Provision Agreement
- 246. Public disclosure
- 247. Punctuality
- 248. Qualitative data
- 249. Quality
- 250. Quality (Eurostat context)
- 251. Quality (IMF context)
- 252. Quality (OECD context)
- 253. Quality control survey
- 254. Quality differences
- 255. Quality index
- 256. Quantitative data
- 257. Questionnaire
- 258. Questionnaire design
- 259. Ratio estimation
- 260. Recommended use of data
- 261. Record check
- 262. Record-keeping error
- 263. Recording of transactions
- 264. Reference document
- 265. Reference metadata
- 266. Reference period
- 267. Reference time
- 268. Refusal rate
- 269. Register
- 270. Registrar
- 271. Registration
- 272. Registration authority
- 273. Registry item
- 274. Registry metamodel
- 275. Related data reference
- 276. Related metadata reference
- 277. Relationship
- 278. Relative Standard error
- 279. Release calendar
- 280. Relevance
- 281. Reliability
- 282. Reporting unit
- 283. Respondent burden
- 284. Respondent load
- 285. Response errors
- 286. Response rate
- 287. Responsible organization
- 288. Revision policy
- 289. Sample
- 290. Sample design
- 291. Sample size
- 292. Sample survey
- 293. Sampling
- 294. Sampling error
- 295. Sampling fraction
- 296. Sampling frame
- 297. Sampling technique
- 298. Sampling unit
- 299. Schedule
- 300. Scope
- 301. SDMX-EDI
- 302. SDMX-ML
- 303. SDMX Registry
- 304. Seasonal adjustment
- 305. Secondary source of statistical data
- 306. Semantics
- 307. Serviceability
- 308. Sibling group
- 309. Simultaneous release
- 310. Source
- 311. Source data
- 312. Special Data Dissemination Standard SDDS
- 313. Special language
- 314. Standard Classification
- 315. Standard error
- 316. Statistical concept
- 317. Statistical Data and Metadata Exchange SDMX
- 318. Statistical error
- 319. Statistical indicator
- 320. Statistical macrodata
- 321. Statistical measure
- 322. Statistical message
- 323. Statistical metadata
- 324. Statistical metadata repository

- 325. Statistical metadata system
- 326. Statistical metainformation
- 327. Statistical metainformation system
- 328. Statistical methodology
- 329. Statistical microdata
- 330. Statistical population
- 331. Statistical processing
- 332. Statistical production
- 333. Statistical standard
- 334. Statistical subject-matter domain
- 335. Statistical unit
- 336. Stewardship
- 337. Stratification
- 338. Structural definition
- 339. Structural metadata
- 340. Structure
- 341. Study domain
- 342. Submission
- 343. Submitting organization
- 344. Supplementary data
- 345. Survey
- 346. Survey data collection
- 347. Survey design
- 348. Syntax
- 349. Target population
- 350. Taxonomy
- 351. Term
- 352. Terminological entry
- 353. Terminological system
- 354. Terminology
- 355. Thesaurus
- 356. Time coverage
- 357. Time of recording
- 358. Time Period
- 359. Time series
- 360. Time series breaks
- 361. Timeliness
- 362. Transparency
- 363. Trend
- 364. Trend estimates
- 365. True value
- 366. Type of data collection
- 367. Under-coverage
- 368. Unit non-response
- 369. Unit of measure
- 370. Unit response rate
- 371. Unit value
- 372. Unit value index
- 373. User needs (for statistics)
- 374. User satisfaction survey
- 375. Validation
- 376. Valuation
- 377. Value domain
- 378. Value item
- 379. Value meaning
- 380. Variable
- 381. Verification
- 382. Weight
- 383. XML
- 384. Year-to-date data

**CURRENT CORRESPONDENCE BETWEEN SDMX CROSS-DOMAIN CONCEPTS FOR
"METADATA STRUCTURE DEFINITIONS" AND MCV TERMS**

Concept	Nearest MCV term
1. Accessibility of documentation	Accessibility
2. Accounting conventions	Accounting conventions
3. Accuracy	Accuracy
4. Classification systems	Classification
5. Comparability/Coherence	Comparability, Coherence
6. Statistical concept	Statistical concept
7. Confidentiality	Confidentiality
8. Contact	Contact
9. Data presentation	Data presentation
10. Date of update	Date of last change
11. Dissemination formats	Dissemination format
12. Frequency and Periodicity	Frequency, Periodicity
13. Institutional framework	Institutional framework
14. Professionalism and ethical standards	Professionalism
15. Quality management (incl. resource management)	Quality
16. Release calendar	Release calendar
17. Relevance	Relevance
18. Revision policy and practice	Revision policy
19. Scope / coverage	Scope, coverage
20. Simultaneous release	Simultaneous release
21. Source data	Source data
22. Statistical processing	Statistical processing
23. Supplementary data	Data dissemination
24. Timeliness and punctuality	Timeliness, Punctuality
25. Transparency	Integrity
26. Validation	Validation

EUROSTAT – SDMX CROSS-DOMAIN CONCEPTS MAPPING

EUROSTAT DISSEMINATION METADATA CONCEPTS		MAPPING TO CURRENT DRAFT OF SDMX CROSS- DOMAIN CONCEPTS
Top level	Child level	
Metadata Update	Last certified without update	Date of update
	Last update of content	Date of update
Contact	Organisation	Contact
	Address	Contact
	Contact name or service	Contact
	e-mail address	Contact
Data coverage	Short description of data domain	Data presentation
	Data breakdown and main variables	Data presentation
	Units of measure	Data presentation
Periodicity	Periodicity of compilation	Frequency and periodicity
	Database frequency	Frequency and periodicity
Timeliness and punctuality	Timeliness	Timeliness and punctuality
	Punctuality	Timeliness and punctuality
Transparency of practices	Legal acts, reporting requirements	Institutional framework
	Rules on confidentiality	Institutional framework
	Internal access	Transparency
	Commentary on the occasion of release	Transparency
Accessibility	Notification of changes in methodology	Transparency
	Release calendar	Release calendar
	Simultaneous release	Simultaneous release
	Dissemination formats	Dissemination formats
	Documentation on methodology	Accessibility of documentation
Quality cross-checks	Related data and quality cross-checks	[No direct concordance]
	References to quality reports	[No direct concordance]
Accuracy and reliability	Overall accuracy assessment	Accuracy
	Quality checks before release	Accuracy
Comparability and coherence	Comparability over time	Comparability and coherence
	Comparability over space	Comparability and coherence
	Comparability with related sources	Comparability and coherence
	Comparability between datasets	Comparability and coherence
	Breaks in time series	Comparability and coherence
Relevance	Rate of available statistics (user needs)	Relevance
	Intended audience and purpose	Relevance
	Supplementary data	Supplementary data
Statistical concepts and classifications	Statistical concept	Statistical concept
	Definition of indicators	Statistical concept
	Classification system	Classification systems
	Conformity with official standards	Classification systems
	Classification coverage	Classification systems
Scope of the data	Reference area / geopolitical entity	Scope/coverage
	Time coverage	Scope/coverage
	Statistical unit	Scope/coverage
	Statistical population	Scope/coverage
Accounting conventions	Reference period	Accounting conventions
	Base period	Accounting conventions
	Basis for recording	Accounting conventions
Nature of basic data	Data source used	Source data
	Type of survey	Source data
	Methods of data collection	Source data
Compilation practices	Compilation	Statistical processing
	Adjustments and weights	Statistical processing
	Data validation	Statistical processing
	Revision policy and practice	Revision policy and practice
Other	Warnings on re-use and limitations	[No direct concordance]

OECD METASTORE – SDMX CROSS-DOMAIN CONCEPTS MAPPING

OECD METASTORE METADATA ITEMS		MAPPING TO CURRENT DRAFT OF SDMX CROSS-DOMAIN CONCEPTS
Top level	Child level	
Source	Contact person and organisation	Contact
	Data source(s) used	Source data
	Name of collection / source used	Source data
	Direct source	Source data
	Source Periodicity	Frequency and periodicity
	Source metadata	Accessibility of documentation
	Date last input received from source	Timeliness and punctuality
Data characteristics and collection	Unit of measure used	Accounting convention / basis
	Power code	Accounting convention / basis
	Variables collected	Statistical concept
	Sampling	Source data
	Periodicity	Frequency and periodicity
	Reference period	Timeliness and punctuality
	Base period	Data presentation
	Date last updated	Date of update
	Link to Release calendar	Release calendar
	Contact person	Contact
	Other Data characteristics and collection	[No direct concordance]
Statistical population and <u>scope of the data</u>	Statistical population	Scope / coverage
	Geographic coverage	Scope / coverage
	Sector coverage	Scope / coverage
	Institutional coverage	Scope / coverage
	Item coverage	Scope / coverage
	Population coverage	Scope / coverage
	Product coverage	Scope / coverage
	Other coverage	Scope / coverage
Statistical concepts and <u>classifications used</u>	Key statistical concepts used	Statistical concept
	Classification(s) used	Classification systems
Manipulation and dissemination	Aggregation & consolidation	Statistical processing
	Estimation	Statistical processing
	Imputation	Statistical processing
	Transformations	Statistical processing
	Validation	Validation
	Index type	Dissemination format
	Weights	Source data
	Seasonal adjustment	Statistical processing
	Other manipulation & adjustments	Statistical processing
	OECD Dissemination format(s)	Dissemination formats
Other aspects	Recommended uses and limitations	Relevance
	Quality comments	[No direct concordance]
	Other comments	[No direct concordance]

**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

**RECENT ACHIEVEMENTS IN METADATA SYSTEMS AT THE ROMANIAN NATIONAL
INSTITUTE OF STATISTICS**

Supporting Paper

Submitted by National Institute of Statistics, Romania¹

I. INTRODUCTION

1. In the framework of Romanian National Institute of Statistics there are some domains in full swing. Among there we can discern the follows: Intra-Community Trade INTRASTAT, Central Database TEMPO and Special Data Dissemination Standard SDDS.

2. These new domains require a special metadata approach. In terms of each topic, the used metadata have different parts. The specific of mention domains claims a real challenge for metadata management.

II. THE INTRASTAT METADATA

3. According to Romanian accessing to European Union, National Institute of Statistic is constrained to assume two different ways for building his foreign trade data collection. The first one will be the registration of intra - Community transactions, an activity that will be carry on NIS. The second one will be the uploading of data about extra - Community trade, according to the records provided from Romanian Customs Authority.

4. In the first case is necessary that NIS will design a register which contents data about Romanian exporters and importers. This register must be considered one of the main components of INTRASTAT metadata system. A government decision established that exporters - importers register will be starting from data furnished by Romanian Customs Authority, in the first year of this statistical survey, and Ministry of Public Finance, in the future, according to information content in the new format of VAT questionnaire.

¹ Prepared by Edmond - Lucian SINIGAGLIA, lucians@insse.ro.

5. Starting with data entrance, the metadata will play a significant role. The transaction will be registered if the input data will be filtered according with metadata related. Therefore will be verified the existence in nomenclatures and registries of the following information: commodity code (according to Combined Nomenclature and List of goods excluded from statistics relating to the trading of goods between Member States to be transmitted to the Commission), partner country (EU countries list), flow (export/import), details about transaction nature (according to custom regime), mode of transport, coding of delivery terms (Incoterm ICE/ECE Geneva).

6. Another metadata used in the data uploading is the interval system. Depending on the types of goods, the range of value is defined in relation to the weight and/or in relation to supplementary units. That gives us the following possibilities (an interval for the value/weight, an interval for the value/supplementary unit, an interval for the weight in kg/supplementary units). There are any cases in which the intervals are fixed by the Combined Nomenclature itself.

7. In contrast with extra - Community trade data, which will be load exhaustively from Customs Authority, the INTRASTAT data will be process in the NIS and the role of metadata will be increase.

III. THE CENTRAL DATABASE TEMPO

8. The Central Database with Time Series TEMPO permits the management in an integrated conception the NIS main data thesaurus and the metadata related. Through the integration process with metadata, the statistical information can be dominated in a good way. Also the data misinterprets are minimized.

9. In TEMPO system the data are stored in multi-dimensional matrixes or hyper-cubs. If a matrix have many dimensions, her values are most detailed. Each matrix dimension has associated a statistical nomenclature or classification.

10. Every data matrix has two compulsory dimensions: the time reference and the unit of measure in which are expressed the matrix values).

11. The data matrixes are classified into a hierarchical structure which has many levels:

- sector;
- domain;
- sub-domain;
- indicator.

12. Each matrix has associated certain metadata categories, such as:

- the indicator definition;
- information about used nomenclatures/classifications;
- methodological items (range, adjustment methodology, break into time series);
- information about data sources;
- information about data confidentiality.

13. The main terms used in TEMPO, joined with data and metadata categories, are described in glossary of items, which is accessible through main menu of application.

14. The database TEMPO consultation is possible owing to a client type program, designed in Delphi C/S 4.0. The access to data and metadata related is made on account of Oracle database, being of NIS server.

15. The first option offered in the main menu is SEARCH, which contains the following options:
- CATALOGUE;
 - NOMENCLATURE;
 - MATRIX CODE;
 - KEY WORDS;
 - PERIODICITY;
 - DATA SOURCE;
 - ADJUSTMENT;
 - LAST SEARCH;
 - LOAD FROM DISK;
 - MATRIXES LIST IN MEMORY.
16. It is obvious that the data access is facilitated by a set of metadata. It is possible to select a data matrix starting with the elements of a nomenclature/classification, such as:
- a nomenclature subset;
 - a position contents in a subset/entire nomenclature.
17. There are any facilities for selecting/deselecting of the nomenclature/subset nomenclature.
18. There is the possibility to retrieve the data starting to key words, which can be looked as a sort of metadata. These metadata, which can be classified as retrieval metadata, are:
- the indicator name;
 - the indicator definition;
 - the matrix name.
19. VISUALIZATION, another option of main menu, provides other facilities, among other things a lot of metadata:
- EXTRACTING SPECIFICATION;
 - DATA;
 - METADATA;
 - GRAPHICS.
20. Choosing the METADATA option, the system shows the metadata categories associated to current matrix, divided into many areas, such as:
- DEFINITION;
 - TIME INFORMATION;
 - METHODOLOGY;
 - DATA SOURCES.
21. In the framework of metadata visualization, there is information related to:
- periodicity:
 - period of start;
 - the last period loaded;
 - period of end.
 - absence of values;
 - record keeping data;
 - successive matrix;
 - number of loaded values;

- definitive;
 - temporary;
 - setting right.
- break into time series:
 - the break reasons;
 - the break period;
 - conversion algorithm;
 - matrix with compatible data.
- adjustment:
 - methodology of adjustment;
 - associated adjusted matrix.
- other methodological notes.

IV. Special Data Dissemination Standard SDDS²

22. In accordance with the national Strategy for Development of Statistics Romania subscribed to the Special Data Dissemination Standard SDDS in May 2005. SDDS is established by the International Monetary Fund to guide member states that have or might seek to have access to the international capital markets in the dissemination of economic and financial data.

23. SDDS, in taking a comprehensive view of the dissemination of these data identifies four dimensions of data dissemination:

- the data (coverage, periodicity, and timeliness);
- the access by the public;
- the integrity of the disseminated data;
- the quality of the disseminated data.

24. The IMF's Dissemination Standards Bulletin Board (DSBB) presents the components of the SDDS concerning:

- the metadata;
- the methodological skills;
- the quality and accessibility, equality of data-users and thus emphasizes transparency in the compilation and dissemination of the official statistics.

25. The responsibility for the accuracy of the metadata, including timely updates, and for the economic and financial data underlying the metadata rests with the subscriber.

26. The National Summary Data Page (NSDP) includes a wide range of macroeconomic indicators (data categories) related the follow sectors of the national economy:

- Real Sector;
- Fiscal Sector;
- Financial Sector;
- External Sector;
- Population.

² A part of this section is presented in a form received from IMF.

27. In relation to the NSDP the SDDS requires the elaboration and maintenance of an Advance Release Calendar (ARC) that comprises provisional information concerning the timing of the data releases for a four months horizon.

The metadata section of SDDS contains:

- Information by data category (select information by accessing a list of data categories);
- Information by country (select information by accessing a list of subscribing countries);
- View information about metadata dimension and metadata elements by data categories and countries;
- View information by key concepts within for one or more metadata elements countries and data categories;
- Advance Release calendar (ARC) information for one or more data categories and countries;
- View summary of observance information: cross - country practices versus the SDDS.

WP. 15
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and EUROPEAN COMMISSION
ECONOMIC COMMISSION FOR EUROPE STATISTICAL OFFICE OF THE
CONFERENCE OF EUROPEAN STATISTICIANS 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 (iii): Metadata and Statistical Cycle

INTRODUCTION TO THE TOPIC AND CONCLUDING ISSUES FOR DISCUSSION

Submitted by Graeme Oakley, Australian Bureau of Statistics
[graeme.oakley@abs.gov.au]

I. BACKGROUND

1. As a result of discussions at the Work Session on Statistical Metadata, held in Geneva on 9-11 February 2004, the ECE secretariat prepared a proposal to the Bureau of the Conference of European Statisticians (CES/BUR.2004/44) for the creation of a Task Force to create and publish a framework for statistical metadata. "The key objective of the work of the Task Force is to organise available information about statistical metadata related to official statistics into a framework that can be used by national statistical organizations in their development of metadata systems. It is expected that international statistical organizations would also gain benefits from this work."
2. The framework would be presented in the form of a manual which "provides theoretical and conceptual approaches to a range of topics related to official statistics, together with information on 'best practice' and examples. It would include inventories of applicable standards, terminology and definitions." An ambitious project plan was proposed with METIS 2006 being the point at which the draft framework manual would be discussed and next steps proposed (Phase 4 of the 6 phase project).
3. What progress has been made? The Task Force developed a draft structure for the manual. The four topics for METIS 2006 represent the Parts A to D of the manual. The task of developing draft material for the sections of the manual were allocated

among participants based on their area of interest and expertise. This allocation method saw people working on Parts A, B and D, but no work has been done on Part C - the subject of this session.

4. Part C of the framework has following subheadings:

- C1. Survey Planning and Design
- C2. Survey Preparation
- C3. Data Collection
- C4. Input Processing
- C5. Derivation, Estimation, Aggregation
- C6. Analysis
- C7. Dissemination
- C8. Post Survey Evaluation

II. PURPOSES OF THIS SESSION

5. Given the situation described above, the focus of the session and the papers from participants is to build up a body of work from which material can be extracted for the manual. Participants might point to other material eg previous METIS papers that could be used. [This would be best done by an email containing details of the resource, and directed to the ECE secretariat.]
6. The purposes of the session are to:
- a. discuss and agree on the content in this part of the framework;
 - b. identify and obtain relevant content for the part related to metadata for each stage of the cycle, how they relate (flow) from stage to stage, and the roles of stakeholders in each stage;
 - c. identify examples of implementations, for example:
 - how data elements defined in C1 (both collected and derived) are used in subsequent stages and are made available for C7
 - creation, capture and dissemination of quality metrics (ie metrics such as response rate - from C3, edit error rates - from C4, RSE's - from C5) are captured and used in C6, C7, C8 and C1
 - process metadata - to drive processes eg time series analysis, imputation
 - d. discuss architecture, design, system development, implementation, administration etc issues of end-to-end metadata management; and
 - e. to identify people to undertake further work on this Part.

III. PROPOSED SESSION OUTLINE

7. The structure of the session will be:

1. Introduction

2. Invited papers

* Statistics New Zealand - Gary Dunnet. "Statistics New Zealand's End-to-End

Metadata Life-Cycle"

* Statistics Canada - Alice Born. "Statistical cycle focus in metadata captured for surveys"

* Statistical Office, Slovenia - Matjuz Jug. "Using Metadata in the Statistical Processing Cycle - the Production Tools Perspective"

* Statistics South Africa - Ashwell Jenneker. "Reengineering projects focussing on metadata and the statistical cycle"

3. *Supporting papers*

* Statistics Finland - Harri Lehtinen. "Dissemination of Statistical Data and Metadata - Process based on Common Structure of Statistical Information (CoSSI)"

* Statistics Netherlands - Harry Goossens. "Metadata as a crucial starting link in new statistical cycles"

* Australian Bureau of Statistics - Graeme Oakley. "Quality Infrastructure System - a Case Study of an E2E Application at the ABS"

4. *Discussion*

IV. ISSUES FOR DISCUSSION

8. It is suggested that the following matters be discussed after the presentation of papers.
 - (a) There has been no development of Part C of the manual beyond the high level break-down list in the Introduction. Does the group agree with the proposed high level content ie C1 to C8? Should there be common sub-headings used in each section and what should they be? To start discussion, here is a proposal:
 - n.1 Introduction - explain the processes covered in the phase of the statistical cycle
 - n.2 What metadata is used and created during those processes
 - n.3 Matters to consider that are relevant to this processing phase eg design issues, architectures, stakeholders and roles
 - n.4 Case studies and references to useful material from any source
 - (b) Do any METIS members have contributions to make to any section eg case studies, examples, additional materials? [These would best be contributed in writing either at or just after the meeting, although there may be something worth clarifying during the discussion.]
 - (c) There is no draft document for this Part of the framework manual (as there are for other Parts) and so we need to discuss 'next steps' to prepare a draft. Who is prepared to volunteer for this work? What would the meeting see as the next steps? [The overall timeframe will need to be considered in the broader context of work on all Parts and will be a separate discussion, probably on Day 3.]

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 (iii): Metadata and the Statistical Cycle

STATISTICS NEW ZEALAND'S END-TO-END METADATA LIFE-CYCLE

Invited Paper

Submitted by Statistics New Zealand, New Zealand¹

I. INTRODUCTION

1. Statistics New Zealand has embarked on a major change initiative focussed on its business processes and systems, called the Business model Transformation Strategy (BmTS). The BmTS aims to facilitate the development of processes and systems to deliver on specific outcomes and priorities, while informing and driving the design and building of new capability and competency models for the future organisation. The paper covers Statistics New Zealand's view of the end-to-end statistical cycle, describes where metadata fits in, and gives an example of an implementation of the proposed data and metadata life-cycle approach.

II. INTRODUCTION TO THE BUSINESS MODEL TRANSFORMATION STRATEGY.

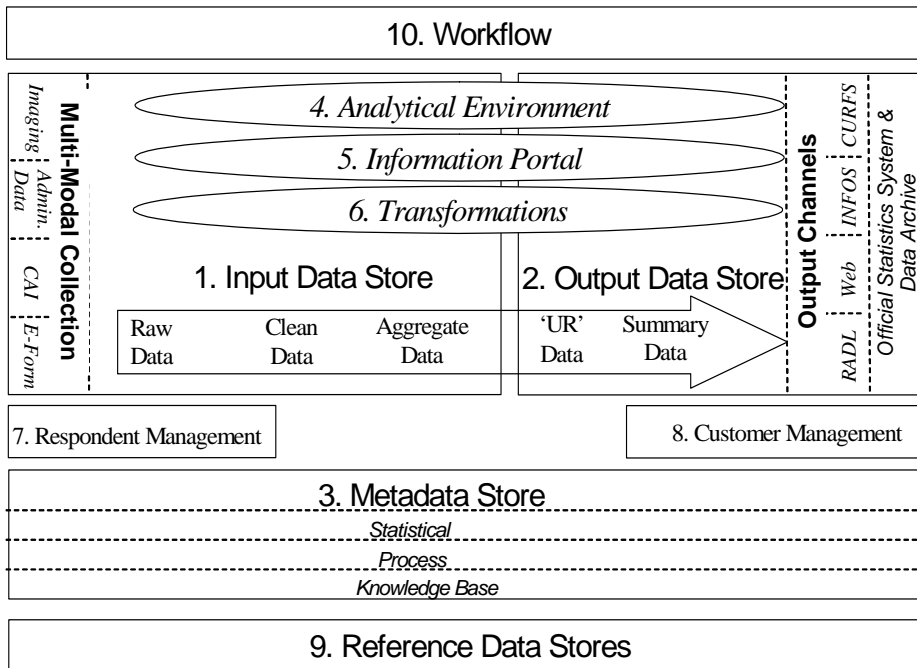
2. The Business model Transformation Strategy is focused on Statistics NZ business *processes and systems* including:

- a. the identification of a need for statistical output;
- b. the design and build of systems for meeting that need;
- c. the collection, processing and analysis of input data; and,
- d. the production and dissemination of output.

3. Its aim is to improve, streamline and standardise in order to optimise the Statistics NZ business model from end to end.

4. In order to deliver on Statistics New Zealand's organisational key strategies the following conceptual view of the ten key business components that are required to operationalise the new business model has been developed.

¹ Prepared by Gary Dunnet, gary.dunnet@stats.govt.nz; and Rebecca Merrington.



5. The BmTS will be judged successful if, at the end of the three year period (2007/08):
 - a. The ten components of the new business model have been identified, solutions designed, and the solutions are accepted by the organisation (and internationally) as consistent with known best practice or best method;
 - b. The ten components of the new business model have been implemented within the six pilot programmes;
 - c. There is a reduction in the operating cost to produce a statistical output (that are operating on a separate subject matter system) by between 10 – 20% after moving to the new end-to-end business model;
 - d. There is a reduction of 50% in the investment (of time and money) required to implement the end to end processes and systems required for a new statistical output;
 - e. Staff throughout Statistics NZ have a good understanding of the business solutions and that business areas have assessed how the business solutions will impact their functions and business processes; and,
 - f. A high level plan to implement the business model across the remainder of Statistics NZ is developed.
6. In order to meet these success criteria, there are four fundamental elements that are seen as critical to the creation of a new business environment for Statistics NZ. These are:
 - a. A corporate approach of Statistics NZ's data resource to prevail over a collection or even subject specific view.
 - b. A robust metadata environment that supports the full lifecycle of Statistics NZ created data (an end to end view of data) and will facilitate the implementation of the OSRDAC metadata and data repository.
 - c. The definition of a set of generic business processes that will be supported by standard technology environments, that provides the flexibility to adapt to new statistical requirements.
 - d. An organisational structure is created that supports the appropriate balance of data management across the Official Statistical System and the specific needs of the various internal business areas and external researchers.
7. The intention is to provide a new business environment that:
 - a. Facilitates the more challenging data collection, integration and analysis necessary to meet the increasingly complex policy and research needs of government and the wider research community

by abstracting the business users and their business processes from the underlying data structures and database systems, moving our statistical staff up the analytical 'value' chain.

- b. Creates the flexibility to respond to changes in users needs and demands, to make use of new data sources or methods and to provide a flexible range of generic information access methods. While also providing the ability to more easily match and confront data in order to increase the quality of Statistics NZ information.
- c. Increases the accuracy, relevance, coherence and interpretability of statistical outputs. It is intended to increase the value from the investment made in official statistics through improving accessibility to statistical information while preserving the confidentiality of data providers and continuing to reduce respondent burden, particularly sub-populations.
- d. Increases the use of administrative data, to make better use of the data that's available, reducing the number of individual collections or the need for new collections to create new statistics.
- e. Simplifies the migration of data and systems as underlying technologies change.
- f. Builds a professional environment that creates a more satisfying working experience.
- g. Provides a standard environment and uniform systems that will allow staff to quickly get up to speed with new subject matter.
- h. Standardises the skills sets and professional development costs of our staff.
- i. Allows Statistics NZ to provide standard information management tools and services for official statistical purposes.

8. In order to develop the new business environment the following principles will be applied across all processes and systems in the End-to-End Model (E2EM):

- a. A standard business process model and E2EM will be implemented for all new Statistics NZ developments, with the aim of removing stove-pipe developments. A process will be put in place to approve changes, exceptions and additions to the E2EM once it is in place.
- b. The systems and processes implemented throughout each of the ten components elements (modules) will be built in a manner that is generic and well documented, utilising externally available modules where possible. The systems will comply with the I&T Strategy. The 80/20 rule will be used when developing generalised systems - core and clerical work will be covered by the 80%, technical complexities will interface with the developed 80%.
- c. Data and metadata must be known, accessible, secure and useable. Data structures and change history will be documented and managed with a life-cycle focus.
- d. There are two key business requirements:
 - i. The production of quality official statistics; and
 - ii. The ability to produce and maintain a quality research and development data archive environment.
- e. A (yet to be determined) consistent efficient approach to methods, processes and systems, will be used when developing, compiling, accessing, analysing and disseminating statistical information. Efficient processes will enable more time for the analysis and dissemination of statistical information or shorter release cycles.

9. The ten key business components that are required to operationalise the new end-to-end business model need to be considered in the context of both the Business Process Model and the statistical data flow through the organisation. Utilising both these contexts could assist in enabling each business area to assess how the business solutions will impact their functions and business processes as the changes will lead to a very different environment for some areas.

III. THE ROLE OF METADATA IN THE BMTS

10. Statistics New Zealand has always recognised the benefits of robust metadata management, including its role in data integration and managing business processes. In fact, during the last large capability enhancement exercise, the IT Strategy, in the mid 1990's, Statistics New Zealand commissioned Prof Bo Sundgren to provide a strategic direction paper on metadata management. From this work and within the IT Strategy project, Statistics New Zealand created metadata infrastructure like CARS and SIM. This infrastructure has performed exceptionally well in some areas (classifications management) and adequately (at best) in others. However, where there are issues with the existing infrastructure, the issue is not about the

design of the infrastructure, but that this infrastructure is in part, poorly maintained, unstructured, difficult to access, unintegrated, passive and not carefully managed.

11. Essentially, the Business Model Transformation Strategy (BmTS) is designing a metadata management strategy that ensures metadata:

- a. fits into a metadata framework that can adequately describe all of Statistics New Zealand's data, and under the Official Statistics Strategy (OSS) the data of other agencies
- b. documents all the stages of the statistical life cycle from conception to archiving and destruction
- c. is centrally accessible
- d. is automatically populated during the business process, where ever possible
- e. is used to drive the business process
- f. is easily accessible by all potential users
- g. is populated and maintained by data creators
- h. is managed centrally

12. From this list, it is obvious that one of the most exciting aspects is that in this environment metadata will be created, used and stored throughout the end-to-end business process. Metadata becomes the heart of "How we do things" as professional statisticians.

13. The use of metadata to drive business processes is not a new idea, but is yet to be fully realised in other statistical agencies. This is because the ideal requires an extensive understanding of the way an organisation operates, how it should operate, and the relationship between these processes and the enterprise's architecture. Under the BmTS, Statistics New Zealand is currently addressing both its business processes and the infrastructure needed to support these processes. As a result, end-to-end data/metadata management will become a reality.

14. Even though this work is pioneering, Statistics New Zealand is convinced that end-to-end metadata management is the only way to fully realise Statistics New Zealand's leadership of the official statistics system and need for standardisation, simplification and integration of statistical information.

15. From investigation of international best practice, and an understanding of the Statistics NZ's existing metadata infrastructure, it is likely that the following will be developed:

- a. metadata requirements inventory (collected from metadata users)
- b. roles and responsibilities charter establishing metadata owners, populators, maintainers and regulators
- c. a policy framework that provides the right incentives
- d. comprehensive metadata model (to define the content of metadata)
- e. standard data item definition framework/template
- f. thesaurus of statistical terminologies which define and organise terms into a structure
- g. taxonomy of relationships between the terminology structure and other metadata and data
- h. metadata storage plan (this may require the establishment of separate repository/ies)
- i. Extract Transfer and Load (ETL) plan establishing methods of metadata population
- j. metadata registry (to organise the metadata defined in the model)
- k. metadata 'viewer' that ensures that data users have easy access to relevant metadata in a timely manner
- l. classifications repository (CARS already exists)
- m. questions Library
- n. metadata store
- o. training and implementation plan

16. Metadata is the cornerstone of the other initiatives within the BmTS, including:

- a. establishing robust metadata management is essential to the structure of Input and Output Data Environments;

- b. understanding what metadata is required to be used, created and potentially stored throughout the end-to-end business process model and then ensuring systems deliver the required outcome (e.g. Respondent Management);
- c. as the driver of the statistical and operational Transformations and processes;
- d. providing access to information (Information Portal, Analytical Environment and Dissemination channels); and
- e. ensuring metadata is reused and updated from different systems/stages in the business process life cycle.

17. While developing an integrated and dynamic metadata management system has many benefits, some of the risks and issues to be managed include, the size of the initiative and that from initial investigation there does not seem to be a one stop shop solution to metadata management.

IV. THE SHAPE OF METADATA AT STATISTICS NEW ZEALAND

18. Statistics New Zealand has defined its metadata in terms of the "Developing a Common Understanding of Standard Metadata Components - A Statistical Glossary" (Joint UNECE/EuroStat Work Session on Statistical Metadata (METIS), March 2002), along with the Australian Bureau of Statistics model. Metadata is defined in terms of being 'definitional', 'procedural', 'operational', 'systems' or 'dataset' (examples given below).

19. In order to ensure that metadata is relevant and useful, it is critical that Statistics New Zealand understands its audience. In 2002, Gareth McGuinness undertook an Statistics New Zealand audience analysis project ("Full Audience Analysis Paper") and identified several broad user groups (Public, Professional, Technical). It is considered necessary to include a fourth group (System) to facilitate a move to a more 'active' metadata environment. So our metadata initiatives recognise the following metadata audiences:

- a. Public - the 'person on the street'
- b. Professional - analysts and researchers
- c. Technical - academics and statisticians
- d. System - computers driving statistical or business processes

20. As a result of technology and prevalent thinking in the mid-1990's, Statistics New Zealand faces the following issues with its existing metadata infrastructure:

- a. metadata is not kept up to date
- b. metadata maintenance is considered a low priority
- c. metadata is not held in a consistent way
- d. relevant information is unavailable
- e. there is confusion about what metadata needs to be stored
- f. the existing metadata infrastructure is being under utilised
- g. there is a failure to meet the metadata needs of advanced data users
- h. it is difficult to find information unless you have some expertise or know it exists
- i. there is inconsistent use of classifications/terminology
- j. in some instances there is little information about data, where it came from, processes it has been under or even the question to which it relates

21. These issues have resulted in the corporate realisation that the significance of metadata to the effective performance of a National Statistics Office has been overlooked with the result that the current metadata infrastructure is not dynamic enough nor suitably integrated.

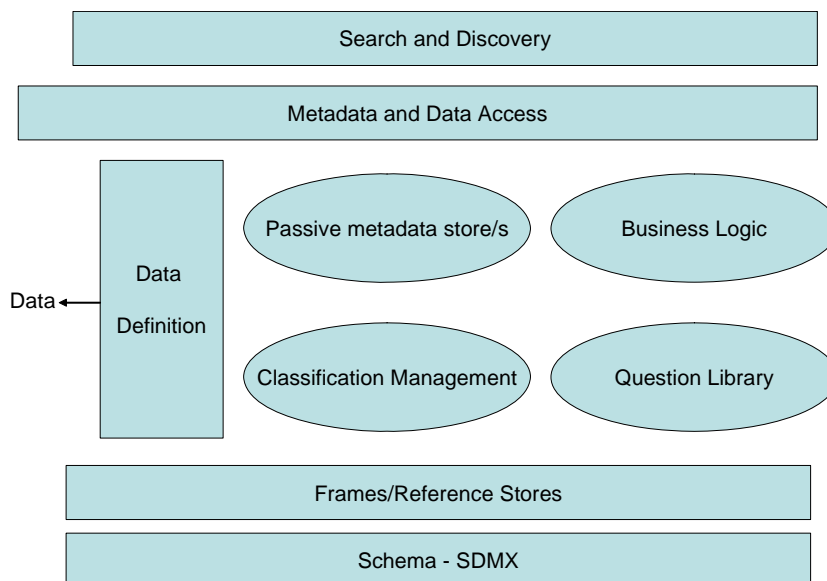
22. However, although some of the existing metadata infrastructure remains passive and unintegrated, the value of the existing metadata infrastructure should not be discounted. Therefore, the metadata management strategy, while addressing the issues highlighted, wherever appropriate re-use either the existing infrastructure or the principles/structure behind it. This is particularly true of Statistics New Zealand's classification management tool - Classifications and Related Standards (CARS).

23. From an assessment of internal and external best practice. The following principles of metadata management have been established. These principles include:

- a. metadata is centrally accessible
- b. metadata structure should be strongly linked to data
- c. metadata is shared between data sets
- d. content structure conforms to standards
- e. metadata is managed from end-to-end in the data life cycle.
- f. there is a registration process (workflow) associated with each metadata element
- g. capture metadata at source, automatically (where possible)
- h. establish a cost/benefit mechanism to ensure that the cost to producers of metadata is justified by the benefit to users of metadata
- i. metadata is considered active
- j. metadata is managed at as a high a level as is possible - managing at the lowest level is prohibitive
- k. metadata is readily available and useable in the context of client's information needs (internal or external)
- l. tracking the use of some types of metadata (eg. classifications)

24. Over the last six months a comprehensive assessment has been undertaken of metadata requirements across both Statistics New Zealand and the broader Official Statistical System (OSS). These requirements have been assess against a number of international models (incl. Dublin Core, ISO 17369 (SDMX), ISO / IEC 11179, ISO 19115 etc.) and we are starting a prototype based around SDMX (Statistical Data and Metadata eXchange). We expect to use SDMX (along with the associated Common Vocabulary Library) as the base for our metadata storage and management.

25. Through the user requirements, the following key metadata infrastructure model has been developed. This model is being continually tested against the enterprise wide metadata requirements. The model is also being used to guide the scope of some small 'metadata related' projects; for example, metadata capture to feed the data definition around the Input Data Environment.



V. THE FIT OF METADATA INTO THE BUSINESS PROCESS MODEL.

26. Statistics New Zealand has created a generic 'as is' Business Process Model (BPM) consists of seven key stages - from the identification of a need at one end, through to the dissemination of statistical outputs to meet that need at the other end. The target end to end model that the BmTS aims to operationalise will focus our resources on the Need, Analyse and Disseminate stages, and reduce the resources spent on the intervening stages (Design, Build, Collect, and Process) once these have been rationalised, streamlined and standardised.

27. Metadata management and the Metadata Store supports all phases of Statistics New Zealand's agreed 'As is' business process model from end-to-end. A more specific discussion about the metadata produced and used at each stage of the business process is outlined below.

[illegible]

VI. PROPOSED METADATA FRAMEWORK

28. The following is the proposed metadata framework. This framework takes the definition of metadata established above and then discusses different attributes of the definition.

Attributes of Metadata	Categories	Types of Metadata				
		Definitional	Procedural	Operational	Systems	Dataset
		<ul style="list-style-type: none"> statistical units/population scope and coverage definition concepts classifications data items - names and definitions, value sets statistical terminology, glossary, thesaurus 	<ul style="list-style-type: none"> objectives collection instruments , instructions, CATI scripts, IFU scripts infrastructure systems used methodologies sample method interviewer instructions imputation and estimation methods question modules 	<ul style="list-style-type: none"> response rates, sample size and distribution sample frame information edit failures coding quality information quality metrics release and various approval sign-offs notes to assist with approval sign-off for dissemination statistical release information 	<ul style="list-style-type: none"> edit rules derivation rules coding rules physical field/database descriptions imputation and estimation rules 	<ul style="list-style-type: none"> datasets - structure, footnotes, titles etc metadata to describe micro and macro output datasets dissemination products annotations technical notes
What is it used for?	<ol style="list-style-type: none"> Assess data quality (all the quality dimensions). Understand how data is created. Identify and locate data Remember production steps. Train new staff. Tune and improve processes. Run software tools. Understand software tools. 	1,3	1,2, 4,5	1,6	1,7	8

Where in the business process is this used?	1. Need 2. Design 3. Build 4. Collect 5. Process 6. Analyse 7. Disseminate (incl. archive)	1,2,3,5,6,7	2,7	2,4,5,7	2,4,5,6,7	2,3,6,7
Who defines it?	1. Statistical Designers (IT, Methodologists, Questionnaire) 2. Input data providers 3. Production statisticians 4. Software tools	1	1,3	2,3	1	4
Who uses it? (in general, not related to access rights)	1. Public 2. Professional 3. System 4. Technical - External 5. Technical - Operational	1,2,3,4,5	3,4,5	3,5	3,5	3,5
How is it used?	1. Active 2. Passive	1,2	1,2	1,2	1	1
How is it populated?	1. Automatic 2. Manual	1,2	2	1	1,2	1
What format is used?	1. Structured 2. Unstructured	1,2	1	1	1	1

VII. CONCLUSION

29. Robust metadata management is the key to the success of the BmTS. Collecting comprehensive requirements and learning from past and international experience will ensure that Statistics New Zealand develops a robust metadata management system.

**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 (iii): Metadata and the Statistical Cycle

METADATA TO SUPPORT THE SURVEY LIFE CYCLE ¹

Invited Paper

Submitted by Statistics Canada,² Canada

¹ The opinions expressed in this paper are those of the author and do not necessarily reflect the official policies of Statistics Canada.

² Prepared by Alice Born, Statistics Canada. Contact: alice.born@statcan.ca The author wishes to acknowledge the contributions from Amie Lee, Carmen Greenough and Susan Ingram of Statistics Canada in the preparation of this paper.

I. INTRODUCTION

1. The Integrated Metadatabase (IMDB) is the corporate repository of metadata for each of Statistics Canada's current 567 surveys. These surveys are the Agency's core activities and the IMDB is the principal mechanism by which they are documented, providing a key information resource for data users and for corporate knowledge management. Under Statistics Canada's Policy on Informing Users of Data Quality and Methodology³, data users are provided descriptions of the underlying concepts being measured, and the methodology and indicators of the quality of data Statistics Canada disseminates. Metadata and related documentation must conform to standards and guidelines issued under this policy. Also, Statistics Canada's Quality Guidelines⁴ provide guidelines on how to describe the concepts and on indicators of data quality that should be reported as part of the metadata.
2. The content of the IMDB is organized around the survey entity. For the purposes of the IMDB, the term "survey" refers to the collection, analysis and reporting of data concerning characteristics of a population. These data may be collected directly from survey respondents, derived from other Statistics Canada surveys and/or collected from administrative files.
3. The IMDB contains statistical metadata and includes both reference and structural metadata.⁵ Reference metadata describe key administrative characteristics, data sources, methodology, and measures of data accuracy of the survey. Structural metadata (or definitional metadata) refer to variables and their definitions, and related classifications. The objective of this paper is to describe, in detail, the common set of reference metadata related to the survey life cycle as presented in the IMDB. The paper illustrates one metadata model developed at a national statistical office in order to meet the requirements for disseminating statistical metadata to its data users. However, as the demand for statistical metadata, particularly from international organizations, increases, there is growing internal pressure to reuse existing metadata in the IMDB and add administered items to the model to fill these requirements.
4. The paper presents a description of the IMDB; details on the metadata supporting a survey life cycle; tools for entering the metadata; versioning rules in the IMDB; and additional administered items external to the IMDB and for future development.

II. DESCRIPTION OF THE IMDB

5. Statistics Canada has implemented a corporate metadatabase that stores metadata on its 566 current surveys and statistical programs. The IMDB contains another 312 records for surveys in various states (e.g., surveys with no publicly disseminated data, amalgamated, etc) for historical purposes. The content of the IMDB has been selected to suit its primary purpose, which is to provide users with information needed to interpret the statistical data that Statistics Canada disseminates. The type of information provided covers the data sources and methods used to produce the data published from surveys and statistical programs, indicators of the quality of the data as well as the names and definitions of the variables, and their related classifications. The metadata supports all of the Agency's dissemination activities including its online data tables, CANSIM and Canadian Statistics, publications and daily data releases. The IMDB has been built to facilitate the maintenance of historical statistical metadata as well as providing a snapshot of the metadata at any survey instance as far back as November 2000 – the starting point of the IMDB.
6. The IMDB model is based on the ISO/IEC 11179 specification and standardization of data elements, the Corporate Metadata Repository (CMR) from the U.S. Bureau of Labor Statistics, and an extension of the American National Standards Institute metamodel (ANSI X3.285). ANSI X3.285 was recently incorporated into the ISO 11179 standard. The CMR model consists of a data dimension, business dimension, administration

³ <http://www.statcan.ca/english/about/policy/infousers.htm>

⁴ <http://www.statcan.ca/english/freepub/12-539-XIE/index.htm>

⁵ Terms taken from SDMX documentation. Structural metadata refers to the description and identification of statistical data and reference metadata describes and qualifies statistical datasets and processing more generally.

and document dimension, and terminology and classification dimension.⁶ For purposes of this paper, Statistics Canada's application in the IMDB of the business dimension of the CMR, which supports the survey life cycle activities of statistical offices, is described in detail.

7. The database is implemented in Oracle 9i and is resident on a central server. Currently, the metadata is published on the Statistics Canada website⁷ on HTML pages generated from Perl scripts and Oracle PL/SQL. These HTML pages are the basis for dynamically generated web pages that directly access the database. The database is kept up to date through a graphical user interface (GUI) tool, implemented in Java, and deployed over the Internet to desktops of the stewards of the metadata (e.g., Standards Division and selected survey divisions in Statistics Canada). Updates are quality assured and registered before they are made available for generation of the external HTML pages.

A. Metadata supporting the Survey Life Cycle

8. The IMDB model defines the entities for describing Statistics Canada's surveys and statistical programs, their content and their methodology, and the relationships between them. The model supports the metadata requirements of many of the phases of the survey cycle including survey design, data collection, input processing, derivation, estimation, aggregation, dissemination and post-survey evaluation.

9. The basic structure of the metadata in the IMDB is illustrated in Figure 1. Each entity is referred to as an administered item. Each of the administered items currently in the IMDB represents a part of the survey life cycle⁸ and the Data Dimension of the CMR (i.e., data elements and value domains). Administered items are defined, and may be reused or shared; and they are also managed, tracked and organized. In order to complete the latter, each administered item is supported by the following "regions", outlined in red in Figure 1. The *stewardship* region (e.g., organization, contact and documentation) supports the administration aspects of the administered item such as the responsible division and information for registration as well as supporting documentation. The *identification* region (e.g., identification and time frame) manages the name of the administered item and the time context for the administered item. The *classification* region (e.g., keyword and themes) manages the classifications and keywords to which administered items are assigned. In Statistics Canada, some administered items (e.g., surveys and questionnaires), data tables, data releases and publications are organized around 27 top themes and 221 sub-themes. Those administered items shown in grey have not been implemented in the current version of the IMDB.

10. In Figure 1, the administered items have been grouped into items that support information about the survey and its "umbrella" statistical activity; the survey methodology; and data elements. The green arrows show some of the relationships between these administered items. In the model, all the administered items describing data sources and methodology (i.e., methodology box) are attached to the survey instance; survey instances are linked to the survey; and data elements (variables) and value domains (classifications) are linked to the data file.

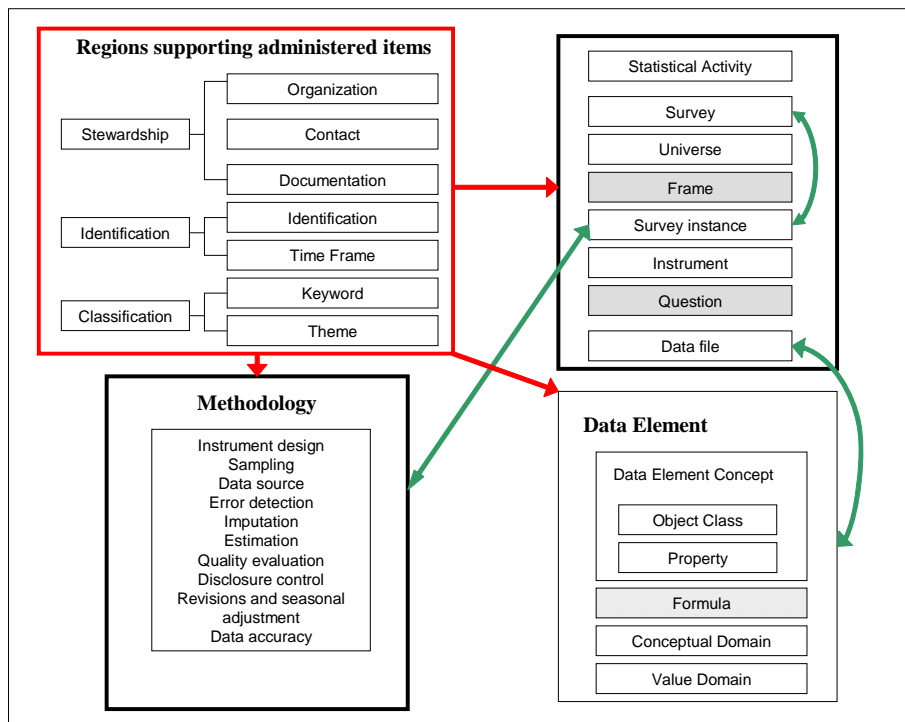
11. The administered items in the current version of the IMDB related to the survey life cycle and covered in this paper are: 1. Statistical activity; 2. Survey; 3. Instance; 4. Universe; 5. Instrument; 6. Methodology; 7. Documentation; and 8. Data Files. These administered items reflect the mandatory requirements for reporting information on data sources, methodology and data accuracy for each survey as stated in the Policy on Informing Users of Data Quality and Methodology.

⁶ Johanis, Paul and Dan Gillman, 2006: Metadata Standards and Their Support of Data Management Needs, Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), Geneva, April 3-7, 2006.

⁷ See www.statcan.ca and in particular, Definitions, data sources and methods module, <http://www.statcan.ca/english/concepts/index.htm>.

⁸ The administered items supporting the survey life cycle in the IMDB match well to the proposed components in Part C of the METIS framework for statistical metadata (see Graeme Oakley, 2006). The IMDB also closely follows the administered items in the Business Dimension Model of the CMR.

Figure 1: Relationship of the survey process to the administered items in the IMDB.



B. Statistical Activity

12. The Statistical Activity administered item in the IMDB represents groups of surveys that share some common features and for which some common explanatory text would be useful to data users. For example, a Statistical Activity record was created for Statistics Canada's Unified Enterprise Survey (UES) Program, which contains general information applicable to 200 separate business surveys into a single master survey program. Another example is the Canadian System of National Accounts.⁹ Not every survey needs to be linked to a Statistical Activity and is only created in consultation with subject-matter areas.

C. Survey

13. The content of the IMDB is organized around the survey entity as opposed to datasets as in other metadata models such as SDMX. A survey, in order to be considered as a record in the IMDB, is defined as a statistical activity that involves the collection, compilation and publication of statistical data measuring characteristics of a population. In the IMDB, surveys are defined as three types:¹⁰

- **Direct:** microdata are collected directly from a respondent with the use of a Statistics Canada collection instrument (e.g., Labour Force Survey);
- **Administrative:** microdata are extracted from administrative sources from an external organization, which were originally collected for their own purposes (e.g., vital statistics from provincial and territorial governments); and

⁹ <http://www.statcan.ca/cgi-bin/imdb/p2SV.pl?Function=getSurvey&SDDS=1735&lang=en&db=IMDB&dbg=f&adm=8&dis=2>.

¹⁰ For purposes of presentation, surveys are referred to as "surveys and statistical programs" on the IMDB web pages as a way of representing all three types of surveys.

- **Derived:** data are derived from other Statistics Canada surveys or other data sources to produce datasets of new derived variables (e.g., national accounts, Gross Domestic Product, price indexes).¹¹

14. The following guidelines are used to determine whether or not a “statistical activity” is a survey, and therefore requiring a record in the IMDB.

15. Activities producing clean microdata serving as data sources to surveys or to analytical studies, and for which no aggregated data are published, do not constitute surveys for the IMDB purposes. Direct surveys and administrative data can be active, discontinued or be conducted one-time only. Derived surveys can only be active or discontinued. One-time only derived statistics are considered as an analytical study and are therefore considered out of scope for IMDB. A compilation of selected data collected from direct surveys or administrative sources does not constitute a derived survey, even if it is produced on an on-going basis. In general, these statistical compendia, such as Statistics Canada’s *Canadian Economic Observer*, are treated as a product and not as a survey.

16. Contrary to direct and administrative surveys, it is difficult to establish clear operational criteria for the designation of derived statistics. The extent of the transformation of the source data to produce new information is the critical factor, which cannot be quantified to establish an absolute rule. In the case of an activity drawing on data collected by others to produce a new dataset, one must ask oneself if referring the users to the metadata in the IMDB on the source surveys will adequately inform them on the quality and methodology of the product. If the answer is yes, the creation of a new derived survey and associated metadata is not called for; if the answer is no, then the statistical activity leading to the product should be designated as a derived survey.

17. The survey administered item contains the following information: the title and acronym of the survey, (i.e., Monthly Survey of Manufacture (MSM)); an overview of the survey that provides a description of the objectives of the survey, the survey population, for whom the data are intended and the use of the data; and the status of activity on the survey (e.g., active, discontinued, transferred or one time only). All surveys are assigned an identification number, known as the Statistical Data Documentation System (SDDS) number in the IMDB. Figure 2 shows the actual attributes as stored in the model.

Figure 2. Attributes of the survey administered item.

Survey	
Survey_AC_Id	INTEGER (PK1)
Survey_AC_Version	NUMBER(5, 2) (PK2)
Survey_AC_Name_en	VARCHAR2(1000)
Survey_AC_Name_fr	VARCHAR2(1000)
SurveyOverview_en	VARCHAR2(4000)
SurveyOverview_fr	VARCHAR2(4000)
Survey_Type	INTEGER
SDDS	INTEGER
Mandatory_Type	INTEGER
Longitudinal_Type	INTEGER
Census_Type	INTEGER
Direct_Type	INTEGER
Derived_Type	INTEGER
Administrative_Type	INTEGER
SurveyPurpose_en	VARCHAR2(1500)
SurveyPurpose_fr	VARCHAR2(1500)

¹¹ Derived statistics are referred to as “statistical programs” in the IMDB.

C. Other Administered Items

18. A Survey consists the administered items related to the survey life cycle that are grouped together through an Instance administered item for each reference period of the survey. Table 1 shows these administered items and their respective definitions. The indentations of each item in the table illustrate the hierarchical relationship between the entities. These administered items are reused or updated in successive survey instances or shared with other surveys and statistical activities in the generation of web pages. Through the Policy on Informing Users of Data Quality and Methodology, the IMDB has a **common metadata set**¹² that is reused for each survey. It is proposed that the administered items in the Statistics Canada metadata set be considered by other national statistical offices and international organizations as a “best practice” to help move towards a common set of metadata items to be used for various metadata exchange initiatives.

19. On the Statistics Canada website, users have access to metadata for each instance of each survey or statistical program for which data are disseminated. The administered items stored in the IMDB have been organized to present general information on the survey (survey title, status, frequency, record number and survey mandate) and metadata related to the survey life cycle for a survey instance (e.g., reference period, data release date, survey instrument (questionnaire), variables, survey description, data sources, methodology, data accuracy, documentation and data file (available internally only)). Quality metrics such as response rates and coefficients of variation are disseminated under Data Accuracy.

20. Figure 3 presents the web page for a survey instance, in this case, the Annual Survey of Manufactures. On the left hand side bar of the web page, there is additional information including links to Summary of changes over time and Other reference periods. The Summary of changes over time presents a chronology of changes to administered items as well as the start date of the survey and Other reference periods gives users access to metadata for other survey instances for which data have been released. We are currently developing a calendar of surveys in the field, that is, those surveys without disseminated data but with metadata for the some administered items including survey, instance, collection instrument, instrument design and collection method. In the Instance administered item, the time frame entity contains fields for the actual start and end dates for data collection of that survey instance. Using these dates, a list of surveys that are currently collecting data will be created and disseminated on both the Definitions, data sources and methods and Information for survey participants modules.

21. The release of data is announced in the *Daily*, the Statistics Canada's official release bulletin. This announcement triggers the need to update the metadata and create a new instance of the IMDB survey record. Release dates for “mission critical” surveys are posted a year in advance and other releases, two weeks in advance. Based on this information, the manager of the IMDB contacts the survey managers in advance for updated metadata.

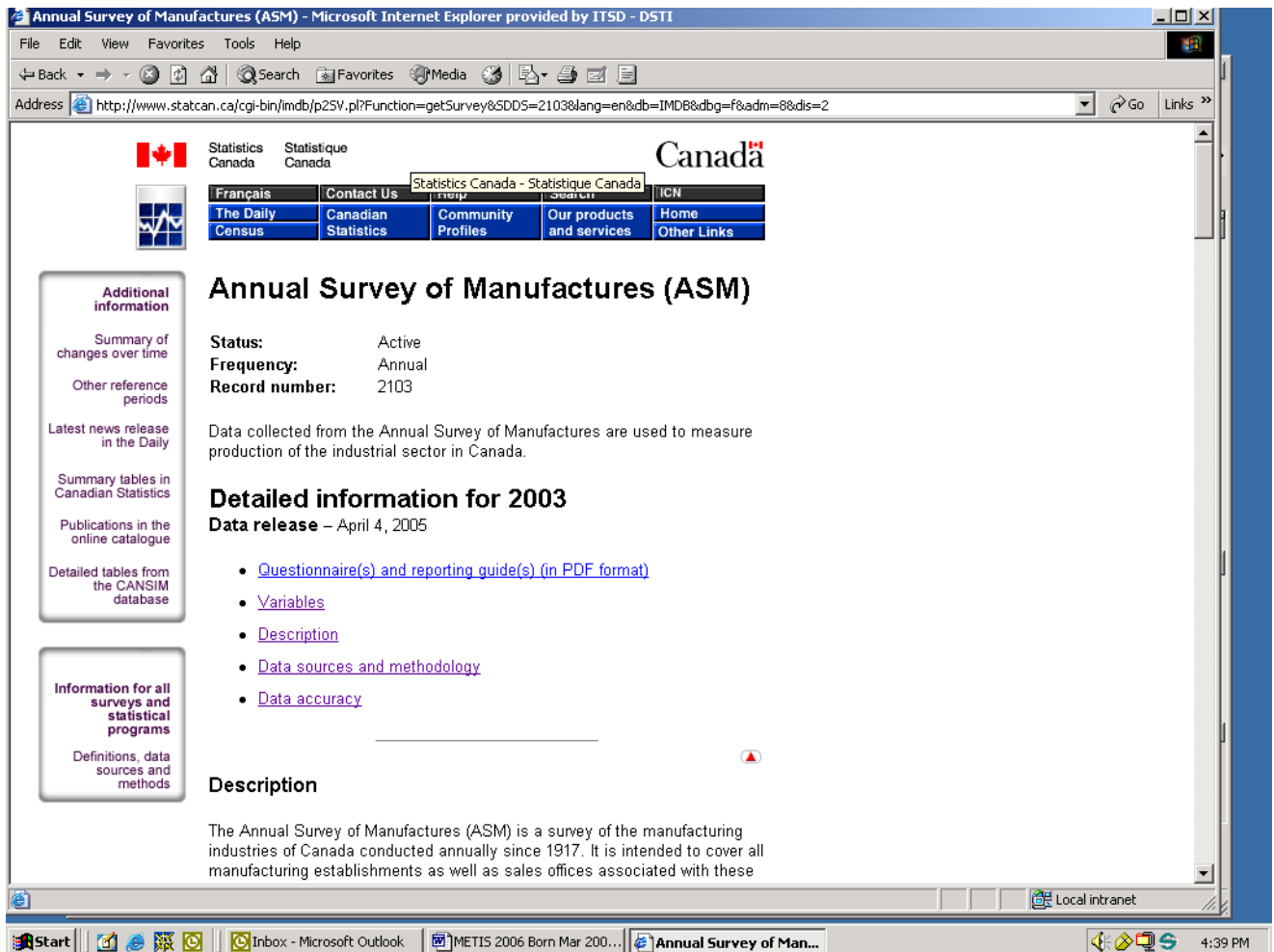
¹² The term metadata set is taken from the SDMX initiative. **Metadata set** is a set of information regarding almost any object that describe the maintainers of the data and structural definitions; describe the schedule on which data is released; describe the flow of a single type of data over time; describe the quality of the data, etc. In SDMX, the creators of reference metadata may take whatever concepts they are concerned with, or obliged to report, and provide a reference metadata set containing that information. Statistical Data and Metadata Exchange Initiative (SDMX), 2005: Framework for SDMX Technical Standards (Version 2.0), November, 2005.

Table 1. Administered items in the IMDB supporting the survey life cycle and their definitions.

IMDB administered items	Definition
Statistical activity	The statistical activity is groups of surveys that share some common processing system or conceptual framework. For example, a statistical activity was created for the Unified Enterprise Survey Program, which contains general information applicable to all surveys in the UES program. Not every survey needs to be linked to a statistical activity. A statistical activity record is created in consultation with subject-matter areas.
Survey	A survey is a statistical activity that involves the collection, compilation and publication of statistical data measuring characteristics of a population. Includes direct surveys, administrative surveys and derived surveys. Provides administrative details about the survey.
Target population (or universe)	The population of units that is actually covered by the survey. Identifies the statistical unit and any of its relevant characteristics that are used (e.g., Canadian population aged 15 and over and not residing in institutions; or, establishments in NAICS industry XXXXXX with revenues over a certain threshold). Where applicable, any differences between the survey population and the target population (i.e., population for which information is desired) of the survey are described.
Survey Instance	Refers to each time the survey process occurs (i.e., each cycle of a survey). For each reference period, a new version of the instance record is created. For example, for a monthly survey, the IMDB will contain one instance record for every monthly cycle of the survey.
Collection instrument	The vehicle used for the collection of data. For direct surveys, it is a questionnaire. For administrative surveys, it is the record layout of the input record. It is not applicable for derived surveys. A questionnaire can be in many forms such as paper version, electronic, etc. Each questionnaire is linked to the instance record to which it pertains. The questionnaire image is copied into IMDB in pdf format.
Methodology	A text description of each of the following aspects of the methodology of a survey.
Instrument design	Method used to design, test and implement survey instrument. It only applies to direct surveys. Description of the design; methods for testing the questionnaire (e.g., review committee, focus group, pilot survey, etc.); and date of last revision of the instrument.
Sampling	Description of the survey units, any stratification used and the sample selection methods. It does not apply to derived surveys.
Collection method	Details on the collection methodology and the type of instrument. Details on method of initial contact and follow-up. Also included is a description of the capture method. For administrative and derived surveys, this item can be used to describe data sources. Collection methods include: data collected directly from respondents with a use of a collection instrument include: electronic data interchange; respondent completed – paper format (mail or fax); respondent completed – touch-tone telephone; Computer Assisted Telephone Interview (CATI); or Computer Assisted Personal Interview (CAPI) methods; or data extraction from administrative files provided by an external organization; or data derived from other Statistics Canada surveys.
Error detection	Methods used to detect errors during collection, capture and processing of the data. Provides details on the types of edits used, ratios applied, etc. and identifies at which stage of the survey process it is done (i.e., during collection, or as part of processing.)
Imputation	Process used to replace missing microdata, and invalid or inconsistent responses identified during editing. Provides details on the type of imputation (e.g., manual, automatic, etc.), imputation rates, the method used (e.g., historical, hot deck, donor, etc.), and software used. Usually does not apply to derived surveys.
Estimation	Methods used to produce estimates for the survey population from collected data. It includes non-response macro adjustments, post stratification,

	calibration, weight-share methods, and variance estimation methods (e.g., direct, Taylor, Jackknife, Bootstrap, etc.). In the case of administrative and derived surveys, procedures and models used to produce the indicators are described.
Quality evaluation	Methods undertaken to evaluate the quality of the final data. Procedures include data confrontation with other published sources, re-interviews, reverse record checks or historical trend analysis.
Disclosure control	Measures taken to ensure that data from the survey does not disclose information concerning any identifiable respondent thus maintaining confidentiality of the respondent data. This summary can include for micro data, removal of respondent, content reduction and content modification; for tabular data, sensitive cells, correction methods such as collapses/suppress cells; and revisions by committees.
Revisions and seasonal adjustments	Methods used to adjust estimates in relation to same estimates for prior periods including benchmarking, calendarization or seasonal adjustments; and procedures for regular revisions to data.
Data accuracy	Data accuracy indicators for the survey. These measures include the coefficient of variation for the key variables in the survey, information on coverage error, response rates and any other relevant data accuracy indicators. Includes response bias and error, and processing errors.
Documentation	Documents useful for the users' understanding of the data can be linked and include user guides, data dictionaries, technical notes, etc.
Data file	Information on the location, format and content of clean data master files that are used as inputs to surveys or as outputs of surveys. The IMDB stores information on clean data master files that are produced for each instance of a survey.
Variables	Description of the meaning of a data point. Based on ISO 11179.
Statistical unit	Definition of the unit about which data are collected (e.g., establishment, household, person and births).
Property	Definition of the characteristic of the statistical unit.
Representation class	Describes the specific form of the representation of the property (e.g., type, name, category, value, area, index)
Classification or unit of measure	A set of allowed values that a variable may take. Classifications are used to represent categorical data and units of measure are used to represent quantitative data (e.g., dollars, tonnes)

Figure 3. Web view of the home page of a survey instance for the Annual Survey of Manufactures (2003).



D. Tools for Loading the Metadata

22. Description text for the administered items is entered into a set of input screens, which is internally referred to as Metastat. These input screens are used to capture information that is common to each of the administered items – identification, description, time frame, documentation, classification (e.g., key words and themes), organization and contact information – previously described as “regions” supporting administered items in Figure 1. Some of these have their own codeset. For example, there are different types of time frames built into the model including: effective period, reference period, collection period, data release and last update. Depending on the administered item, a subset of these time frame types is presented on the input screen (see Figure 5). The description text pertaining to each of the administered items is captured on its own set of input screens. Below are input screens for selected administered items (e.g., Survey, Survey instance (cycle) and Collection method (Data source)) with selected “tabs” (e.g., Identification, Time Frame and Description), respectively (Figures 4, 5 and 6).

23. Every system built in Statistics Canada must provide both an English and French interface and editable (data entry) fields for both languages since the metadata published in both languages. When the application is started, the user selects the language of the interface. This allows coded text fields appear in the language of the interface. Every editable field is displayed allowing data entry to be done simultaneously in both English and French.

Figure 4. Input screen for the survey administered item in the IMDB – Identification tab.

The screenshot shows the 'Survey' window with the 'Identification' tab selected. The 'Name' field is 'Annual Survey of Manufactures'. The 'Registration Status' is 'Not specified', 'Administration Status' is 'Preliminary Validation', and 'Dissemination Level' is 'Public'. The 'Registrar's Comments (English)' field contains text about a change at the Universe level. The 'Registrar's Comments (French)' field is empty. At the bottom are buttons for Close, Delete, Save, and Cancel.

Survey

ID: 2103 Version: 2.0

Name: Annual Survey of Manufactures

Directive:

Identification Description Time Frame Documentation Classification Organization

Identification Administration

Registration Status: Not specified

Administration Status: Preliminary Validation

Dissemination Level: Public

Registrar's Comments (English): A change at the Universe level has caused the versioning of this survey record. (cg 28/04/05) - Summary of change note has been captured at universe level. (changed NAICS 1997 to 2002.)

Registrar's Comments (French):

Close Delete Save Cancel

Figure 5. Input screen for survey instance (cycle) administered item – Time Frame tab.

The screenshot shows the 'Cycle' window with the 'Time Frame' tab selected. The 'Name' field is 'Annual Survey of Manufactures - 2003'. The 'Time Frame' tab contains a table with columns: Type, Start (English), End (English), Start (French), End (French), Survey Stat., and Frequency. The table has four rows: Data release, Reference period, Last update, and Display year. At the bottom are buttons for Add, Remove, Close, Delete, Save, and Cancel.

Cycle

ID: 14033 Version: 6.0

Name: Annual Survey of Manufactures - 2003

Directive:

Identification Description Time Frame Documentation Classification Organization

Type	Start (English)	End (English)	Start (French)	End (French)	Survey Stat.	Frequency
Data release	April 4, 2005		4 avril 2005		Not Specifi...	Not Specifi...
Reference period	2003		2003		Not Specifi...	Not Specifi...
Last update	instance versioned Apr. 1/05 ...		-		Not Specifi...	Not Specifi...
Display year	2003		2003		Not Specifi...	Not Specifi...

Add Remove

Close Delete Save Cancel

Figure 6. Input screen for Collection method (data source) administered item – Description tab.

The screenshot shows the 'Data sources' window in the IMDB Administration software. The window has a menu bar with 'File', 'Statistical Elements', 'Methodology', 'Reference Lists', 'IMDB Administration', and 'Help'. Below the menu bar is a toolbar with a 'Data sources' icon. The main area contains fields for 'ID' (14037), 'Version' (4.0), and 'Name' (Annual Survey of Manufactures - 2003 onward). Below these fields is a 'Directive' field. A set of tabs includes 'Identification', 'Description' (selected), 'Time Frame', 'Documentation', 'Classification', and 'Organization'. Under the 'Description' tab, there are two sub-tabs: 'Description' and 'Instance'. The 'Description' sub-tab is active, showing two text areas: 'METHODOLOGY_SUMMARY_EN' and 'METHODOLOGY_SUMMARY_FR'. The English text area contains the following text: 'Data are obtained from two sources: questionnaires that are mailed out and administrative files. Since the survey collects a wide range of information for over 250 manufacturing industries (based on the NAICS), the response burden is substantial. Using administrative files, where possible, reduces both the survey response burden and data collection costs, while maintaining the necessary level of accuracy. Mail out occurs in November of the reference year (for establishments with fiscal year-ends of April to'. The French text area contains the following text: 'Les données sont extraites de deux sources : les questionnaires expédiés par la poste et les dossiers administratifs. Puisque l'enquête permet de recueillir une vaste gamme de renseignements pour plus de 250 industries manufacturières (selon le SCIAN), le fardeau de réponse est appréciable. Le recours aux dossiers administratifs permet de réduire, lorsque possible, le fardeau de réponse ainsi que les coûts de collecte des données, et de conserver le niveau d'exactitude requis.' At the bottom of the window are buttons for 'Close', 'Delete', 'Save', and 'Cancel'.

III. VERSIONING RULES IN THE IMDB

24. Since metadata do not remain static over time, the metadata model must be able to accommodate these changes. For managing the metadata for each cycle of the survey and changes to the survey through its life cycle, versioning or “time travel” has been built into the IMDB. In this section, the procedures for revisions and versioning of administered items are presented.

25. When the content of an administered item requires revisions, several procedures for implementing them are available in the metadata model. These procedures include:

- Create: creation of a new administered item (e.g., new survey);
- Update: replacement of prior content with revised content (i.e., for errors in text and addition of more completed text); and
- Version: new version of the same administered item while the previous version of the administered item is stored in the IMDB.

26. Although the Create and Update functions are relatively self-explanatory, the versioning function consists of different business rules for each administered item. Versioning is used when there is a need to retain changes over time for each item.

27. At the time an administered item record is created in the IMDB, the system automatically assigns an administered item identification (ID) and a version number (i.e., AC_ID 123 AC_Version 1.0 becomes AC_ID 123 AC_Version 2.0). While the administered item ID remains unchanged for the entire life cycle of the administered item, the version number changes each time the item is versioned. Versioning allows the descriptive text or elements of the administered item to be revised without overwriting the previous version of the metadata. Business rules for versioning of an administered item (in bold in Figure 7) have been developed for specific to the types of changes. Below is a more detailed description of the versioning rules.

A. Versioning of a Survey

28. Changes to the characteristics of the survey (e.g., name, status, frequency, type, statistical activity, Collection Registration Number (CRN), theme assignment and divisional assignment) result in the creation of a new version of the survey administered item in the IMDB. The only exception is changes to the survey

objective. A change in the survey objectives (i.e., Survey Purpose attribute) results in the creation of a new survey (i.e., assignment of a new SDDS number in the IMDB). In Metastat, the reason for the versioning of the survey is explained in the Version Revision Description field of the new version under the Identification tab. For example, in the case of the change in survey frequency, the note would state: *“In the past this survey was conducted on a monthly basis. It became an annual survey as of January 2004 because.....”* The changes are recorded as part of the Summary of changes over time for the survey on the web version of the record.

B. Versioning of a Survey Instance

29. The only reason for versioning an instance record is a change in the reference period, which is generally triggered by the release of new data to the public. The instance record is versioned for each cycle of a survey and the links are established to the pertinent administered items that need to be updated for that reference period. An instance record is not versioned for reference periods during which there is no survey activity. The reason for the survey's inactivity is added to the Version Revision Description field under the Identification tab of next active instance. In cases where the survey was conducted but the data were not released, an instance record representing the reference period is still created in the IMDB. Text indicating that no data were released by Statistics Canada is displayed in the Data Release field.

30. For a monthly survey, the IMDB will contain one instance record for every monthly cycle of the survey, for an annual survey, one instance for every annual cycle, and so on. For each reference period, a new version of the instance record is created. This enables unchanged information to be carried over from one reference period to the next and only changed information needs to be updated.

31. Every instance record has various administered items linked to it (Figure 7). These include the collection instruments, the various methodology items and the data file. Some administered items may vary from one period to another, while others may remain stable. The new information in these changed administered items is captured separately and then linked to the instance record. For example, an instrument record is versioned each time a new image is produced even if the change is only the date printed on the questionnaire image. The instance administered item thus becomes a central location of all links to each administered item and provides the full picture of the entire survey process for each reference period. In general, changes to instrument, methodology and data files are released at the same time as a new version of the survey instance is released since changes are all triggered by the release of new data to the public.

C. Versioning the Target Population

32. A change to the target population of a survey results in the creation of a new version of this administered item as well as a new version of the survey. This also includes changes in the statistical unit and classifications. For example, under the Summary of Changes over time for the Annual Survey of Manufactures, the following is presented:

Target population – Prior to reference year 2000, the Annual Survey of Manufactures (ASM) provided estimates of principal financial statistics for all incorporated manufacturing businesses that had employees and had sales of manufactured goods equal to or greater than \$30,000. With reference year 2000, the universe was expanded to cover all manufacturing units. This change added approximately 60,000 units to the ASM universe.

D. Versioning of a Methodological Administered Item

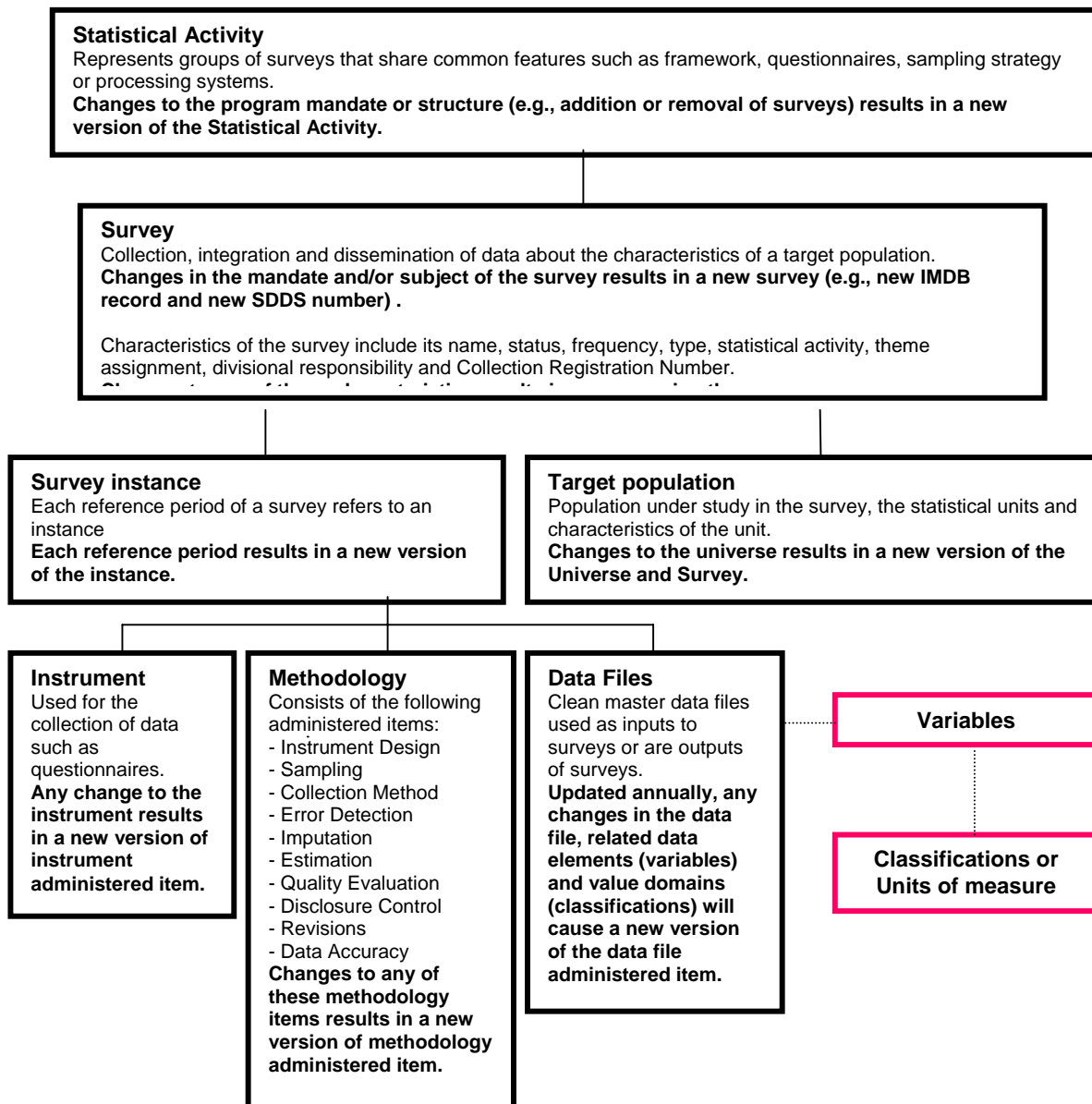
33. A methodology record is versioned only when there is a change of method. Improvement in the text does not constitute a new version of the methodology. For some items of the methodology group, the descriptive text contains reference to specific dates and/or reference periods. For these cases, the methodology records require versioning. In many cases, the sample size is reference-period specific, and the Sampling record is versioned for that reference period. Specific collection dates are also related to a specific reference period and Data sources, which contains these dates, is versioned that survey instance. Data accuracy is another

administered item that is often versioned for every instance of the survey since it contains measures that are reference-period specific.

E. Versioning of a Statistical Activity

34. Versioning of a Statistical Activity occurs when there is a change in the program mandate and/or structure, including the addition of surveys from other survey programs. Any changes to the description of the Statistical Activity or in the set of surveys that are linked to it will cause the creation of a new version of the Statistical Activity record.

Figure 7. Versioning rules for the administered items in the IMDB that support the survey life cycle.



IV. ADDITIONAL ADMINISTERED ITEMS AND FUTURE WORK

35. The Business Dimension Model of the CMR supports a number of additional administered items, which complete the necessary information to describe the survey life cycle. These are not directly stored in the IMDB but are connected through external links.¹³

A. Systems

36. In the CMR, there is an administered item for Systems, which provides metadata for both hardware and software systems. At Statistics Canada, metadata related to this administered item is stored in the Agency's Software Register (SR). The SR is a list of all application systems in use at Statistics Canada and the list of software products which are used in these applications. Although not directly linked to the IMDB at this time, the SR contains the survey identification number (i.e., Statistical Data Documentation System (SDDS) number) and the functional responsibility for managing the survey (i.e., FRC code) stored in the IMDB. However, it would be more useful to link the Agency's SR to the SDDS and its program element (PE) structure, which relates the Agency's expenditures and budget by programs or projects. This integration of the IMDB, the SR and corporate financial planning systems will permit analysis of application development and software diversity across the Agency and identify which surveys are dependent on what applications. The analysis can be used to monitor development (which are capitalized) and maintenance costs of applications; promote better coherence and reuse of software components across surveys and statistical programs; and identify surveys or sets of surveys that may be vulnerable because of dependencies on applications that are beyond their expected life or software that is no longer supported.

B. Products

37. The CMR also contains an administered item class called Products, which is linked to Data Sets. At Statistics Canada, this item is managed by the Common Object Repository (COR), which is external to the IMDB but contains links to the IMDB throughout the Statistics Canada website. Figure 8 illustrates the relationships between the metadata and the various disseminated products. For example, there is a direct link from the data release in the Daily to the survey instance (Figure 9). Surveys and attached metadata, and questionnaires are also organized by the subjects, thereby enhancing accessibility of the metadata for the user (Figure 10).

C. Corporate Planning and Post-survey Evaluation

38. With the IMDB, information on surveys can be reported by FRC (division), by subject, or other dimensions such funding source (cost recovery/base budget). As part of the further enhancement of the IMDB, links to corporate reporting systems are being evaluated particularly to support evaluation and audit of statistical programs.

39. Statistics Canada is currently developing a systematic approach to conducting a post-survey evaluation. Referred to as the Quality Management Assessment (QMA), managers responsible for the delivery of a survey or statistical program will provide a description of the data quality management tools that are used in the various processes (e.g., questionnaire testing, interviewer monitoring); an assessment of the impact of changes (i.e., as the result of a survey redesign) to these processes on quality, costs, response burden and confidentiality; and a description of how accuracy is managed (i.e., how sampling and non-sampling errors are prevented, identified, corrected and evaluated). The QMA will be developed within the Agency's Quality Assurance Framework.¹⁴ It is anticipated that the results of the QMA will provide valuable input to divisional quadrennial

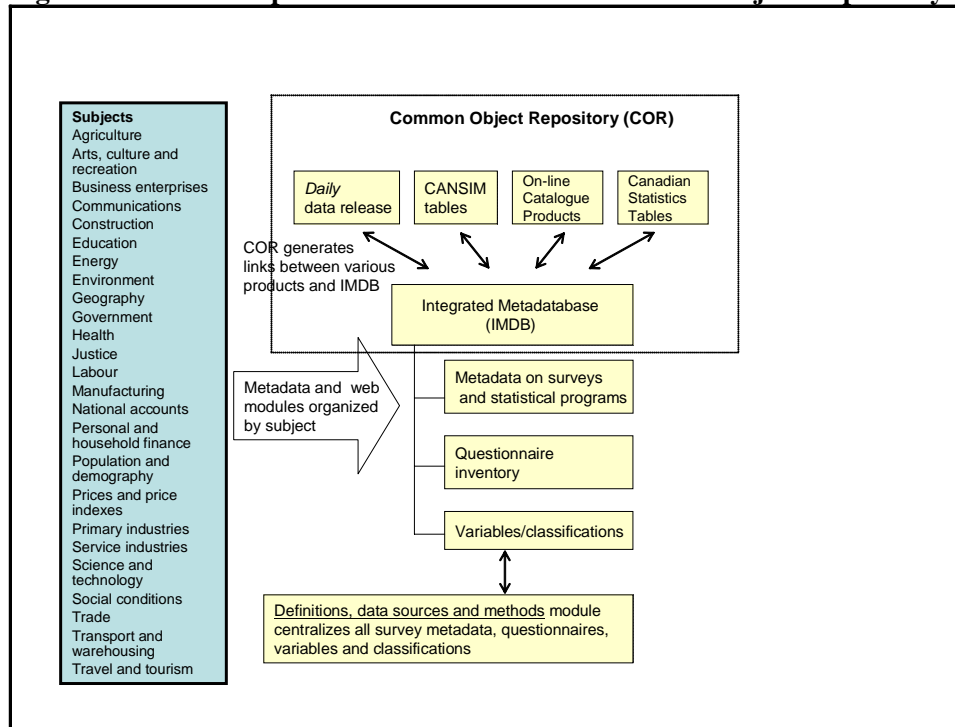
¹³ See Figure 2. Business Dimension Model in Working Paper 7 (Paul Johanis and Dan Gillman, 2006).

¹⁴ <http://www.statcan.ca/bsolc/english/bsolc?catno=12-586-X&CHROPG=1>

and biennial program reviews and identify the need to add administered items and quality indicators in the IMDB to meet the needs for reporting quality management practices in the Agency.¹⁵

40. As part of the QMA project, there will be links to the relevant metadata on data sources, methods and data accuracy stored in the IMDB. Here we can use the IMDB to analyze and evaluate the practices across surveys and statistical programs. For example, we could produce information on how many and what surveys have moved from X-11-ARIMA to X-12-ARIMA for seasonal adjustment, or on the use of generalized systems for imputation or estimation. Other measures of data quality can be assessed across surveys and over time. They may include time lapsed from data collection to data release (timeliness), time required for respondents to complete a questionnaire, response rates, and refusal rates. All of this information is stored in the IMDB.

Figure 8. Relationship between the IMDB and Common Object Repository.



¹⁵ Julien, Claude, 2006: Quality Assessment Management at Statistics Canada, European Conference on Quality and Methodology, Q2006.

Figure 9. Link in the Daily to the survey instance.

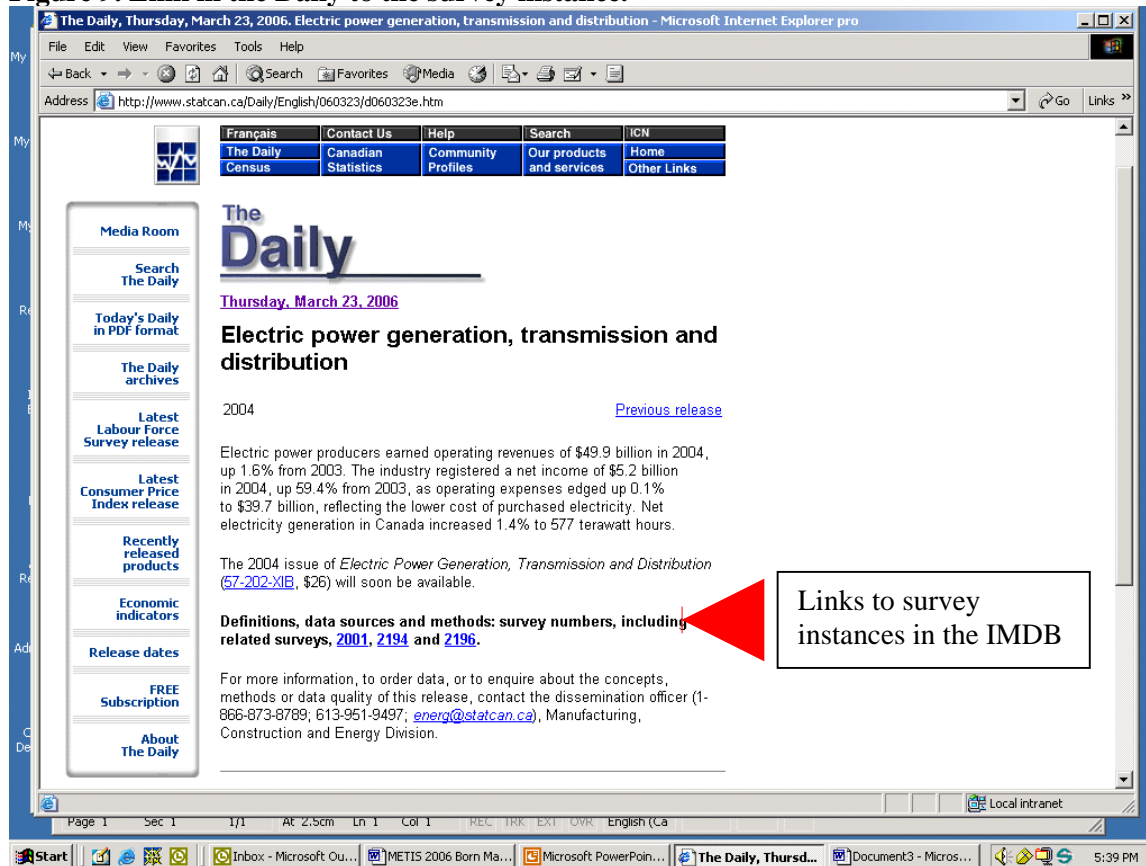
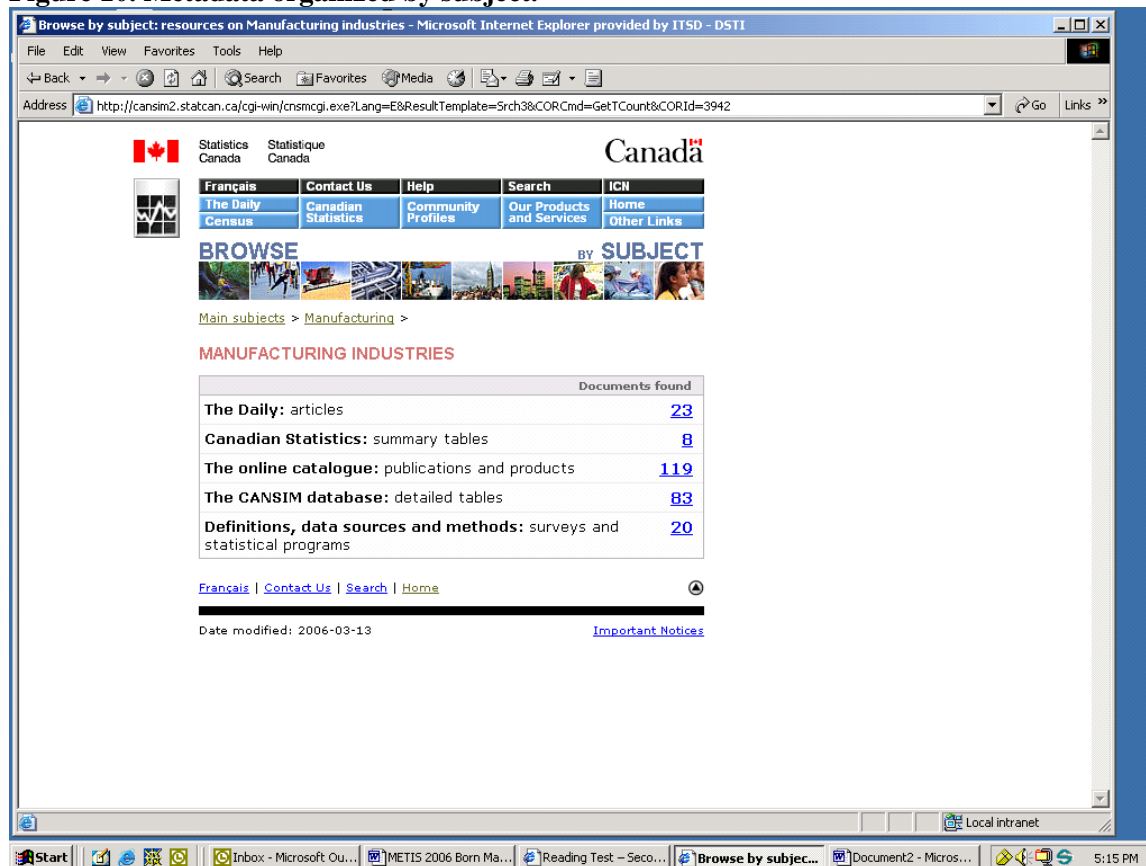


Figure 10. Metadata organized by subject.



V. CONCLUSIONS

41. The IMDB is becoming the single source of metadata for describing Statistics Canada's surveys and statistical programs. This means that survey managers have to supply metadata on their individual surveys only once to the stewards of the metadata, Standards Division. Since IMDB is built on a common metadata set with reusable administered items and attributes, survey managers can reuse the descriptions for different survey cycles and across other surveys they might manage. Also, the use of a common metadata set presents a "common look and feel" to data users accessing the metadata through our website.

42. While the most of the content in the IMDB was determined by the Policy on Informing Users of Data Quality and Methodology, both internal and external users have indicated other requirements when it comes to statistical metadata. The IMDB has been designed and continues to be developed to meet these needs. Now that the metadata is complete for each survey, other users can access those administered items that meet their requirements. In addition to supporting the information requirements for disseminated data, the IMDB is being used as a source of information for standardizing survey processes and content, corporate and financial planning, quality management at the survey level, survey respondents, international data exchanges and data researchers.

43. The purpose of this paper was to provide an example of a set of statistical metadata that supports the survey life cycle as a contribution to the METIS manual on statistical metadata, currently in development. Hopefully, this has been achieved.

VI. REFERENCES

Johanis, Paul and Dan Gillman, 2006: Metadata Standards and their Support of Data Management Needs, Joint UNECE/Eurostat/OECD Work Session on Statistical Metadata (METIS), Geneva, April 3 to 5, 2006.

Johanis, Paul, 2005: Documenting Data Elements in Statistical Agencies, JSM 2005, Minneapolis, August 7, 2005.

Johanis, Paul, 2000: Statistics Canada's Integrated Metadatabase: Current Status and Future Plans, Work Session on METIS, November 28-30 2000, Washington, D.C., United Nations Statistical Commission and Economic Commission for Europe Conference of European Statisticians.

Julien, Claude, 2006: Quality Management Assessment at Statistics Canada, European Conference on Quality in Survey Statistics, Q2006,

Lee, Amie, 2003: Integrated Metadatabase (IMDB) Architecture, Statistics Canada, Ottawa, November 5, 2003.

**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 (iii): Metadata and the Statistical Cycle

USING THE METADATA IN STATISTICAL PROCESSING CYCLE – THE PRODUCTION TOOLS PERSPECTIVE

Invited Paper

Submitted by Statistical Office of the Republic of Slovenia¹

I. INTRODUCTION

1. In 2004, SORS started the action plan for the migration of statistical production from the mainframe computer to a local environment. The production tools used in the local environment offer much better possibilities for use (and reuse) of metadata. The paper will show the usage of different types of metadata in three main production environments used in the office: Blaise, SAS and Oracle.
2. In 2000 we've introduced Gentry tool, giving subject-matter experts possibilities to define the Blaise instrument (working Blaise survey specific data entry-editing application) – not by writing a code, but simply entering the parameters, describing the survey questionnaire and its data fields into the CAPI questionnaire. From this CAPI questionnaire, the survey questionnaire is generated. The “generated” metadata are later used through the process like metadata from other, not generated Blaise datamodels. Generation is based on some common in-house standards for questionnaires. It covers basic metadata (description of fields, blocks, etc.), but it is successful and efficient on a large number of surveys, and at the same time it enforces standardized processing.
3. In 2004 the Statistical Office of the Republic of Slovenia started the project of developing and implementing standard editing and imputation tool, written in SAS. The reason was first of all the growing demand for quick and efficient flow and results of the statistical process and the possibility of cost reduction. Metadata are used for the presentation of input matrices of the edit coefficients and for the presentation of the matrix. The plan is to set up such a system that these two matrices, together with the data, would be the only input for the whole system. Output metadata are quality indicators which indicate the number of values that have been changed during the process as well as the impact of these changes on final results. Through the values of variable indicators it is also possible to see in which stage of the process the data value was changed.

¹ Prepared by Matjaž Jug, Pavle Kozjek and Tomaž Špeh

4. In 2005 the new Oracle database, containing data and metadata for structural business statistics (based on several secondary and primary sources), was built. It is based on metadata rich model with classifications, source indicators, process indicators and variable definitions. The process requires a robust loading mechanism with the possibility to change data sources for existing variables and add new variables, general imputation and editing procedures, standard extraction procedures and analytical possibilities.

II. USING THE METADATA IN BLAISE APPLICATIONS

A. Metadata in Blaise environment

5. At SORS, the Blaise system is used as a general data collection and editing tool, completely covering CATI and CAPI mode of data collection, as well as all data entry and editing from paper forms (about 110 forms) and other data sources, where the microediting phase is necessary.

6. There are two general rules about metadata in Blaise (Pierzchala - Manners, 2001):

- a. all metadata defined once
- b. no data without metadata.

7. The additional Blaise capability of metadata manipulation allows generating customized data export routines and downstream data definitions from the metadata in the instrument (working questionnaire application). Thus it is possible to accomplish multiple modes of data collection (e.g. CATI, CAPI) and data editing with the same system, and export data including metadata in any way you need them, all from one metadata specification.

8. Metadata generated from the Blaise datamodel are in general the basic metadata to describe data structure and support processing. Some components are not mandatory; they depend on the application developer: if a certain part of metadata (e.g. question text) is defined in a datamodel, then it is used for generation of metadata for the next phases in a process. Only the key elements that define the datamodel are required.

B. Gentry, an in-house developed tool for high-speed data entry

9. In general, most of Blaise capabilities concerning metadata are used at SORS. But as a first step to get the survey metadata, the correct Blaise datamodel (as a source of metadata, used later in the process) should be prepared. Usually this should be done by the Blaise application developer. We prefer to have it generated, at least the major part. Currently, our central metadata repository Metis can not provide all information needed to generate the survey Blaise datamodel, and there is no general procedure available to use the metadata from Metis.

10. To bridge the absence of a structured metadata source, we offered subject-matter people possibility to generate the Blaise instrument (survey specific data entry-editing application) – not by writing the code, but simply entering the parameters, describing the survey questionnaire and its data fields into the CAPI questionnaire. From such a collected database (which is in fact a repository of basic metadata of all surveys in the system), the survey data entry applications can be generated. Used tools: Blaise, Manipula and VB.

11. This is the basic principle of Gentry, an in-house developed tool for high-speed data entry from paper forms. It generates ready-to-use applications for data entry (including the double-keying). Of course, difficult and complex datamodels can not be generated: Gentry datamodel specification is based on some common in-house standards for paper questionnaires. In the system only basic process metadata are used (description of fields, blocks, etc.), but the basic needs of the process are fulfilled. The system has been used in production for about six years and successfully covers all SORS data entry from paper forms. Using standardized and straightforward solutions it was also a good example for later developed system for data editing.

12. The system for data editing was made on the same principles as Gentry and using the same tools. For better connection between the two systems, the generator for data restructuring was developed. (Typical transformation from “form=many records” to “form= one record”, and vice versa). Again, some parameters need to be entered in a CAPI form, and upon them the new meta descriptions are generated, using Manipula (the Blaise subsystem that covers batch processes) and VB interface.

13. The most used options to prepare data for further processing (data export from Blaise) are simple ASCII or ASCIIRELATIONAL (with tables separated) files, both with generated meta descriptions for import into target systems (at SORS mostly SAS and Oracle). To support the large, frequent and changing data flows, the stronger integration of Blaise applications with other applications can be used, using BCP (Blaise Component Pack) which also facilitates access to relational databases through OLE DB. To conclude, generating metadata (although only basic, to support the processing) is one of the key features of data entry and editing systems at SORS. Although implemented on a rather basic level, it contributes a lot to efficiency and transparency of the systems, and probably most important, it enforces standardized approach to application development.

Figure 1: GEntry (Generator of applications for high-speed data entry): a part of Blaise questionnaire for survey datamodel specification. Based on values entered, the high-speed data entry application is generated.

Blaise Data Entry - \\sursobd\studio\Razvoj\VnosBL\def_gen

Forms Answer Navigate Options Help

Tip podatka v polju:

☐ 1. Celo število - vrsta podatka na DOS-2: 5,8,9,12-19
☐ 2. Decimalno število
☐ 3. Alfanumerični niz znakov - vrsta podatka na DOS-2: 3,4,6,10,11
☒ 4. Konstanta - vrsta podatka na DOS-2: 20

	Ime	Vk	Tip	Zacetek	Konec	StDec	Dolz	Vnos	Poseb	PreVnos
polje[1]	RAZ		4	1	4		4			0501
polje[2]	ZAP		1	5	10		6	1	1	
polje[3]	MAT7		3	11	17		7	1	1	
polje[4]	MAT3		3	18	20		3	1	1	
polje[5]	MES		4	21	22		2			12
polje[6]	LET		4	23	24		2			99
polje[7]	VK		3	25	25		1	1	1	
polje[8]	spr	1	3	29	40		12	1		

Old 3/15 Modified by rules Clean Navigate def_gen

III. THE ROLE OF METADATA IN AUTOMATIC EDITING SYSTEM IN SAS

C. Metadata in SAS environment

14. In 2004 the Statistical Office of the Republic of Slovenia started the project of developing and implementing a standard editing and imputation system. The reason was first of all the growing demand for quick and efficient flow of the statistical process and the possibility of cost reduction. This project is part of the statistical infrastructure project, the goal of which is to provide standard statistical tools and methods for supporting the statistical process at SORS.

15. The statistical process is about relating multiple data sources and bringing them together to produce statistical data. Metadata provide the definition across data sources that make this possible. In addition, metadata enable you to trace what moved when, how it was changed, what business rules were applied, and what impact those changes might have. These are critical issues. Failure to place enough emphasis on metadata will result in problems later on; often at great cost.

16. Our system is based on two well-known methods – the Hidiroglou-Berthelot method for detection of outliers (H-B method) and the Fellegi-Holt method for errors localization and data imputation (F-H method). It has been applied in the case of two short-term business surveys: the Monthly Survey on Turnover, New Orders and Value of Stocks in Industry and the Monthly Survey on Wages.

17. This paper handles the automatic data editing process and its metadata within the statistical infrastructure, and its relationships with the other statistical processes. It considers its contribution to the management and evaluation of the processes themselves, and to the measurement of the quality of the statistical data. It was concluded that two kinds of process metadata are needed:

- a. those relating to its progress through the statistical process - process metadata and
- b. those relating to the options and parameters which need to be applied to the automatic editing system - methodological metadata

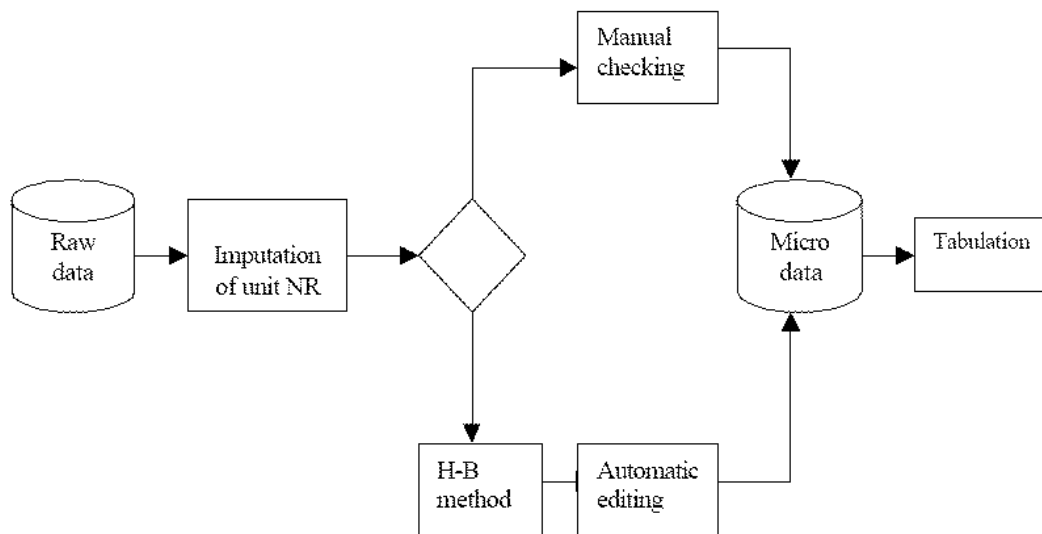
D. Process metadata

18. There should be no need for any human intervention in deciding which process should be applied to the data when, or in what order. This therefore needs to be determined by metadata accompanying the data, and interpretable by the data management systems and the statistical tools. This tool must be able to: recognize the data which are due to be subjected to editing and/or imputation; recognize which editing method should be applied, and with what parameters. In our case metadata are used for the presentation of input matrices of the edit coefficients and the matrix on the correlates. These two matrices, together with the data, are the only input for the whole system. We are planning to set the computer program for derivation of implied edits out of the basic set of edits. For the case of the two discussed surveys this work has still been done partly manually. The output metadata are quality indicators which indicate the number of values that have been changed during the process as well as the impact of these changes on the final results. Through the values of variable indicators it is also possible to see in which stage of the process the data value was changed. The quality indicators are classified into three levels. The first level denotes the data source, the second level denotes if the data were changed and the third level denotes how they were changed.

E. Conceptual metadata

19. The process of exchanging data from Blaise to SAS or Oracle is based on metadata as described in the section II. On the other hand, SAS and Oracle tables have common attributes as the names, types and lengths of variables, which simplify the data and metadata flow between SAS and Oracle.

Figure 2: Statistical process in the case of Monthly Survey on Turnover, New Orders and Value of Stocks in Industry. Data items imputed or changed in each phase are marked with proper process indicator (see annex).



IV. METADATA CONNECTED WITH THE DATA IN ORACLE DATA WAREHOUSE

F. Datawarehousing at SORS

20. At SORS we've started to build a common statistical data warehouse with the Foreign Trade database, implemented in Oracle RDBMS. During several projects between 1999 and 2005 additional databases for final statistical microdata were built in SORS, all of them based on dimensional data model and Oracle technology. Users have access with different analytical tools, for example Oracle Discoverer, Microsoft Access and SAS Enterprise Guide.

21. In the dimensional model data are structured in a form suitable for on-line analysis and tabulation. As long as the data warehouse was used only as a supplement to traditional production systems (situated on mainframe) the basic metadata (mainly classifications) were good enough. However, with the need for corporate data management solution in LAN we have started to add additional metadata into the star-schema model.

G. Metadata-rich production warehouse

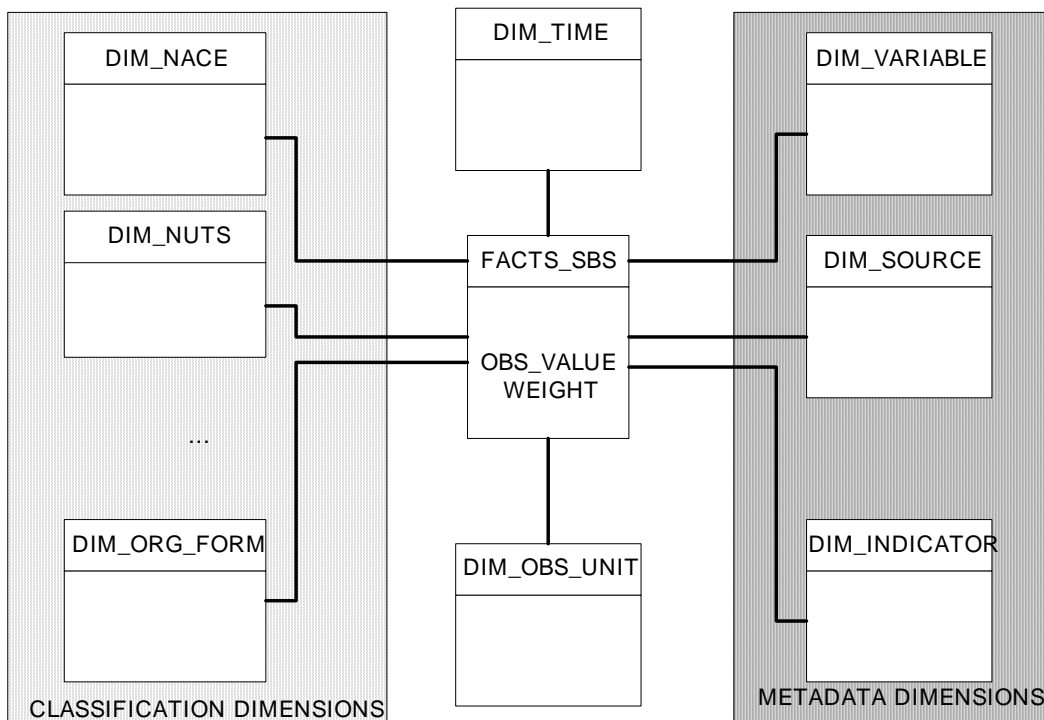
22. In 2004 with the small pilot covering five surveys in the field of waste statistics the new metadata rich dimensional model was introduced. The idea was based on similar data models in NASS and ABS (Yost, Cotter, 2003). The next project (started in June 2005) was modernisation of structural business statistics. The database, built in one of the earliest data warehousing projects as the analytical database for final statistical data, was upgraded in the productional data warehouse containing metadata linked with data.

23. The main requirements included a metadata rich model with classifications (for on-line tabulating), process indicators (describing whether observation was reported, derived, edited or imputed; indicating the type of data source and method of collection/derivation or imputation), variable definitions (with variable code, short name, description, unit of measure, etc.) and data source indicator (with the possibility that some particular continuous variable can be loaded from more than one source, depending on period or part of population). Thin fact table design was selected since in structural business statistics the described metadata elements are observed on the level of a single observation value. In the central fact table there are only the

columns for weighted and unweighted observation value and foreign keys for dimension tables populated with metadata.

24. Such a model with metadata directly linked with data in a structure suitable for analytical tools offers powerful analytical possibilities for users (Lubulwa, 2003), better possibilities for using metadata in data processing modules (for example SAS) and better control over the statistical process.

Figure 3: Conceptual data model of metadata-rich structural business statistics star schema



V. LESSONS LEARNT

25. Metadata usage in productional tools has shown many opportunities but also a lot of new challenges.

H. The role of central repositories for metadata

26. Central classification server (Klasje) and central metadata repository (Metis) are natural places to store metadata used by production tools. It is much easier to build the production process architecture if there are centrally managed metadata stores. However, using metadata stored in a central repository in production tools means that all metadata in the repository have to be exact, complete and collected and managed in a very consistent way. In the opposite case the production process will not run smoothly or will even break. The needed level of quality can be achieved using corporate business rules, good user interfaces and permanent control. In SORS classification server Klasje is used as a primary source for production tools and there are activities to implement the similar process for other types of metadata (variables, questionnaires etc.).

27. Metadata suitable for usage in production tools of course need to be in a structured form, not in plain text. That is one of the reasons that central repositories should be used instead of template-based documentation. However, not all types of metadata are feasible to put into a metadata repository. For process metadata observed on observation value level, the most suitable place to be stored is together with data. For some other types of metadata (variable definitions, classifications, etc.) it is very convenient to store it in a data

warehouse with direct connection with the data, but the place from where it should be loaded into a data warehouse are central repositories.

I. Harmonisation of metadata concepts

28. In order to use them in the production process, metadata concepts should be harmonized. The usual problem is that people are using metadata already but they are using local versions (for example own variable names, etc.). With the introduction of central metadata management there is a tension to keep “legacy” metadata in the same form without harmonizing it. For efficient metadata – based production process also the cultural change is needed.

J. Technical considerations

29. From technical point of view the possibilities for metadata exchange and system integration are better and better. For example from SAS environment it is possible to generate a database in Oracle and populate it with data without the need to write single line of SQL code. The possibilities for metadata exchange are one of the main drivers for using the commercial tools instead of developing custom applications.

VI. CONCLUSION

30. At SORS we try to follow the golden metadata rules, where possible. In Blaise we are using conceptual metadata to define the data structure and generate data-entry applications. In SAS metadata are used as an input defining the rules for automated editing and process metadata are produced automatically. Both conceptual and process metadata are loaded in the Oracle metadata-rich datawarehouse environment to be accessible for statisticians.

REFERENCES:

Pierzschala, M., and Manners, T., Revolutionary Paradigms of Blaise, Proceedings of the 7th Blaise User's Conference, Washington D.C., 2001

http://www.blaiseusers.org/ibucpdfs/2001/Pierzchala_Manners_IBUC_Paper.pdf

R. Seljak, T. Špeh: Automatic Editing System for Two Short-Term Business Surveys,

<http://www.unece.org/stats/documents/2005/05/sde/wp.43.e.pdf>

P. Tate: The Role of Metadata in the Evaluation of the Data Editing Process,

<http://www.unece.org/stats/documents/2005/05/sde/crp.5.e.pdf>

M. Yost, J. Cotter: The Impact Of A Dimensional Data Warehouse On Survey Processing Systems, Statistical Input Data Warehouse Workshop, Canberra-Murramarang, 2003

G. Lubulwa: Towards a better IDW: What Every Analyst Wants, Statistical Input Data Warehouse Workshop, Canberra-Murramarang, 2003

APPENDIX: CLASSIFICATION SCHEME OF THE PROCESS INDICATOR

The process indicator is a 3-level classification (xy.zz). The first level (x) describes whether data was reported directly from reporting unit or in different way. The second level (y) shows status of the data after the statistical processing and offers information which data was changed. The third level (zz) is reserved for the information about the method, used for statistical processing. Classification is stored in classification server and used by production tools to automatically set the proper status. Data, stored together with »indicator dimension« in data warehouse offer the complete information about processing of each data value.

1st level (x)

Mode of data collection

- 1 data provided directly by reporting unit
- 2 data from administrative source
- 3 data computed from original values
- 4 imputed data – imputation of non-response
- 5 imputed data – imputation due to invalid values detected through the editing process
- 6 data missing because the unit is not eligible for the item (logical skip)

2nd level (y)

Data status

- 1 original value
- 2 corrected value

3rd level (zz)

Method of data correction

- 11 correction after telephone contact
- 12 data reported at a later stage

Reporting methods

- 11 reporting by mail questionnaire
- 12 computer assisted telephone interview(CATI)
- 13 telephone interview without computer assistance
- 14 paper assisted personal interview (PAPI)
- 15 computer assisted personal interview (CAPI)
- 16 paper assisted self interviewing
- 17 computer assisted self interviewing
- 18 web reporting

Imputation methods

- 10 method of zero values
- 11 logical imputation
- 12 historical data imputation
- 13 mean values imputation
- 14 nearest neighbour imputation
- 15 hot-deck imputation
- 16 cold-deck imputation
- 17 regression imputation
- 18 method of the most frequent value
- 19 estimation of annual value based on infraannual data
- 21 stochastic hot-deck (random donor)
- 22 regression imputation with random residuals
- 23 multiple imputation

Method of computation of derived values

- 11 production value calculated out of the deflated turnover and change of the stocks
- 12 production value calculated out of the deflated turnover
- 13 value calculated out of the given proportion
- 14 turnover calculated out of the tax authorities data

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 (iii): Metadata and the Statistical Cycle

RE-ENGINEERING PROJECTS FOCUSSING ON METADATA AND THE STATISTICAL CYCLE

Invited Paper

Submitted by Statistics South Africa, South Africa¹

I. PURPOSE OF THE PAPER

1. This paper briefly introduces the current work on metadata management being undertaken in Statistics South Africa (Stats SA). It provides the reader with an understanding of the anticipated scope, implementation issues and identified benefits of metadata management through the entire statistical value chain. It also presents the work done thus far in developing the requirements for implementing a metadata store as part of the End-to-end Statistical Data Management Facility (ESDMF).

II. BACKGROUND

2. Stats SA, as South Africa's official statistics agency, is responsible for informing users of the concepts, classifications and methodologies used in collecting, processing and analysing data, the quality of that data, and all other relevant features describing the data.

3. The vision of Stats SA is to be *the preferred supplier of quality statistics*. The achievement of this will promote a culture of evidence-based planning and decision-making in pursuit of socio-economic development and good governance. The development and application of world-class standards, classifications, methods and procedures are central to the drive for quality statistics.

¹ Prepared by T.J.Lukhwareni, josephl@statssa.gov.za, A.C. Jenneker, ashwellj@statssa.gov.za and E. Gavin lizg@statssa.gov.za

4. Stats SA has long recognised that the centralised storage of data is a key component for improving the quality of data. The prospect of developing a data warehouse has been debated since at least 1997. Data warehousing options were evaluated in consultation with other statistical agencies. In 2004, Stats SA adopted the business case for the development of such a facility. The requirement for the data warehousing infrastructure was to consolidate the management and processing of data and metadata. Two key deliverables were identified:

- a central data storage facility in which statistical data is maintained in a standard manner, together with a set of tools for retrieval, analysis and report-generating;
- a central metadata store where data and information needed to interpret the data is stored according to standard, uniform and agreed fields and formats. This would include a store for classifications, concordances, code files and sample frames.

5. A business requirements gathering exercise indicated that, while a data warehouse may be a necessary element in addressing quality issues throughout the statistical value chain, it is not sufficient to solve these issues. Proceeding from this premise, a consultant retained by Stats SA subsequently proposed the high-level conceptual scope shown in figure 1. Statistical production units within Stats SA also gave input on the high-level conceptual scope.

6. The End-to-end Statistical Data Management Facility (ESDMF) proposed consists of a number of functionally integrated but modular conceptual components that will address specific aspects of the quality improvement through the statistical value chain. The high-level scope diagram outlines six main components in the ESDMF: the Collections Environment, Central Data Store (CDS), Dissemination Area, Workflow Control, Metadata Store and Access Control. (Respondent and client management are out of the scope of this facility).

7. The aim of the Workflow component is to control the definition and execution of formalised processes in a consistent, repeatable and standardised manner that can support delivery of improved statistical products. The processes defined for execution in the Workflow component will support the complete statistical value chain.

8. This component will act as the link between the actions performed within a specific project and the metadata stored in the Metadata Store. Metadata and other data such as control measures will be used to validate completed tasks, and to govern tasks to be performed. The Workflow component will record, in an audit trail, all actions performed on the data from collection to dissemination.

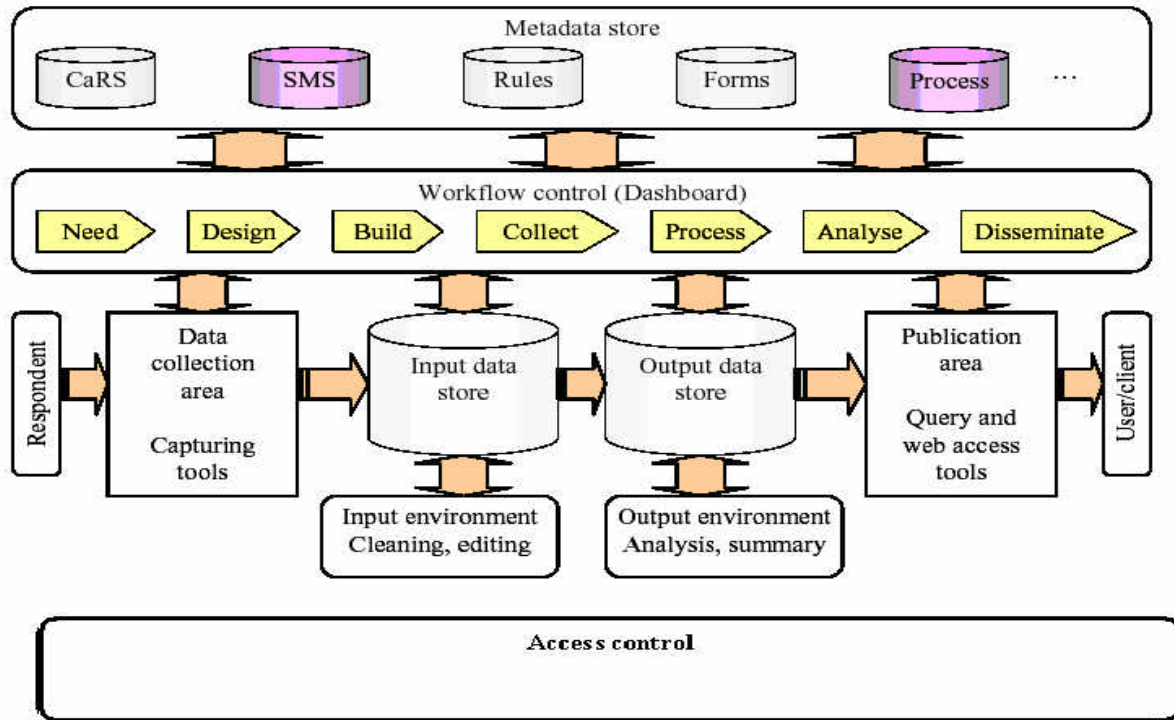


Figure1: High level conception of the End-to-end Statistical Data Management Facility (ESDMF)

CaRS: Classifications and Related Standards

SMS: Survey Management System

III. METADATA MANAGEMENT

9. “Metadata management refers to the content, structure, and designs necessary to manage the vocabulary and other metadata that describes statistical data, designs and processes. ... (It) includes the development of metadata models, building metadata registries to organise the metadata, development statistical terminologies which define and organise terms” (Bargmeyer and Gillman, METIS 2000).

10. The management of metadata includes the following elements:

- understanding what metadata is;
- understanding the scope of metadata;
- establishing metadata owners, maintainers and regulators;
- ensuring that data users have easy access to relevant data and metadata in a timely manner;
- determining appropriate technical storage environments i.e. the metadata store;
- registration of metadata items.

A. The definition of metadata

11. Strictly speaking, there is no such thing as a perfect or absolutely accurate dataset. One of the functions of metadata is to document data structure, assumptions and limitations so that data are not used inappropriately, and so that users can easily understand limitations on use. A standard format for metadata is required to facilitate appropriate use and understanding of the data. This can be achieved by producing metadata that fits the content of data and the needs of authors, information providers and end-users.

12. Stats SA adopted the definition of metadata proposed by Statistics New Zealand (SNZ) [5]. Metadata is “data about data, and refers to the definitions, descriptions of procedures, system parameters, and operational results which characterise and summarises statistical programs. Metadata may be passive (descriptive), i.e. the

form of documentation, which is used by agency staff, or may be active (prescriptive), i.e. determining the actions of automated surveys processes”.

13. Metadata can be further categorised into various categories. Stats SA distinguishes five separate categorisations:

- Definitional metadata: metadata that describes the concepts in the data, e.g. classifications, question and question modules for collection instruments and statistical terms;
- Operational metadata: metadata arising from and summarising the results of implementing the procedures, e.g. response rates, edit failure rates, costs and other quality and performance indicators;
- System metadata: active metadata used to drive automated operations, e.g. publication or dataset identifiers, date of last update, file size, access methods to databases and mapping between logical names and physical names of files;
- Procedural/methodological metadata: metadata relating to the procedures by which data are collected and processed, e.g. sampling, collection methods and editing processes;
- Dataset metadata: metadata used to describe, access and update the dataset and data structures, e.g. textual description, data cell annotations, dataset title, keyword and distribution information.

14. These metadata categories may be interrelated. In particular, the metadata used to describe fully a statistical dataset could include elements of the first four categories listed above. Such information could be necessary for the appropriate use of the dataset. An analogy is that of the evolution of the documentation of geographical information, to include the explicit recording of processing steps (the *lineage* of the data) and quality indicators.

15. Metadata usually includes information about the metadata itself. This describes important features about metadata documentation such as its author, technical form, publisher, etc. This kind of information may have predefined descriptions of the information to be included and how this information is to be represented. In this context, metadata can be described as “metadata about metadata”. Metadata about metadata is needed, for example, to facilitate the author’s work by documenting important details, and to keep a record of important facts about the metadata updates and related issues.

B. The Metadata Store

16. Stats SA’s Metadata Store will manage the definition and updating of metadata needed to maintain standards and to record and assist in controlling the processes throughout the statistical value chain. It must be centrally available and provide a single source of approved and reusable metadata for metadata creators and users. The Metadata Store must allow new metadata to be defined to support growth in this central repository of metadata. It will include systems such as the Classifications and Related Standards (CaRS), acquired from SNZ, which provides a central repository for classifications, concordances, concepts and definitions, and codefiles for use by all components within the organisation.

17. Metadata occupies an important role in every phase of the statistical value chain and, if developed appropriately, can enhance the quality of data and the statistical production process. Stats SA’s proposed Metadata Store serves several important functions for development and storage of systematic and coherent information and data management. Besides its annotative purpose, the metadata store has an important role in maintaining and improving data quality through the following functions:

- Providing a centralised store to enable and encourage collaborative sharing and re-use of definitions and standards throughout the organisation;
- Ensuring a consistent interpretation of standards and definitions across all aspects of surveys and over time;
- Improving data quality by enforcing compliance to workflow controls and procedures;
- Minimising human error by regulating and automating downstream statistical processes.

C. Metadata in the statistical value chain

18. Stats SA is in the process of developing a metadata entity map that highlights the grouping of metadata elements within the statistical value chain (see Annex 1). This map also provides a link to the sources of metadata and a brief description of each element.

19. The statistical value chain can be defined by the following phases:

- Need: the need phase is entered when a new project is defined. In this phase the request is evaluated, the feasibility for a new project is investigated, the project is planned and a budget is developed.
- Design: in the design phase the project is planned in detail, the survey methodology is developed, the questionnaire is designed, the sample defined, and the data capturing tool is designed.
- Build: in the build phase the data capturing tools are developed, tested and implemented.
- Collect: in the collect phase data is collected and captured using the data capturing tools.
- Process: in the processing phase the captured data is processed. This phase includes validation, correcting, editing and imputation. The result of this phase is a clean dataset.
- Analyse: in the analysis phase all analysis on the data is performed. The result of this phase is a publishable dataset.
- Dissemination: In the dissemination phase publications are created from the dataset produced by the analysis phase, and disseminated in various forms.

D. Metadata Standard

20. A metadata standard outlines the characteristic properties to be recorded, as well as the values the properties could or should have. Standardisation of metadata documentation facilitates wider information sharing and usage. The use of metadata standards enables producers to describe datasets fully and coherently. The adoption of a standard also facilitates data discovery, retrieval and use. If a standard is used, finding a specific piece of information in a metadata record is far easier than if no standard is used. Standards also enable automated search and retrieval functionality.

21. Developing or adopting a metadata standard is a complex process often involving different and not always complementary interest groups. There is often significant overlap – and possibly even inconsistency - as one standard is extended into the territory of another. A number of bodies with which Stats SA interacts, both international and local, are involved in generating standards. These include the International Standards Organisation (ISO) the International Monetary Fund (IMF) and the national standard setting body, namely the South African Bureau of Standards (SABS).

22. Adoption or adaptation of an existing metadata standard holds the dual advantages of minimising development effort and ensuring a common understanding of the metadata by the existing community of users of that particular standard. A further consideration for Stats SA is to limit the burden of compliance by alignment to existing metadata standards as far as is possible. For example, many of the datasets disseminated by Stats SA have an explicit geographic component. In terms of national legislation, metadata on these datasets must be made available in accordance with a prescribed standard for documenting all geographic information datasets developed using public funding.

23. Stats SA is involving the data production components from economic, social and population statistics components as well as supporting functions, such as the ICT and geographic divisions, in the development of metadata standards to meet the organisations needs without overburdening the originating components.

24. Stats SA's Data Management and Information Delivery (DMID) project is evaluating the following metadata standards and systems:

- **ISO/IEC 11179**
ISO/IEC 11179 is an international standard for the specification, and standardisation and management of data elements through metadata registries. ISO/IEC 11179 gives the description of a registry of metadata describing any data.

Part 4 of this standard provides guidance on how to develop unambiguous data definitions. A precise, well-formed definition is one of the most critical requirements for shared understanding of an administered item. Well-formulated definitions are imperative for the exchange of information.

Part 6 of the standard provides instruction on how a registration applicant may register a data item with a central registration authority, and how to allocate unique identifiers for each data item. Maintenance of administered items already registered is also specified in this part of the standard.
- **SANS 1878**
The South African Spatial Metadata Standard SANS 1878 is a profile of the international standard ISO 19115: Geographic Information/Geomatics - Metadata. This standard defines almost 300 metadata elements, with most being listed as 'optional'. The metadata within the standard is an aggregate of the following elements: identification, constraints, data quality, maintenance information, spatial representation, reference system, content information, portrayal catalogue reference, distribution, metadata extension information and application schema information.
- **SCBDOK**
Statistics Sweden's metadata system consists of a number of tools and templates. SCBDOK [10] is a documentation model that specifies the basis for the different statistical production methods. The SCBDOK template is the cornerstone of Stats Sweden's metadata system. Metadok, a software tool, provides a system for creating formalised metadata for the purposes of describing final observation registers in SCBDOK.

E. Principles and benefits of metadata management

25. Benefits of investment in a metadata management programme and associated metadata store include increased re-use of corporate assets, improved impact analysis associated with these assets, increased quality for decision-making, reduced development and maintenance time, greater success in deployment of new organisation capabilities, improved user access and usage, and better understanding of corporate assets. Metadata serves as the link between business processes and data, applications and the technical infrastructure.
26. The principles guiding Stats SA's design of a metadata store include the standardisation of metadata across the organisation in so far as possible, the reuse of metadata and the alignment of metadata standards with existing standards used by key stakeholders where possible. The aim is to limit the compliance burden in order to engender organisational support and rapid uptake of new metadata standards.

IV. RECOMMENDATION AND CONCLUSIONS

27. Management of metadata is not just a technical issue for the data producers. It provides end-users of statistical products and data with adequate information for proper use. Conforming to a metadata standard is important to ensure that users can find, understand and share data.
28. Continuous updating at every stage of the statistical value chain is necessary for an efficient metadata management system. It is not enough to update and capture metadata once; it has to be checked and, if necessary, modified at every stage of the statistics value chain. Technical solutions ensuring that metadata can easily be registered and updated together with the data, at one place, is a precondition for an efficient metadata system. This is a great challenge for statistical agencies working with disparate metadata systems.

29. Management of metadata at Stats SA has moved from being an afterthought, to becoming the core of the improvement of quality of the data and production process.

V. REFERENCES

- [1] Gillman, D. (2004) *Metada Registries, Data Harmonization and Maximizing Use of Warehouse*, US Bureau of Labor Statistics.
- [2] ISO/ IEC Standard 11179 (1997). *Specification and standardization of Data elements. International Standard Organization*, Geneva, Switzerland. ISO/IEC 11179 standards.
- [3] Jenneker, A. (2004), *The Business Case for a Data Warehouse in Statistics in South Africa*, Draft version 0.04, Statistics South Africa , Internal electronic document: DMID_DW_BC_V0-04.doc.
- [4] Johanis, P ‘et al’ (2003). *Paper for the Open Forum 2003 on Metadata Registries* 20-24 January 2003, Santa Fe, New Mexico, USA.
- [5] Merrington, R (2004) *Creating a New Business Model for a National Statistical Office of the 21st Century*, Statistics New Zealand.
- [6] Oakley, G. (2004) *Report of visit to Statistics South Africa to advice about the Data Warehousing Project* 31 May to 4 June 2004.
- [7] Oakley, G. (2004) *Revised Paper from Nov 03 IRMC discussion of ‘Strategy for End-to-End Management of ABS Metadat’*.
- [8] Lukhwareni, T.J. ‘et al’ (2005). *Management of Metadata in National Statistical Agency*. Commonwealth conference paper.
- [9] Statistics Norway (1998). *Guidelines for statistical metadata on the Internet. Statistical Journal of the United Nations ECE* 15 (1998) 169-176
- [10] Sundgren, B. (2000) The Swedish Statistical metadata system. Eurostat conference, 2000.
- [11] Lukhwareni, T.J. ‘et al’ (2006). *Data Quality Policy, Statistics South Africa*.
- [12] Madonsela, S.F. ‘et al’ (2006). *Metadata Entity Map, Statistics South Africa*, draft.
- [13] Lindblom, H. and Sundgren, B. (2004) *The metadata system at Statistics Sweden in an international perspective*.

ANNEX 1: METADATA WITHIN THE STATISTICAL VALUE CHAIN

<p>1. Need</p> <p>1.1 Determine information requirement</p> <ul style="list-style-type: none"> • Evaluate the request • Investigate the feasibility of the project • Identify stakeholders • Consult stakeholders • Acquire feedback from key stakeholders • Create proposal and approval, etc <p>1.2 Develop detailed project plan</p> <ul style="list-style-type: none"> • Create detailed task plan • Acquire feedback from key stakeholders • Revise plan • Approve plan <p>1.3 Develop Budget Plan</p> <ul style="list-style-type: none"> • Prepare initial budget and plan • Outline and quantify costs and benefits • Obtain Approval 	<p>2. Design</p> <p>2.1 Develop survey methodology</p> <ul style="list-style-type: none"> • Determine detailed objectives • Determine survey options • Perform design analysis • Produce design recommendations • Sign off survey methodology • Produce survey methodology specification <p>2.2 Design operational requirements</p> <ul style="list-style-type: none"> • Produce collection requirements • Produce processing requirements • Produce output requirements <p>2.3 Design and testing questionnaire</p> <ul style="list-style-type: none"> • Develop and test questionnaire • Deliver questionnaire recommendation <p>2.4 Sampling</p> <ul style="list-style-type: none"> • Perform reselection • Perform sample maintenance • Confirm sample
<p>3. Build</p> <p>3.1 Printing questionnaire</p> <p>3.2 Build technology solution</p> <ul style="list-style-type: none"> • Analyse or refine technical specifications • Create system solutions • Detailed module design • Code application <p>3.3 Test technology solution</p> <ul style="list-style-type: none"> • Programmer testing • Peer review • Client acceptance testing • Client sign off <p>3.4 Implement technology solution-training and piloting</p> <ul style="list-style-type: none"> • Controlled release • Support and training for users • Documentation 	<p>4. Collect</p> <p>4.1 Field work</p> <ul style="list-style-type: none"> • Receive questionnaires • Maintain contacts details • Receive administrative data • Follow ups <p>4.2 Manage respondents</p> <ul style="list-style-type: none"> • Review sample • Pre-contact respondents <p>4.3 Close off collection</p> <ul style="list-style-type: none"> • Monitor response rate • Sign off collection
<p>5. Process</p> <p>5.1 Capture data into electronic form</p> <ul style="list-style-type: none"> • Batch questionnaire • Input data • Validate data <p>5.2 Perform macro editing</p> <ul style="list-style-type: none"> • Compare data with data from 	<p>6. Analysis</p> <p>6.1 Examine source data</p> <ul style="list-style-type: none"> • Ensure data structures are correct • Assess whether observed changes reflect expectation • Compare data with data from previous periods • Compare data with other data source

<p>previous periods</p> <ul style="list-style-type: none"> • Compare data with other data sources • Investigate outliers <p>5.3 Run imputations or estimations</p> <ul style="list-style-type: none"> • Tax modelling • Identify unlinked and special treatment • Run imputations • Monitor imputations <p>5.4 Produce clean datasets</p> <ul style="list-style-type: none"> • Prepare handover report • Peer review • Release for analysis 	<ul style="list-style-type: none"> • Investigate outliers <p>6.2 Produce statistical results</p> <ul style="list-style-type: none"> • Create estimates for total survey population • Produce derived data • Apply seasonal adjustments • Apply experts knowledge adjustments • Customise analytic applications(where required) • Derive sampling errors <p>6.3 Validate results</p> <ul style="list-style-type: none"> • Assess whether estimates reflect expectations • Compare derived data with derived data from previous period • Compare derived data with other source data • Assess quality measures • Check output have been calculated correctly <p>6.4 Interpret results</p> <ul style="list-style-type: none"> • Determine explanation <p>6.5 Prepare content for dissemination</p> <ul style="list-style-type: none"> • Confidentialise data • Indicate quality of data • Produce content • Approve content for release <p>6.6 Conduct quality control</p> <ul style="list-style-type: none"> • Document main findings for process improvement • Review Methodology • Get client feedback
<p>7. Dissemination</p> <p>7.1 Receive and validate draft content</p> <ul style="list-style-type: none"> • Receive dataset/content • Perform quality assurance of data content • Perform editorial changes • Editorial and content sign off <p>7.2 Manage and load dissemination repositories</p> <ul style="list-style-type: none"> • Load repositories <p>7.3 Prepare pre-release for publishing</p> <ul style="list-style-type: none"> • Prepare and populate tables • Apply corporate formatting standards 	

<ul style="list-style-type: none">• Prepare electronic distribution• Prepare hard copy outputs <p>7.4 Manage first release</p> <ul style="list-style-type: none">• Manage releases <p>7.5 Handle customer enquiries</p> <ul style="list-style-type: none">• Receive enquiry• Categories enquiry• Respond• Customer follow up	
--	--

**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 (iii): Metadata and the Statistical Cycle

**DISSEMINATION OF STATISTICAL DATA AND METADATA - PROCESS BASED ON COMMON
STRUCTURE OF STATISTICAL INFORMATION (COSSI)**

Supporting Paper

Submitted by Statistics Finland¹

I. INTRODUCTION

1. Statistics Finland has been developing XML-based data dissemination for a couple of years now. The dissemination system is based on the model of common structure of statistical information (CoSSI), and the dissemination is based on XML documents compatible with it. The CoSSI model covers different ways of statistical data organisation (statistical data matrix and statistical table), statistical publications (monthly and quarterly publications, press releases, etc.) and quality declarations. The structuring of the metadata connected to statistical data is also implemented within this system.

2. The metadata part in the CoSSI model is divided into document metadata, statistical metadata and processing metadata. Document metadata is information about the producer of the document, the document's content, date, statistical topic, etc. Statistical metadata is information vital for the interpretation of numerical statistical information, and describe the variables in a statistical table or matrix. This metadata information is useful for the user in the dissemination process by helping the interpretation of statistical figures, and for the producers of statistics when metadata are transferred between statistical production stages. It could be also used for bilateral exchange of statistical information between statistical agencies.

3. The metadata, publications, tables and matrixes will be stored in an XML database. This database is also an archive for all published information. Statistical publications are prepared in the statistics departments of Statistics Finland. People responsible for publications will have access to the XML database via an XML editor. Using this XML editor they can make publications, and include tables, figures and metadata in them, and also update and write new metadata documents describing their statistics. So a publication in XML format, defined by the CoSSI model, includes all the statistical data, metadata, text, figures and language versions in a single XML file. This XML file will then be automatically converted to HTML files for the website, PDF files for printing or to some other appropriate format.

¹ Prepared by Harri Lehtinen (harri.lehtinen@stat.fi).

4. PC-Axis is also starting to use the CoSSI model as XML format. For this the CoSSI model contains a special processing metadata section called pxmeta. This PC-Axis metadata information can also be stored in XML format in the same XML database as all the other information. By doing this, database (PX-Web) publishing can also be incorporated into the same process as publication production.
5. Using XML and XML tools, such as XML database and XML editor, we can unify the production of information for different publication channels. This also gives us the opportunity to include rich metadata in statistical information and publish metadata along with the statistics they describe.

II. COSSI - COMMON STRUCTURE OF STATISTICAL INFORMATION

6. At Statistics Finland we have been developing a common structural definition of statistical information (CoSSI ²), which covers different ways of statistical data organisation (statistical data matrix and statistical table), publications, and within which the structuring of the metadata connected to statistical data is also implemented.
7. The CoSSI defines the structures of statistical data (matrices and tables), metadata (document and statistical metadata, and quality declarations), and publications. XML DTDs have been selected as the technical means for implementing these structures. The CoSSI model is comprised of several DTDs that can be modularly combined for different types of documents. The basic document types are a statistical table (CALS), a statistical matrix (XDF) and a publication. These documents are XML documents that are compatible with the CoSSI model and also contain the metadata and the language versions necessary for describing a set of statistics.

A. Metadata in CoSSI model

8. In the CoSSI model, metadata are divided into four categories:
 - a. Document metadata
 - b. Statistical metadata
 - c. Quality declarations
 - d. Processing metadata
9. The division is based on the content and character of metadata. Document metadata describe the content of a document, its creator and identifiers connected with the document. Statistical metadata contain descriptions of the variables that are present in statistical data and tables, calculation rules and any classifications that may apply to a variable. The quality declaration is a standard format description of the data collection, the data and the applied statistical methods. Processing metadata are metadata for statistical software applications.

B. Document metadata

10. The document metadata module has all the elements of DublinCore, as well as some metadata specific to Statistics Finland. The document metadata DTD cover the metadata that are needed for the publications, tables and matrices of Statistics Finland (electronic and paper dissemination).
11. Document metadata provide information about the person and organisation having produced the document, the content of the document (e.g. subject, keywords, language, source) and information specific to Statistics Finland, such as the series and category of Official Statistics of Finland. Productional metadata, processing instructions, etc., are not contained in document metadata.

² More detailed description is presented in Rouhuvirta and Lehtinen, Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003.

12. Document metadata is a module that is used in tables, matrices, statistical metadata, and publications for describing document metadata. The document metadata module can also be used for the production of information for bibliographical information systems.

Document metadata	
Creator	Person
Subject	
Keywords	
Content description	
Publisher	Organisation
Contributor	Person
Date	Published modified
Type	
Format	
Language	Main and other language
Document information	OSF and Category
Identifier	URN, URL, ISBN, ISSN, DOI, Number
Rights	
Coverage	
Relations	
Source	

Figure 1. Content model of document metadata

C. Statistical metadata

13. Statistical metadata are data (matrix or table) specific. A statistical metadata document describes the variables, and their operational definitions and classifications in matrices and tables. If data (matrix or table) are changed as a result of tabulation or other procedure, these changes are added to the statistical metadata document. Thus, if a variable is edited, the pertinent metadata are added to the statistical metadata.

14. Statistical metadata are largely presumed to be textual. This is also the case when statistical metadata are presented with conceptual symbols as a formula. Our cultural environment requires metadata to be multilingual. Within the CoSSI model this has been solved as part of the structuring of statistical information.

15. The description of statistical metadata does not contain information about the processes that guide the production of statistics or about the monitoring of this process, or about the technical descriptions of data that are required by the diverse software programs used in the processing of statistical data. The document identification and other metadata required in archiving are described in another component of the CoSSI model that covers metadata concerning documents and file copies.

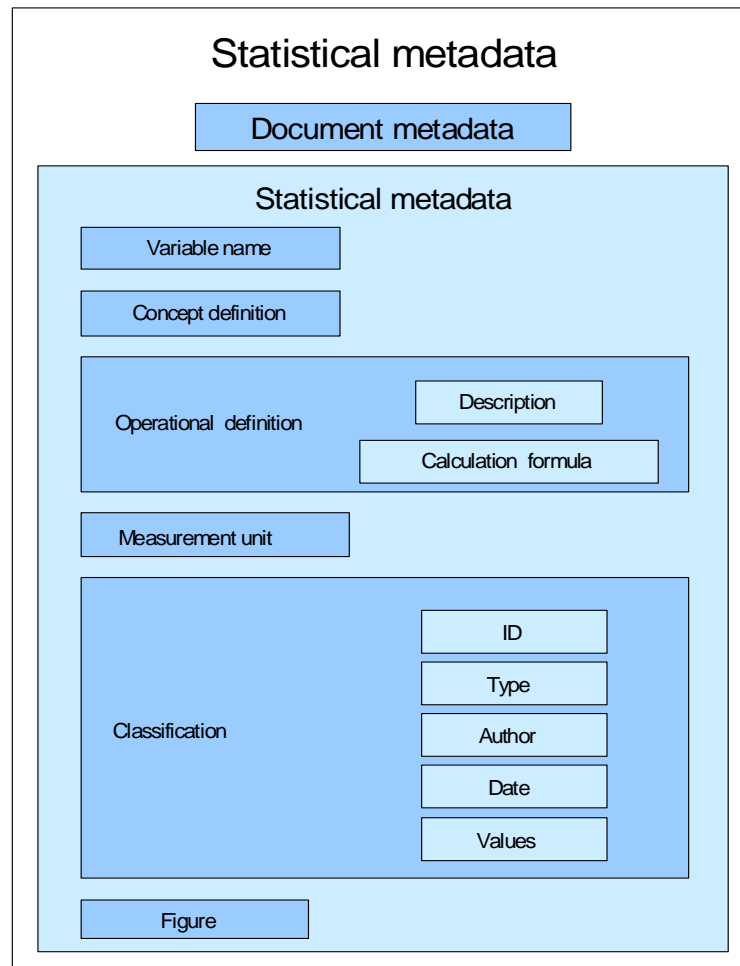


Figure 2. Content model of statistical metadata

16. The model of statistical metadata is very simple but rich in information content. Its main benefit to the production of statistics is that statistical metadata retain a consistent structure right from a survey questionnaire to the eventually disseminated statistics. Because of their simple structure, statistical metadata can be easily exploited during different stages of statistics production and the same model of statistical metadata can be used in all stages from the collection of data to their publication and dissemination. As the same model is used, it does not impose any necessity to re-enter, convert or otherwise re-process statistical metadata, which remain unchanged and go together with the numeric data they describe from the start to the finish of the data processing procedure.

17. The second major benefit is the extendibility of the data, for instance, in situations where statistical standard classifications are used. The model of statistical metadata provides a set and standardised ways of making standard classifications part of them.

D. Quality declaration

18. Quality declaration, or quality metadata, describe the following aspects of statistics:

- a. Relevance of statistical data
- b. Description of the methods used in statistical surveys
- c. Accuracy of information
- d. Timeliness and promptness of the information published
- e. Accessibility and clarity of information
- f. Comparability of statistics
- g. Consistency

19. In publications, tables and matrices, quality metadata can be provided as a quality declaration, which is a module in the CoSSI model.

E. Processing metadata

20. Processing metadata are metadata intended to be used by a certain software application to guide its functioning. Examples of these would be certain PC-Axis keywords that instruct the application to show the figures in a table at the accuracy of certain number of decimal digits or create the heading of a table automatically according to predefined rules. The question so far studied in respect of processing metadata is how data in the PC-Axis file format could be included in an XML format document compatible with the CoSSI model and which PC-Axis keywords should then become processing metadata. In future, it would be possible in principle to expand the CoSSI model with the processing metadata of other statistical applications, such as SAS and SuperStar, depending on the production environment.

III. XML-BASED PUBLISHING SYSTEM

21. Statistics Finland has been developing XML-based publishing system which sets out XML documents, publications, tables, matrices and metadata that are compatible with the CoSSI model. The system converts these XML documents automatically into the formats required by different dissemination channels.

22. An output format compatible with the tables and matrices of the CoSSI model is needed for statistical software applications for XML-based publishing. The main applications Statistics Finland uses are SAS and SuperStar, and PX-Edit for PC-Axis tables. At the moment PC-Axis tables in matrix and table formats can be produced with PX-Edit and output format for matrices and tables has also been developed for SAS. In the newest version of SuperStar there will be an output in the CoSSI table format (CALS).

23. The actual production of publications takes place at the statistical operating units where statistical experts write the text, select the tables for the publication and produce the statistical graphics. Epic software was selected as the editor for XML-based publishing, and was tailored during year 2005 to function as the production editor for publication documents conforming with the CoSSI model. In the tailoring the user interface of the editor was made as user-friendly as possible and functions were added to it that support e.g. importing of external XML tables compliant with the CoSSI model into a publication, production of language versions, completion of metadata and writing of text. Technically, XML is hidden in the editor, so the editing environment is quite similar to that of familiar word processing programs.

24. An XML database into which statistical metadata compatible with the CoSSI model are saved as XML documents has now been taken into test and piloting use. As tables are made, descriptions of the variables selected for them can be retrieved from the database and thus included in the dissemination. There is also a connection into the XML database from the Epic editor, so statistical metadata can also be retrieved via the editor. Statistics Finland chose XML database called eXist to be the database for publications, statistical data and metadata.

25. In consequence, a monthly or quarterly publication written with the Epic publishing editor becomes a document compliant with the CoSSI model and contains all the material of one publication, i.e. text, tables, statistical and document metadata, figures and language versions, in one XML file. These publication originals in XML format are saved in an XML database (eXist), which becomes the publication archive. Published tables and matrices are also saved in XML format into the archive.

26. Before its publishing, a publication in XML format still has to be converted into the format required by the used dissemination channel. For publishing on Statistics Finland's website, an XML publication is converted fully automatically into HTML and PDF formats. The conversion into HTML format produces a set of HTML pages from one XML publication so that the caption text of the publication forms the start page and the contents listed under it form links to other parts of the publication. The conversion produces sets of HTML

pages in all languages that are present in the XML publication. Besides HTML versions, PDF documents are also produced in each language, and these can be offered to customers as printable versions on the HTML pages. The conversion into PDF format produces one PDF document per each language version in the XML publication.

27. The author of the publication can check the final HTML and PDF versions that will be disseminated prior to their publication. When the publication is ready, the author has at his or her disposal a publishing program for defining when the publication should be released and which files should be published.

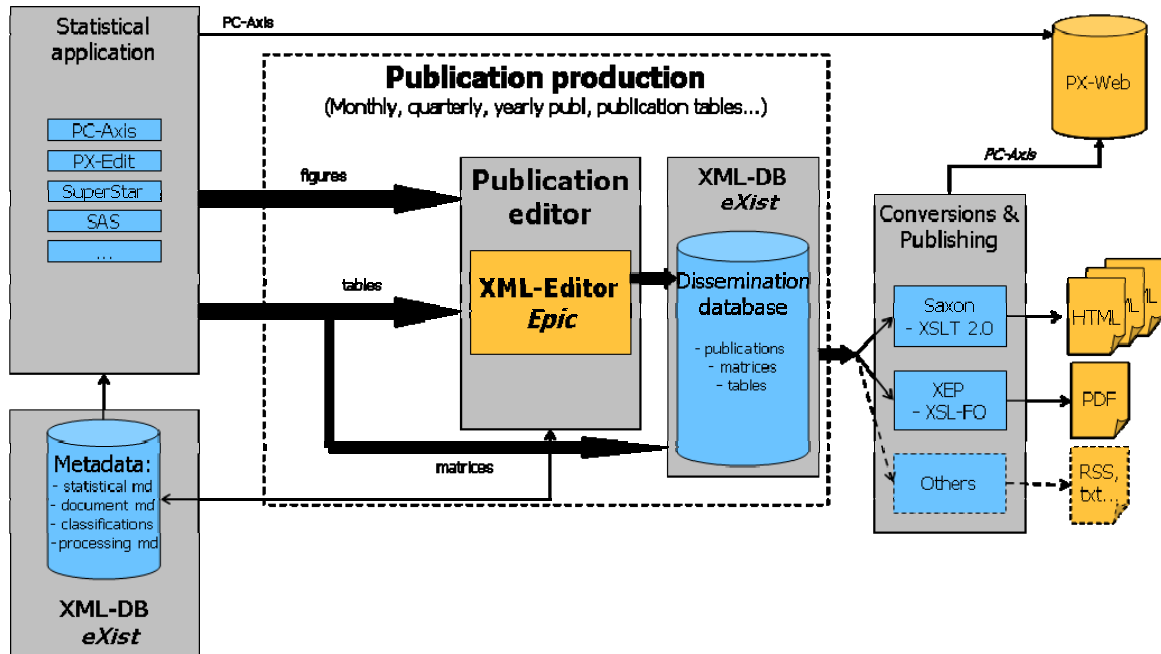


Figure 3. XML based dissemination of statistical data, publications and metadata

IV. EXAMPLES

A. XML database - eXist

28. Open source database called Exist has been chosen as the XML database for statistical publications, statistical data (tables and matrices) and metadata. The database structure has been divided according to the statistics that Statistics Finland produces. Underneath each statistics there are folders for publications, statistical data and metadata of the statistics. In figure 4 there is a view of the database as it appears in the Epic editor database connector interface.

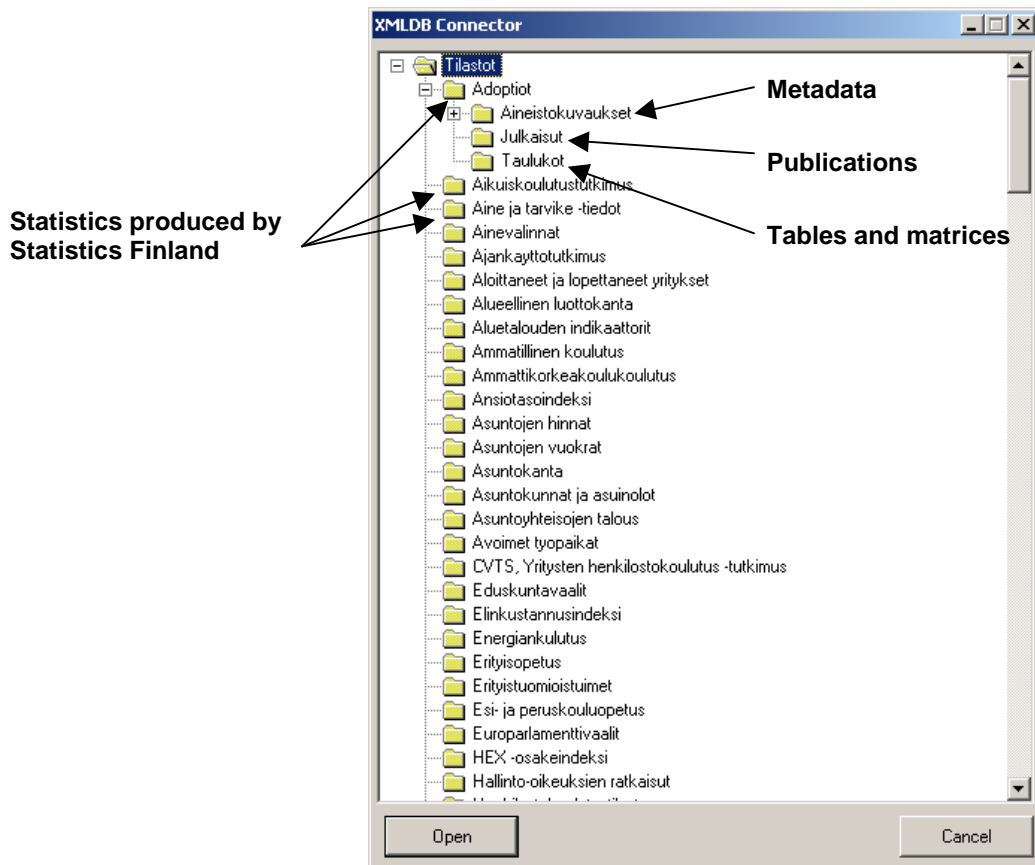


Figure 4. View of the XML database in Epic editor

B. Epic editor

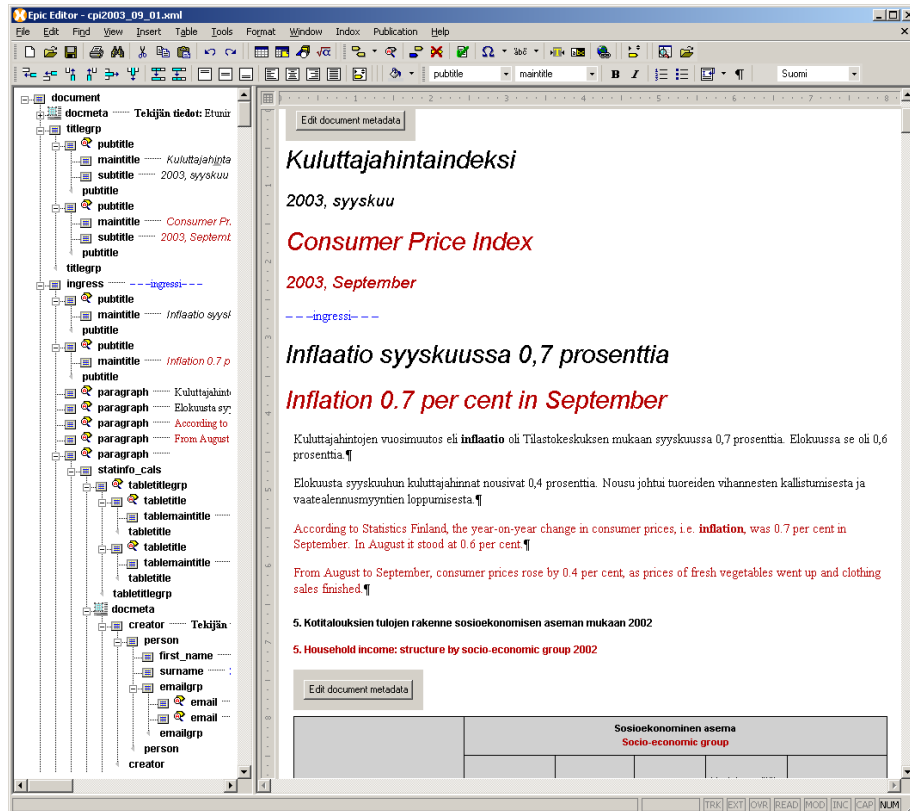


Figure 5. View of a statistical publication in Epic editor

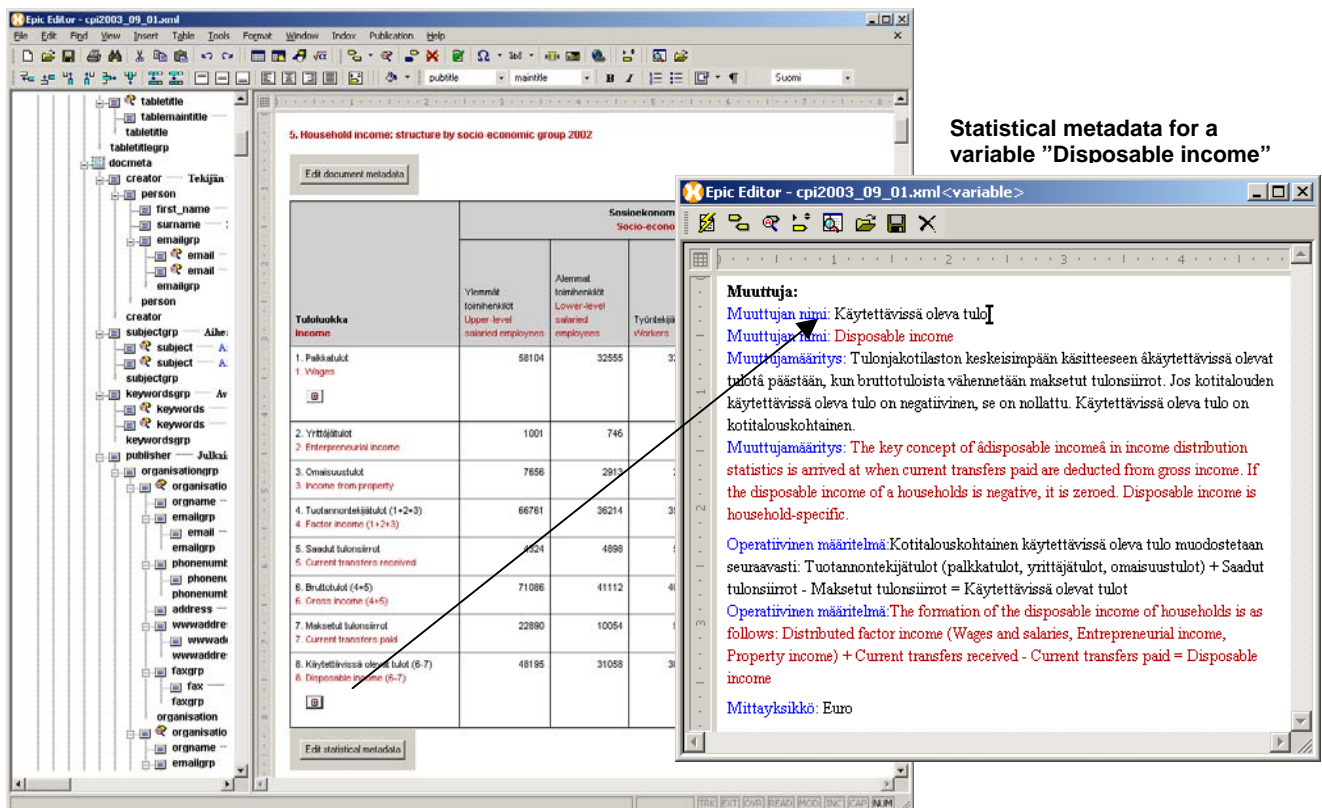


Figure 6. Statistical metadata for a variable in a table

C. HTML output

The screenshot shows two browser windows from Statistics Finland. The left window displays the 'Consumer Price Index' for September 2003, with an inflation rate of 0.7 per cent. Below this, it shows '5. Household income: structure by socio-economic group 2002' with a table of income components. The right window shows the 'Statistical metadata for a table "5. Household income: structure by socio-economic group 2002"'. A black arrow points from the table in the left window to the metadata page in the right window, with the text 'Link to the statistical metadata' next to it.

Income	Socio-economic group			
	Upper-level salaried employees	Lower-level salaried employees	Workers	Employers and own-account workers in agriculture
1. Wages ^c	58104	32555	32427	76
2. Entrepreneurial income	1001	746	675	292
3. Income from property	7656	2913	2356	76
4. Factor income (1+2+3)	66761	36214	35459	447
5. Current transfers received	4324	4898	5180	74
6. Gross income (4+5)	71086	41112	40639	521
7. Current transfers paid	22890	10054	9917	105
8. Disposable income (6-7) ^c	48195	31058	30722	416

1) Socio-economic groups are comparable with the previous years only on the basis of the year 2002.

Contents

- Inflation 0.7 per cent in September
- Maps
- Comparative international data

Statistical metadata for a table "5. Household income: structure by socio-economic group 2002"

Variable: Wages and salaries

Concept definition:

Wages and salaries refer to the compensations as money or benefits in kind received by households or persons during the year. The acquisition costs, excluding travel costs, of wages and salaries are deducted from them. The concept of wages and salaries used in income distribution statistics comprises pay for regular working hours, as well as overtime pay and income from a secondary job.

Operational definition:

Wages and salaries = cash income + benefits in kind based on employment relationship + reimbursement of costs based on employment relationship - wage and salary acquisition costs (excl. travel costs)

Variable: Disposable income

Concept definition:

The key concept of disposable income in income distribution statistics is arrived at when current transfers paid are deducted from gross income. If the disposable income of a household is negative, it is zeroed. Disposable income is household-specific.

Operational definition:

The formation of the disposable income of households is as follows: Distributed factor income (Wages and salaries, Entrepreneurial income, Property income) + Current transfers received - Current transfers paid = Disposable income

Figure 7. HTML output of a statistical publication with statistical metadata

V. REFERENCES

Rouhuvirta H., An alternative approach to metadata – CoSSI and modelling of metadata, CODACMOS European seminar Bratislava 7th October 2004, Project IST-2001-38636.

Available on the web at:

http://www.stat.fi/org/tut/dthemes/papers/alternative_approach_to_metadata_codacmos_2004.pdf

Rouhuvirta, H. and Lehtinen, H., Common Structure of Statistical Information (CoSSI) - Definition Descriptions, 2nd December 2003, Version 0.9, Statistics Finland 2003.

Available on the web at:

http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf

Rouhuvirta, H., Conceptual Modelling of administrative register information and XML - Taxation metadata as an example. Conference of European Statisticians - Work Session on Statistical Data Editing, Ottawa, Canada, 16-18 May 2005

Quality Guidelines for Official Statistics, Statistics Finland, 2002.

Available on the web at: <http://www.stat.fi/qualityguidelines>

WP. 21
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and EUROPEAN COMMISSION
ECONOMIC COMMISSION FOR EUROPE STATISTICAL OFFICE OF THE
CONFERENCE OF EUROPEAN STATISTICIANS 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 (iii): Metadata and Statistical Cycle

**QUALITY INFRASTRUCTURE SYSTEM - A CASE STUDY OF AN E2E
APPLICATION AT THE ABS**

Supporting Paper

Submitted by Graeme Oakley, Australian Bureau of Statistics
[graeme.oakley@abs.gov.au]

I. INTRODUCTION

1. The Australian Bureau of Statistics (ABS) has the objective, as part of its electronic publishing of metadata vision, to produce information about the quality of a statistical dataset alongside the actual data. The information about the quality of the data would follow a framework similar to that used by the OECD eg relevance, accuracy, coherence etc. It would have a descriptive component (text, relatively static) and metrics such as response rates of the survey, standard error of main variables (numeric, changing with each production cycle).
2. This case study explores the architecture of the Quality Infrastructure System (QIS) to capture the data that contributes to the quality metadata metrics from various processes through the statistical cycle. This data is associated with descriptive metadata from the Collection Management System to create a Quality Declaration which is associated with statistical products published to the web site. This development is still a work in progress,

however, several components are in production and deployment of back-end processes into surveys has commenced.

II. OBJECTIVES OF THE QUALITY INFRASTRUCTURE SYSTEM

4. The QIS has its origins in a project called 'Making Quality Visible' which aimed to help ABS progress its mission to “provide a high, quality, objective and responsive national statistical service”. The 'Making Quality Visible' focus has four main objectives. Firstly, the ABS is committed to improving dissemination by increasing the information available externally about the quality of ABS data. The second aim is to educate users about the different dimensions of quality and how to use information on data quality to make informed use of data. The third objective is to publish and promote guidelines and framework about quality. The fourth objective is to use information about data quality internally to manage and improve statistical processes.

5. The Quality Infrastructure System (QIS) is primarily focused on the first and fourth objectives, by providing an integrated architecture to define, collect, store, access and analyse quality measures. The objectives are summarised as being to:

- understand the performance of statistical processing functions
- understand the quality of statistical outputs
- enable methodologist and subject matter specialists to drill down to understand why performance, or quality, is what it is
- manage quality and statistical processes more efficiently
- deliver to users (either internal or external) reports on performance and quality that *Make Quality Visible*.

6. It is envisaged that these objectives will be met by creating a repository of measures integrated with the Corporate Metadata Strategy. The system will:

- automatically capture quality measures
- store quality measures
- enable desktop viewing of quality measures
- enable dissemination of quality measures both within and outside the ABS.

7. The implementation of the repository will be achieved by:

- defining the set of interactions that must occur with a repository of quality and performance measures
- defining and building Application Program Interfaces (APIs) to support those interactions
- populating it with performance and quality measures (for both within cycle and end of cycle measures) that have been defined by one authority
- defining metadata for quality measures. It is envisaged that a basic set of measures, aimed at clearance documentation and quality statements, will be defined by Methodology Division but that others may be added by users for their own purposes.

8. The information stored in the repository will:

- be equally applicable to business surveys, household surveys, the population census, and indeed any other collection operations

- have useable interfaces for a range of purposes, especially exploring, viewing, presenting, and analysing measures
- be seamlessly interfaced with existing or forthcoming IT tools, as well as other corporate metadata repositories
- be supported by definitional metadata unambiguously defining quality data items and active metadata.

III. SOME DETAILS OF THE SYSTEM

9. The QIS project has a strong metadata focus. While quality measures are a class of metadata in themselves, they require their own definitional and operational metadata. The Methodology Division acts as the registration authority for standard quality measures. This aids in consistency and comparability across collections. Subject Matter areas are still able to define their own custom measures.

10. The key principles of the ABS Metadata Management framework that are invoked in this project include:

- requirement for a 'registration authority' for each metadata element, which is the single authoritative source;
- reuse metadata; and
- capture metadata at source.

11. Metadata related to quality measures that needs to be stored:

- definitions of quality measures
- operational metadata enabling derivation of derived quality measures
- quality measures required to be collected
- levels at which quality measures should be collected
- frequency at which quality measures should be collected
- content of customised reports (including what quality measures to include, when to report, who to report to)
- structure and content of clearance documentation
- registration authority details.

12. Central to the development of QIS was the production a repository for the storage of quality measures. The repository is built around a “star schema“ data warehouse. This model allows for a variety of different types and levels of quality measures to be stored for both economic and household collections. A star schema is composed of a central “fact table”, a database table that contains the observed values of all quality measures and series of “keys” that link each quality observation or “fact” to a number of different “dimensions”. These dimensions are themselves database tables, containing the keys and classificatory details. Dimensions used in the QIS repository include: quality measure (e.g. response rate, standard error), geography (e.g. state, part of state) and industry. Wherever possible standard classifications are used, for example the industry dimension uses the Australian and New Zealand Standard Industry Classification (ANZSIC). The repository is built using the Oracle rdbms.

13. A 'services' architecture has been employed. Two different service types have been developed to store quality measures. The PutQM service enables existing systems (such

as provider management system or estimation system) to send quality measures directly to the QIS Repository. Existing donor systems undergo minor modification to enable them to send quality measure mail messages to the ABS business process management infrastructure. A “subscription service” monitors this process management infrastructure and automatically loads quality measures to the repository as they arrive.

14. The LoadQM services are the services that source quality measures from existing systems and load them to the QIS Repository. LoadQM services, independently from existing systems, access data stores these existing systems create and load these measures to the QIS Repository. These services do not require any re-engineering of donor systems. The PutQM services are preferred as it is based on a single integrated system.

15. On-line analytical products, such as Oracle Discoverer or SAS, can access this repository. Generic services have also been generated to allow standardised access to the repository. The StoreQM service allows quality measures to automatically copied from QIS to other parts of the ABS Corporate Metadata Repository. For example, the Collection Management System is a centralised store that records details on the concepts, sources, methods, timing, collection procedures and output of each ABS collection. The StoreQM service allows for the relevant quality measures to be included in this documentation.

16. The GetQM service is a standard service that allows analytical systems to specify the quality measures required and retrieves these from the QIS repository. These analytical systems allow the user to explore the quality measures, drilling down to finer levels of details or making comparisons across data collections and collection cycles. These services again use the ABS business process management infrastructure to access the quality measure repository.

17. Diagram 1 illustrates the system.

18. How does the dissemination end of the statistical cycle work? Diagram 2 shows the production of the Quality Declaration (QD). The reader should note that ABS does not currently produce and publish a QD - that is work in progress. We still have to finalise the content of the QD in terms of the static text (eg related to relevance, coherence) and the metrics that vary from collection cycle to cycle. Once that is completed then the development of a QD template and product production process will follow a model that we already use for our Directory of Statistical Sources.

Diagram 1. The Quality Infrastructure System

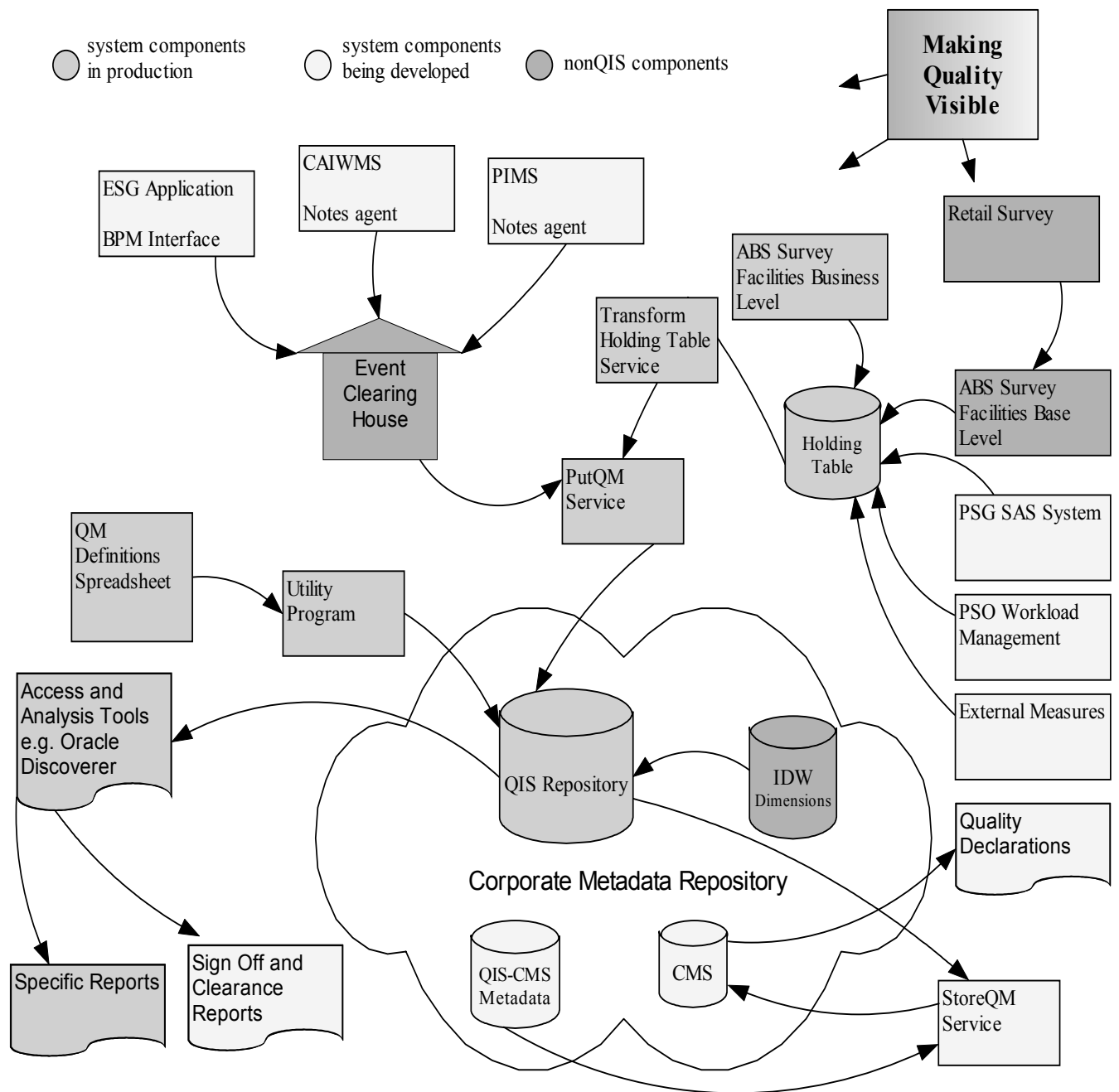
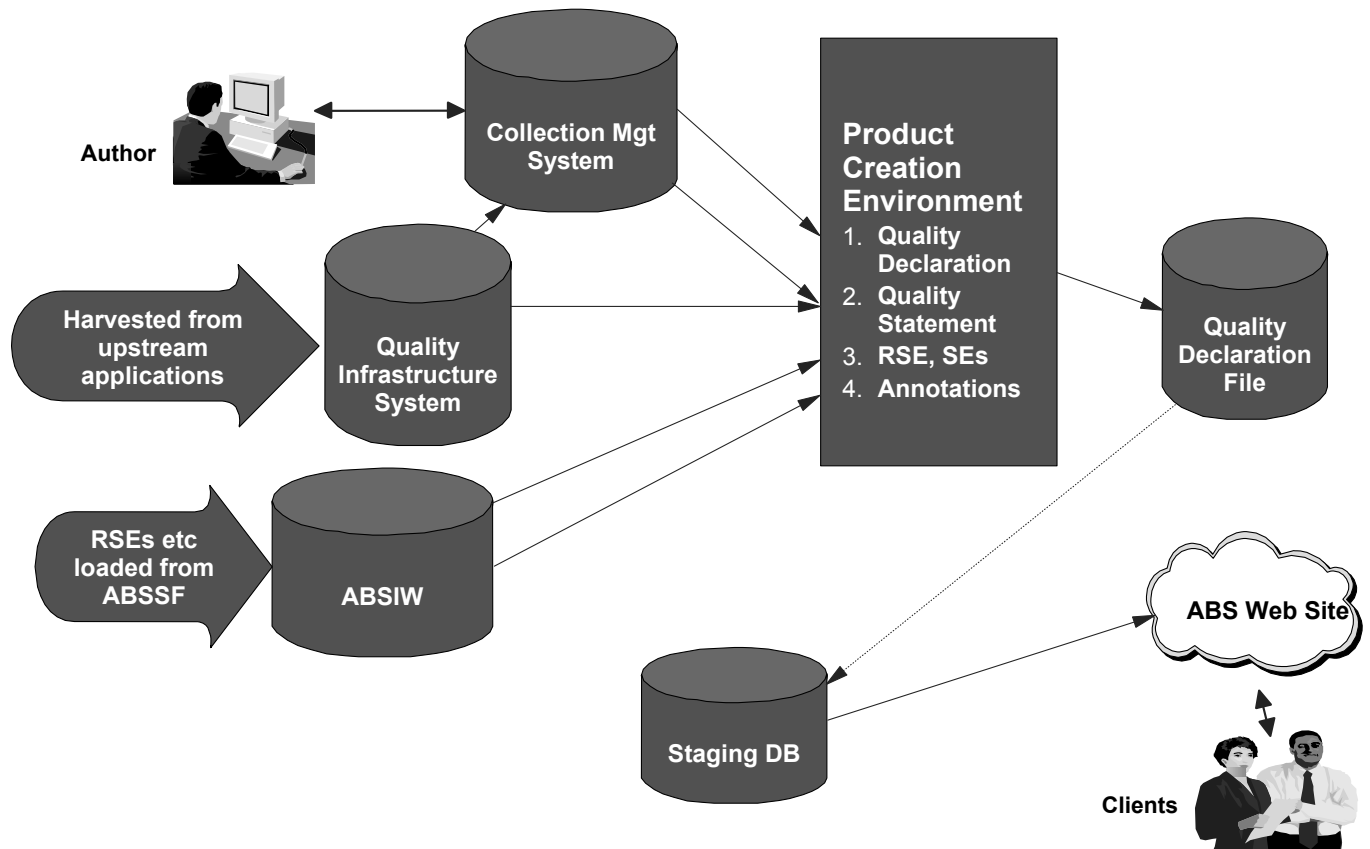


Diagram 2 Production of Dissemination Products

Quality Declaration



IV. CONCLUSION

19. What are some of the issues that are being addressed by this project? They are probably a microcosm of the metadata issues faced by any project that spans a large number of different collection domains and types, being both cultural, people and technical. Some examples of these issues are:

- agreeing on definitions for quality metrics across a large number of subject matter survey areas. Historically, collection areas have developed their own management information system and the definition of data elements is not always consistent, although the same data element name may be used. ABS Methodology Division is attempting to implement standard definitions and terminology, backed up by a 'registration process'. This process is a lengthy negotiation to establish agreement and buy-in, and has very little to do with the tools and systems that are provided. This is a common metadata management issue that requires time, persistence, senior management endorsement for the objectives of the project and good change management processes.
- deployment and timetabling issues. Deployment of the quality infrastructure system has to fit in with the 'business as usual' processing cycle ie not interfere with the core work of the collection. Again a good change management process is needed because the collection area may not perceive any direct benefit to them. The benefits accrue to others eg methodologists who will eventually only

have to work with one system to access quality information from many surveys. ABS is expecting a lengthy process for deployment in order to take into account survey processing schedules - could be up to 18 months to establish the bulk of the collections into the system.

- differences between internal and external clients with respect to quality metadata. The quality metadata required for association with disseminated products is a subset (and possibly derivation) of the basic observation data collected into the quality metadata repository through the statistical processing cycle. Internal users such as methodologists and survey managers are most interested in the detail and being able to drill down. This leads to tensions in terms of focus on the development of facilities (effort devoted to internal requirements ahead of dissemination) and the preparation of descriptive and analytical text (comes second to internal processes).
- efficiency and size of the database, throughput. The infrastructure could produce many data observations during a collection cycle when all the end-to-end processes are instrumented. The database could become large and so careful management by the database administrator will be necessary to ensure that updating efficiency and access efficiency are given appropriate weight. Also, the architecture based on services and an 'event clearing house' could lead to significant network traffic which means attention to throughput will be needed.
- security. Probably the major issue to be considered because the quality information being collected will include early estimates with standard errors, ahead of the official embargo time. Therefore, only a limited number of officers should have access to the quality data. The quality infrastructure system repository needs a robust security model which ensures that the survey manager controls who can have access to sensitive quality measures, and BI tools are set-up to prevent access. [Note that only ABS staff could possibly access the repository because it is on the internal network, but our enterprise risk management strategy requires a 'need to know' approach to access to embargoed statistical data.
- appropriate allocation of resources to areas that have to instrument systems. Availability of resources is always an issue. In this case, many processes in the statistical cycle could require instrumentation, although the increasing use in the ABS of generalised processing engines means that instrumenting one processing engine can pick up many surveys. Still resources are required to do this work and then the survey area will need to specify some 'process metadata' to 'drive' the particular process for their survey eg specify details for some core quality metrics and decide if they want to define collection specific metrics.

20. The Quality Infrastructure System has a number of components in production, the core (phase 0) quality metrics have been agreed and registered, and instrumentation is starting in a few of the processes that service a number of surveys, such as the ABS Survey Facilities (providing estimation and imputation service), and the Economic Surveys Data Centre (providing measures related to population and sample sizes, responses etc). Work has commenced in defining requirements for internal reports, such as 'clearance documents', and the external dissemination products.

**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 (iii): Metadata and the Statistical Cycle

METADATA AS A CRUCIAL STARTING LINK IN NEW STATISTICAL CYCLES

Supporting Paper

Submitted by Statistics Netherlands ¹

I. INTRODUCTION

1. The urgency of using administrative data as the primary data source for statistics is becoming more and more important. Statistics Netherlands even has the legal obligation to first investigate all possibilities for using administrative data sources before sending out/setting up a survey. In the Netherlands now several large (national) projects are being implemented to develop unique basic registrations based upon the motto 'singular (data) collection, multiple use'. At the moment Statistics Netherlands is using more and more datasets from these registrations as primary source for statistics and some statistics are even already completely based upon administrative data sources.
2. In this context it is necessary to develop and implement a different and new start of the statistical cycle. Integrating administrative data in the statistical cycle as (partial) substitution for survey data is long and difficult process. Besides the cultural/emotional fact that often it is very difficult to cope with changes and the new developments, the highest threshold that has to be crossed is gathering maximum knowledge about the unknown new data sources. The biggest challenge is making the right match between the administrative data and the survey data that have to be replaced.
3. Metadata play an essential part in this process, it is the key to a successful implementation of administrative data as the primary data source for your statistics. Looking at the statistical cycle from this perspective, there following three phases are most affected: a) data collection/input b) data transformation and c) output/using the prepared administrative data. It is crucial to gather the right metadata in each phase and then use it for the right purpose.
4. This paper focuses on various aspects and different roles/values of metadata as an essential element for the (right) use of metadata in the implementation of administrative data in the statistical cycle. The paper is organised as follows:

¹ Prepared by Harry Goossens, hgoo@cbs.nl

- In Section II we describe several problems and challenges we face in the implementation process related to metadata;
- Section III looks into various aspects of metadata, more generally;
- Section IV describes a method how to determine the right metadata for the various purposes;
- In Section V some results are presented;
- In Section VI finally some persisting difficulties and further developments are addressed to.

II. PROBLEMS & CHALLENGES

5. Although Statistics Netherlands already have successfully implemented several administrative datasets as primary source for statistics and even some statistics already are completely based on administrative data (for example the financial statistics on small enterprises / SFKO), it is still very difficult to realise new succesful implementations.

6. One of the main reasons for this fact is the growing complexity of the internal statistical process. Statistics Netherlands started in 1999 with the implementation of a new, uniform statistical process. This project, called IMPECT, aimed to redesign and standardize the various statistical processes. This resulted in new internal dependencies. Parallel to the change of processes also the organisational structure was changed, resulting in a more efficient structure. This was the end of the so-called 'stovepipes', specific organisational units for each individual statistic and branch of business (see also van Velzen, 2005).

7. The result of these changes was also that the clear often 1:1 relation between input and output transformed in at least 1:n. The bigger gap and the fact that link between input and output is no longer clear and direct make it much more difficult and complex to compare and match new input variables from external data sources with the output variables. Based upon the old structure with direct relations, the matching of variables mostly could be done from a strong focus on just the specific value of the single data elements, without the need of also looking at the process as a total. In the new situation the insufficient use of metadata is becoming a growing negative and aggravating factor in the implementation process(es).

8. Another factor is that in addition to the legal obligation for Statistics Netherlands to use as much as possible administrative data, there are large projects in progress aiming to develop a general system with an integrated set of national basic registrations for all governmental use. Although this makes much more datasets available and accessible, those datasets are part of a large and complex data infrastructure, which on the other hand makes use more difficult. Using this data demands very good knowledge of the data, process and subject in the broadest possible way.

9. In our modern with the all ICT possibilities regarding data management, the challenge of growing complexity is in fact a very common phenomenon that by the changes at Statistics Netherlands has only become a bit bigger and more specific. The last, also more common problem that must be noticed in this context is the fact that administrative data sources are almost always are collected for usage other than statistics. This means that you need to know that specific usage, including a certain expertise on that specialty.

10. Looking at all these challenges, we concluded if you successfully want to use administrative data it is inevitable to start with investigating the metadata first, before looking into the actual data. It is best doing this from an overview of the whole process that is concerned. And in addition to that, the more complex the process/environment is, the more it is necessary to do this in a structured way. This can be expressed in on central question: *'How can we first determine and then define the metadata we need to get the right information and knowledge about (new) administrative data sources that is needed for a succesful implementation of those data sources ?'* Since it is much easier to find something if you know what you are looking for.....

III. ASPECTS OF METADATA

A. Definition of metadata

11. Studying metadata often lead to discussions about the definition what is metadata and what is not. Starting from our central question we find this not really relevant. Much more it is important to know what information you want need and from there, how you can gather the required metadata. Therefore, it is essential to focus on the metadata that is essential for the specific process that is being handled. Mostly a massive package of often unstructured (meta)data is available, so first it is important to define the minimal set of metadata that is necessary and zoom in on that. After that, use the metadata together with the usage of the actual data, where this concerns both the use within the own internal process (such as data transformation for imputation, validation etc.) and the external use by clients (making statistics). Hereby it is important that you secure both the data as the corresponding metadata, as the meta reports the transformation of the actual data and sometimes also transforms.

B. Kinds of metadata

12. Although it seems to be very obvious, it is important to realise that there are different kinds of metadata and that there is not just one type of metadata. The more you study the metadata of a specific subject/dataset, the more differences you will discover. Therefore it is necessary to determine different categories to get a good overview of the specific metadata/information that is needed. In this context often the terms micro- and macro-metadata are used, but there are no standards. Depending on the dataset you are handling and what for you can just distinguish more or less (sub)categories that is convenient for you. It is mainly a question of using good common sense.

13. In the Dutch basic registrations project the following categories are used:

- *technical metadata*: specifications of variables, field length, data types etc., often defined in ICT specifications of information systems;
- *content metadata*: definitions en descriptions of the meaning of data elements, often defined in catalogues, as well as general knowledge about the specific subject. If for example using VAT data, you not only need to know the definition of ‘turnover’ that the owner of the data source uses in this dataset, it is also important that you have thorough knowledge of tax laws in general and VAT more specific
- *metadata about the process* (not to mix up with process data !):
for good understanding and correct use of external data sources it is necessary to have good understandings of the process that leads to the data source that is delivered to you; not only what transformations, calculations etc. are performed by the supplier, but also the primary purpose for which the dataset is collected and which administration is held about that process.

C. Purposes of metadata

14. Before starting to use an administrative data source there a several aspects and criteria that need to be checked and approved. That can be more factual criteria as well as process related criteria. Either way you need information based upon which you can decide if a criterion is approved or not. From this point of view it is very helpful to first determine specific purposes you want to use the metadata for.

15. Here also there is no standard but much more it is predetermined within each specific process for which you are investigating the metadata. For example commonly used factual purposes are *quality of the content of dataset*, *completeness of the dataset* or *timelags between and reliability of deliveries*. Process related purposes often concern *matching input variables from the source with the expected output for statistics or monitoring process steps*.

D. Organisation

16. The organisation of the management of metadata is another crucial aspect for making the metadata accessible and useable. Therefore it is important to spend sufficient time and effort on it. Often the focus lies on technical solutions and developing systems for metadata management. Although these are important parts, they bring little advantage if not everyone involved is convinced of the importance of metadata. Awareness and discipline in following the procedures are minimal equally important as the tools. But this in fact is a theme for itself, so we will not further discuss it in this paper. There is a lot of documentation already available see on this topic, for instance 'Wallgren & Wallgren' or the Swedish R&D report 2001.

IV. DETERMINATION METHOD

17. In spite of success in the past, recently it was problematic for us to give a quick and thorough answer on an apparently simple question. In a project concerning statistics of large enterprises, the goal was to define a dataset as a consistency matrix of 14 key variables on the level of enterprise groups. Therefore we had to match these 14 variables unambiguously with the source files of the underlying statistics (mainly based upon survey data) as well as with the basic statistic files build from tax data.
18. Explaining and discussing this with the (angry) project manager the reason why this action was so problematic became quite obvious. Positive was that the management of the metadata is being done, but not in a structured and coordinated way. The fact that it is being done at various places and in various ways, not systematically, using different codesets, with no coordination (= no organisation of the management of metadata) made it very difficult to gather the information we need to make that clear match.
19. Though the project of developing the system for national basic registrations focuses on a completely different process, namely the integration of various distinctive datasets into an overall system for generic usage, it faced the same problem. Participating in a special working group we had to answer the following question: *'What minimal set of metadata is necessary to make the system work.'* Trying to answer this question we experienced that it is very difficult to limit the huge amount of metadata to the really essential subset. To do so we first set up a structured way of working, partially based on the experiences from Statistics Netherlands.
20. We started with the distinction of the different main steps of the process and for each step we determined the various purposes metadata is needed for. All these different purposes did we combine to one set for the total process. Then we determined the various kinds of metadata we could find, looking at the total process. In this determination action it is important not trying to be complete, resulting in too much details, you must maintain a certain level of abstraction. The next step was to make a cross reference table, setting out the purposes facing the kinds of meta. For each cell we then described in common language what information we wanted to now and subsequently started to determine which metadata (elements) provide that information and where (at which registration) that meta is administered.
21. The first time we filled out such a table it was a vast struggle as we sometimes were being trapped in discussions whether a (set of) elements were metadata or actual data, or we had defined enough/the right purposes, but after a while we managed to stay in the right perspective and the method worked well. We experienced also that if you point out a metadata element in a cell it is beneficial to see which other cells (= purposes) that specific element is also needed and see if it perhaps changes during the process. And though at the beginning it is a lot of work that seems to take (too) much time, if being handy with this way of working and not going too detailed, results are coming up quickly.

22. For the basic registration project we defined the following cross reference table of 6 purposes facing 4 kinds of metadata. As it is now being worked out in detail, this contains only examples that should illustrate how it is used.

KINDS \ PURPOSES	PURPOSES					
	Collecting data, Input in diverse registrations	Delivering data of the system, data exchange within system	Quality check and assurance	Usage of the delivered data	Viewing	Archivation
Technical Specs						
Catalogue Specs (definitions, general descriptions)	<i>What is the exact meaning of an element</i>			<i>Which services does the system provide</i>	<i>What information does the system provide</i>	
Description of the administration, how it is set up and managed	<i>What accuracy, timelags are used</i>		<i>What accuracy, timelags are used</i>			
Meta of the process of the administration, how is the progress, status information			<i>What status information is available</i>			

23. Looking at the process of implementing administrative data sources we determined 3 process steps :
- Input, gathering and basic preparing the data for use
 - Developing a new method of statistics that must be applied
 - The actual implementation in the statistical cycle
- As mentioned before the definition of purposes and kinds is very related to the specific process when handling an administrative data source.
24. Based upon these experiences in applying administrative data, we are now setting up a new start of the statistical cycle, which aims to first determine and specify the metadata in a structured and more general way, before looking into the data itself. This way of working also provides a good and easier start to set up a clear and good organisation of the metadata management. Yet this way of working is just the beginning, there are still several aspects that now must be handled using this method.

V. RESULTS

25. Until the statistical year 2003 Statistics Netherlands made the annual structural business statistic completely based upon survey data. As a consequence of the new statistical law a project was started to develop a new method that makes it possible to use as much as tax data as possible.
26. This statistic is made for 6 branches, where each branch is split up in groups of enterprises with homogeneous main activities, the so called kernel cells. Further we use 9 size categories, based upon the number of employees. For the various combinations of branch and size there are 8 basic questionnaires, consisting of a generic part and a branch specific part. Depending on several possible combinations of branch, size and cell, in total there about 200 different questionnaires.
27. First the project focused on the small enterprises, i.e. having < 10 employees, being ca. 92 % of all enterprises and representing ca. 20% of the turnover. Based upon the FESS method (Fiscal Estimate Sbs System) the project made it possible that over the year 2004 (send out in 2005) Statistics Netherlands could realize a reduction of 16.000 questionnaires, on a total of 86.500 over all categories. For the small enterprises this means that 33 kernel cells can be completely based on tax data and 22 partially, resulting in only sending out 1.480 questionnaires, i.e. 8 % of the original statistical sample.
28. Recently the second phase of the investigation studied the middle class enterprises, i.e. up until a maximum of 100 employees (category 4-6). For this category, it was identified that 10 % – 15 % of the kernel cells can be based on tax data. Before it is possible to implement this method for other categories, it is necessary to investigate if it is stable over a minimum of 3 years.
29. For large enterprises it is not possible to use this method. The main reason is that as a result of the complex enterprise structures, the number of enterprises that can be unambiguously linked to the tax data is too small to be representative.

V. FUTURE DEVELOPMENTS

A. Persisting problems

30. One of the most persisting thresholds in the process of implementing administrative data is the cultural or emotional aspect. One of the main characteristics of making statistics is the fact that statistics must be vast and stable over a long period, so every new change is seen as a disturbance and threat. When investigating the possibilities of using administrative data as substitution for survey data they always first look at the 'NOTS', the things that are not possible, don't match, instead of looking at the things that can be realised. This cultural aspect needs constant attention and just by coming up with practical results you can convince people.
31. Another problem is that if you want to use administrative data for statistical means that data source often is collected for complete different primary objectives. It is very desirable but also very difficult to gain some influence on the original primary goal and put in some statistical elements.
32. The last crucial issue that should shortly be mentioned is more organisational. If you start to use large external data source you have to develop a business and data architecture that is suited for this new way of working (see the Swedish R&D report).

B. Future

33. At the moment within Statistics Netherlands, we are investigating which other possibilities we can find to use more administrative data in the broadest possible way. As mentioned above, the first results lead to reducing the amount of questionnaires being sent out for the business statistics. In addition to that there is now a project investigating the large enterprises. Further we now also are redesigning the questionnaire itself, so that enterprises have less questions to answer.
34. Another important development is that we participate in the National Taxonomy project, together with the tax collecting service and other ministries. The main goal is to define a general and harmonized set of data elements (variables) that are used in various processes in all governmental use. Based upon that harmonisation, it will then be possible for companies to deliver one uniform set of data to a central point of public authority, directly from their administrative software systems. All further governmental use must then be from that set.
35. The biggest difficulty here is that several different primary goals of collecting data (for instance tax versus statistics) must be combined and used the same definitions, which could lead to a lot of overlap. So every participating party must be willing to compromise. But who really wants to change?

References

- J. van Velzen (2005), *The embedding of a uniform statistical process*;
UN/ECE Work session on Statistical Data Editing, Ottawa
- Statistics Sweden (2001), *The future developement of the Swedish register system*;
Final R&D Report of the Register Project
- A. Wallgren & B. Wallgren, *Register Based Statistics – administrative data for statistical purposes*.

Implementation

Short introduction to Session IV:

By Max Booleman

The Statistical Metadata Framework was developed by the task force following the recommendations of the METIS meeting of 2004. The Framework includes a chapter about practical experiences of national statistical offices (NSO's) that have recently implemented or re-engineered their statistical meta information systems.

Within this chapter the discussion on this topic will focus on:

- implementation planning and management;
- identification of recommended practice in the implementation of metadata systems, etc;
- usability considerations;
- infrastructure development options (e.g. build, buy, sharing and collaboration between NSO's, open source software);
- updating and improving statistical processes;
- change management:
 - influencing corporate culture (includes communication plans);
 - transition planning and management.

But any other business related with practical work on metadata could be added. See below for some examples.

In December 2005 every NSO has been requested to send their best practices to the task force. Two countries already uploaded some papers (Australia, Sweden). These papers will become available on the Internet.

To identify the papers it is very helpful to use a template containing at least and very brief the following items:

- Author and organisation
- What did you do (Project plan)?
- Why did you do it (Business case)?
- How did you do it (Project process)?
- What are the results (Project results)?
- What are the good and bad experiences?

The task force will develop such a template.

Some examples of 'any other business'

Some interesting topics will be addressed but it is certainly not a complete list:

- Co-ordination: Sometimes after but also within the implementation phase metadata can be used to achieve textual co-ordination. Are there some experiences?
- Language versions: metadata could support different language versions of concepts. But do we need also a formal and a popular version? Do we need also for different domains different naming of concepts? What are the experiences and practices?

- With the help of metadata all kind of relations between concepts, populations could be recorded and checked: profit = income – costs; paid salaries by enterprises = received salaries by persons; national total = sum of regions etc. The same could hold for statistical units: Persons belonging to households and working in enterprises. Is it already used in some NSO? Will it help to develop a coherent and consistent statistical system?
- Could the dimensions of an indicator be fuzzy or should it be crystal clear defined? Maybe it is more easy to develop an exact model but is it less easy to present our fuzzy indicators within that model.

Metadata could fulfil different functions. Related to statistical output do NSO's distinguish with the help of metadata:

- Versions of indicators (corrections, adjustments, revisions): differences in quality within more or less the same process;
- A system of products: single source, multiple source, integrated: different process, more or less the same indicator;
- A system of products: business cycle, monthly, quarterly, annual: different focus but nearby the same phenomenon;
- A quality dimension: for example do we expect or assume 'later' means 'better' for different versions or products?

Related to the statistical production process, do NSO's use metadata

- To describe their processes in a standardized way?
- To steer their processes to make the process more robust and reliable?

**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.

**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

DEVELOPING A SYSTEM FOR DESCRIPTION OF MICRODATA AT STATISTICS SWEDEN

Invited Paper

Submitted by Statistics Sweden¹

I. ABSTRACT

1. At present the Statistics Sweden metadata system consists of a number of tools and templates. The Quality Declaration template and the SCBDOK template for documentation of the production processes are the cornerstones of the metadata system. The software tool Metadok is a system for creating formalised metadata for the purposes of describing final observation registers in SCBDOK. However, owing to increasing demand for metadata, the support Metadok can offer is now considered inadequate for coordinating metadata and standardisation.
2. The aim of the ongoing project is to build a new model and a corresponding software tool that will provide an overview of the contents of the microdata registers at Statistics Sweden. Another important aspect is to facilitate documentation by making it easier to reuse and modify existing documentation.
3. The system is going to contain metadata about object classes, populations, variables and value domains including classifications. It will be a tool for standardisation of populations, variables, value domains and for documenting Microdata.
4. A search tool function making it possible to search for metadata about macro and microdata by different criteria or combinations of criteria like object class, variables, reference times, statistical products as well as subject matter areas will be a key function in the system. This and other types of functionality will support staff in several situations such as:
 - Evaluation if a need for information can be fulfilled with existing statistics or existing microdata without having to collect additional data, thereby reducing respondent burden and costs.
 - Survey design (frame construction, auxiliary information in estimation or editing).
 - Documentation (recycle existing metadata, step-by-step documentation of new systems/products).
5. For external users and researchers the system will also provide a better overview of the data at Statistics Sweden.

¹ Prepared by Klas Blomqvist and K-E Kristiansson.

6. The system is built on a conceptual model that is based on existing metadata systems and templates used at Statistics Sweden such as SCBDOK and Metadok and the demands for metadata within the organisation. It is influenced by ISO-11179 and ongoing work on a variable model in the Neuchâtel group.

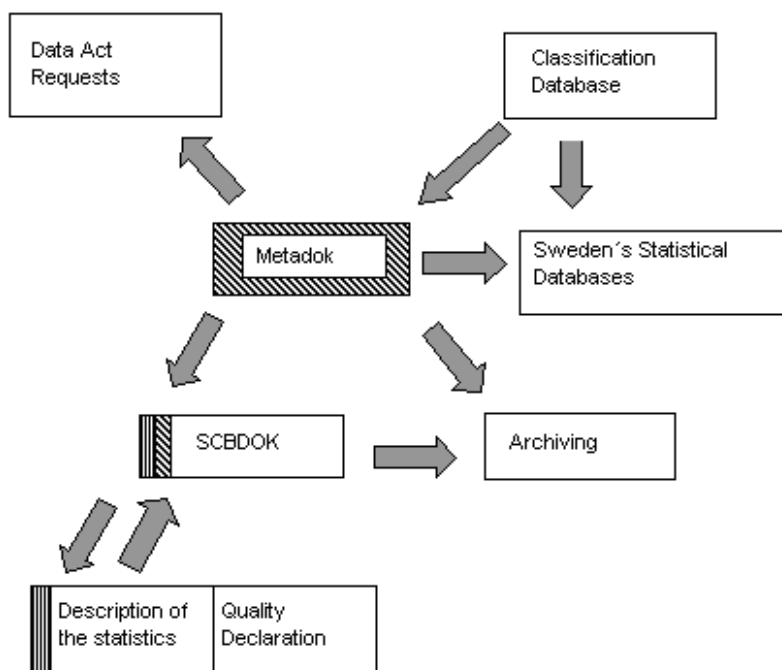
7. The first version of the system, which is for testing and filling the database with content, is to be released in January 2006.

8. The paper discusses the project goals, the conceptual model and the functionality of the system.

II. THE PRESENT SITUATION, AN OVERVIEW

9. The metadata system at Statistics Sweden today consists of several tools and templates. For each of these there are guides that describe the functionality and interaction with other parts of the system. The figure below gives an overview of how these interact.

Figure 1. The present metadata system at Statistics Sweden



10. A brief description of the different parts in the system is presented below.

A. Description of the statistics

11. The purpose of this template is to provide a short description of the quality of the statistics and other basic facts. It contains one section with general information and a quality declaration section. All official statistics should have a Description of the statistics according to law.

B. SCBDOK

12. Since 1994 all final observation registers and production systems that Statistics Sweden is responsible for should be documented in SCBDOK. The purpose is to provide a detailed description of the creation of a statistical register from data collection to dissemination. SCBDOK is a Word template. Some of the information in SCBDOK is included in Description of the statistics.

C. Metadok

13. Metadok is a program that is used for describing a physical database for a microdata register. The metadata in Metadok are formalised and can therefore be used by other software. The purpose is to make multi usage of the documentation possible, such as dissemination on the Internet, Data Act Requests, archiving and aggregations in PC-Axis. Metadok can also be added as a part of SCBDOK, for describing the content of the final observation register. The Metadok documentation is published on The Statistics Sweden Webpage as a part of Sweden's Statistical Databases. In order to reduce the documentation workload it is possible to import already existing metadata from several sources for example from the Classification database.

D. The classification database

14. The Classification database contains national as well as international classifications, such as regional classifications and activity classifications.

III. THE PROJECT

A. Background

15. The current system for documenting microdata *Metadok* has some limitations when it comes to content as well as functionality. The most important are listed below.

- It is difficult to get an overview of the metadata due to that the variable content of a statistical product is tied to one register at a time.
- The information in the registers is not sufficient for evaluating the possibilities for matching data from different surveys or registers.
- Lack of information regarding comparability over time.
- The insufficient information about the registers and the difficulties to get an overview of the data makes coordination and standardisation very difficult and time consuming.
- The support for documenting in Metadok is inadequate.

B. Goals

16. One important aspect is to develop a system for documentation of microdata that has an improved functionality and higher quality of the content than in Metadok. The system is going to work together with other parts of the metadata system. The system will:

- Provide an overview of metadata for the microdata.
- Provide information for evaluating the possibilities for matching data from different surveys or registers.
- Operate as a tool for increased standardisation and harmonisation and therefore give better possibilities for data coordination.
- Provide a better support for documenting so that the overall quality of the documentations can be increased.

C. Delimitation

17. The project does not deal with harmonisation and standardisation of content. Support and future development of the system is not a part of the project either.

D. Working practice in the project

18. The system is going to be an important tool for a large part of the staff at Statistics Sweden. Therefore user groups consisting of interested persons were tied to the project from the beginning. The user groups were categorised in subject-matter statisticians, methodologists and system developers. These groups took an active part in developing the content and functionality of the system in order to be a useful tool for them in their

everyday work. Other requirements have also been collected by the project group and taken into account when developing the system.

E. Project organisation

19. The project group is led by staff from the Methodology unit (Research and Development Department). In the project group there is also staff from the IT and Register and Microdata for Researchers units (also from the Research and Development Department) and the Information and Publishing Department. The project has a Reference group consisting of representatives of all departments at Statistics Sweden and a Steering committee where the current situation in the project is reported on a regular basis.

F. Examples of frequent work situations where VHS is a tool

20. The work with the collection of requirements and in the user groups has resulted in several areas or situations where the system can give a substantial support for the production.

General

- New enquiries.
- Consider whether an enquiry can be dealt with using already produced statistics or collected data.
- Searching for metadata on microdata that consists of relevant object, variables and population.
- Can different registers be used in order to make an integration register.
- Comparability over time, information on time series.
- New surveys.

Survey design

- Frame construction.
- Auxiliary information in the estimation.
- Auxiliary information in editing, coding and imputation.
- Reduction of over coverage in surveys considering certain subpopulations.

System development

- Modelling databases.
- Using code tables from production systems in connection to editing and coding.
- Search and provide information of where data is stored physically.

Documentation

- Reuse of old documentations
- Starting point in already existing metadata
 - Classifications and standards
 - Documentations made by others
 - Searching the metadata
- Document once – reuse

G. Extended demands for functionality and content in VHS.

21. The above-mentioned demands means extended requirements for functionality and content in VHS.

22. *The system shall be coordinated with other parts of the metadata system.*

This means that:

- There is a connection to other types of documentation such as SCBDOK, Description of the statistics, the system development tool SiP and a future questionnaire database. VHS also has to be connected to the product database, the archiving system, the system for Data act requests and the Statistical database SSD.

23. *The system shall provide information for evaluating the possibilities for matching data from different surveys or registers.*

This means that the system will contain:

- Content metadata, i.e. on variable, value domain, object class, and population
- Historical information about variables and value domains
- Comparability over time – inform about breaks in time series and changes in definitions.
- Accurate and up to date information on technical metadata, i.e. column name, server, database and table

24. *The system shall provide an overview of the microdata repository.*

This means the system will contain:

- A search function for survey design purposes, for example flexible search functions for selecting variables.
- The search function should be non hierarchical with multiple search entry possibilities.

25. *The system shall be a support tool for increased harmonisation and thereby better possibilities for coordinated use of data.*

This means the system will contain:

- Standards for object classes, variables and value domains.
- Consistent and distinct definitions.

26. *The system should give appropriate support for documentation so that quality in all documentations can be increased.*

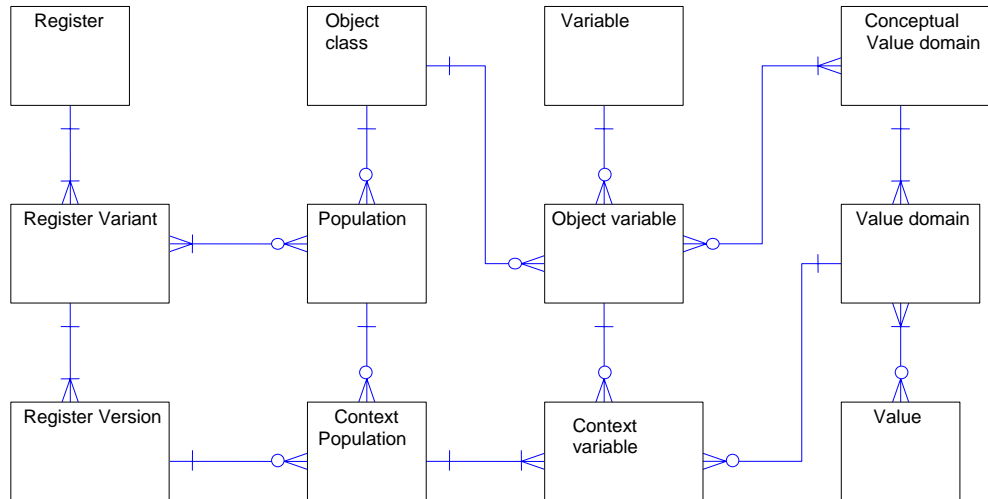
This means that in the system you:

- Will be able to document variables over time, without having to document a whole register.
- Will have one metadata container for each statistical product or enquiry, from which different types of documentation can be created.
- Will be able to create a metadata holder that does not belong to a specific statistical product, register or enquiry.

IV. THE CONCEPTUAL MODEL

27. The development process of the model has to a large extent been driven by the content demands and requirements stated above as the purpose of the project. Already defined metadata concepts and terms in the organisation have been used when possible. ISO 11179, Information Technology -- Metadata Registries (MDR) has been used as an input in the making of the model as well as ongoing work in the Neuchâtel group.

Figure 2. The figure below shows a summary of the core of the conceptual model



V. CONCEPTS

28. **Object class:** An *Object* class is an abstraction of an object. An object is an in itself independently existing entity. When used in a statistical context it becomes a statistical object. There are two types of objects that can be isolated (described) in the model.

- *Register object:* The Object class that is connected to the physical register.
- *Target object:* The Object class that the target population and the survey population consists of.

Example: person, local unit, organisation.

29. **Population:** Description not related to a specific survey round of the quantity of objects that the survey intends to collect data about. The description can be related to either the Population from which the register objects are originated, i.e. the Register population and/or the Population from which the target objects are originated, i.e. the Survey population.

- *Survey population:* The part of the frame population that is included in the target population.
- *Register population:* Description of the Population of Register objects that the Register encompasses totally or contains a sample from.

Example: persons in Sweden on December 31, local units in Sweden at January 1, organisations in Stockholm during the year.

30. **Context population:** Description of the quantity of objects that the survey intends to collect data about in the specific survey round. Context population concerns Context register population and Context survey population. The realisation of Register population and Survey population gives the actual populations for the current survey round.

Example: persons in Sweden at December 31 2005, local units in Sweden at January 1 2005, organisations in Stockholm during 2005.

31. **Register:** An overall *denomination* for the register that is the basis for the statistics.

Example: The population register, the business register.

32. **Register variant:** Common non-survey round dependent description of the Register variant that is used in the survey. One Register can have several Register variants. Different Register variants are distinguished by different definitions of their Register population or differences in variable content.

Example: The population register complete variant, the business register sample frame variant.

33. **Register version:** A Register version is a realisation of a Register variant at a certain point of time/survey round.

Example: The population register complete variant 2005, the business register sample frame variant 2005.

34. **Variable:** A Variable states a characteristic that can be connected to one or several Object classes.

Example: monthly income, annual turnover, industrial activity.

35. **Object variable:** A Variable that has been connected to an Object class.

Example: person monthly income, local unit annual turnover, organisation industrial activity.

36. **Context variable:** A Context variable is connected to an Object variable and has a connection to a Context population, a Value domain, a Register version and has an information source.

Example: monthly income for persons living in Sweden December 31 2005, local unit annual turnover 2005, organisations in Stockholm after industrial activity 2005.

37. **Conceptual value domain:** Presentation of level of detail and definition of a set of categories for a value domain and its categories conceptual meaning.

Example: SNI 2002 (Swedish Standard Industrial Classification 2002).

38. **Value domain:** The level of detail for a Variable and its representation with measure unit or alphanumerical specification (code) and definition of the different categories in the given level of detail.

Example: SNI 2002 (Swedish Standard Industrial Classification 2002) 5-digit level, 0- (Swedish kronor).

A. Technical information

39. The system provides links to where (i.e. server information, column name etc) the data is physically stored. This means that the user does not have to change names in their production systems as long as they link to the correct metadata in the system. Data is not a part of the system.

B. Sources

40. On all Register levels and Context variable there is information on sources for the given Register level or Variable.

C. Structure

41. The application as based on the model gives the following content structure for a given Register:

- Register123
 - Register variant A
 - Register version A1
- Content:*
- Population A1

- Variables
- Value domains
- Technical information
- Sources
- Register version A2
 - Content:*
 - Population A2
 - Variables
 - Value domains
 - Technical information
 - Sources
- Register variant B
 - Register version B1
 - Content:*
 - Population B
 - Variables
 - Value domains
 - Technical information
 - Sources

VI. WHAT'S GOING ON AT PRESENT

42. The collection of requirements for the system considering content and relations to other systems is closed. Based on this the database is now close to stable and the application is under construction. Due to this a lot of issues concerning functionality arise on a continuous basis. A beta version of the application was released in January. The user groups are now working actively giving input to the development of the application. In July a new version will be released ready for starting to fill the system with essential basic metadata about registers that are sources for many other registers and surveys. After that the system will continue to be filled completing it with the rest of the registers.

Figure 3. The figure below shows a “print screen” of the application version 1.

Registerversion	Population	Variabler	PUL	Tekniskinformation	Arkivering	Standard
	Population			TidFrom	TidTom	
Totala hushållsutgifter under ett år	Kosthushåll HUT			2003-01-01	2003-12-31	Produktstand...
Totala hushållsutgifter under ett år (ej dagli...	Kosthushåll HUT			2003-01-01	2003-12-31	Produktstand...
Totala hushållsutgifter under en två vecko...	Kosthushåll HUT			2003-01-01	2003-12-31	Produktstand...

Variabel
 Kortnamn:
 Namn:
 Definition:
 Beskrivning:
 Standardnivå:
 Baserad på:
 Summerbar: ☒ Summerbar
 Rekommenderat kolumnnamn:
 Rekommenderad datatyp:
 Utgivare:
 Tillgänglig:
☒ Kopplingsinformation finns [Visa mer](#)

Objektvariabel
 Namn:
 Definition:

A. Version 1

43. In January 2006 a first version was released. It has very limited functionality and is primarily used for testing and to get input from user groups. Information can be added in this version, but it is not intended to be used in production. The database contains Classifications, some standardised object classes, and standardised variables and a few test examples.

B. Version 2

44. Version 2 will be released in the beginning of the summer. This version will have increased functionality and support more working operations. It will contain registers, register variants, register versions and frequently used value domains. At this stage migration of metadata from Metadok can be started. How and to what extent metadata is to be migrated from the Metadok system is not yet established.

45. When the migration of metadata from Metadok is finished for a register the metadata has to be adjusted and completed in accordance to the structure of the system. Variables and value domains can then be harmonised in the system.

C. Version 3

46. Version 3 is a version where users can start using the application in the production. How much content the system will have from the start is determined by the level of ambition set by the departments in the migration process.

VII. HARMONISATION AND STANDARDISATION

47. The project has not dealt with actual content standardisation and harmonisation. To be able to fully take advantage of the functionality of the system there is a great need for content related work, mostly with standardisation issues. The project is responsible for the development of the system, not the content standardisation and harmonisation of the content. The system will as mentioned earlier be a tool for this and is also to a large part dependent on the content. If the system is to be used effectively there has to be information in the system to reuse.

**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

IMPLEMENTATION OF METASTORE AT THE OECD

Invited/Supporting Paper

Submitted by Russell Penlington & Lars Thygesen, OECD ¹

I. INTRODUCTION

1. MetaStore is the tool for managing statistical metadata in OECD. It supports a set of corporate principles for statistical metadata, giving guidance to how statistics should be described in order to allow users to use and understand the statistical data thus enhancing overall quality. The principles contain the following key notions:

- Attachment level: It must be possible to attach metadata to all levels of a statistical dataset: The whole dataset, a dimension, a dimension member, a time series, and one or more observations. Metadata should be attached at as high a level as possible to avoid duplication and inconsistencies.
- Metadata concepts. Metadata should be labelled under a common set of 41 metadata concepts which have been designed to cover all aspects of data characteristics, and to be aligned to similar concepts used in other statistical organisations; however, it is by no means mandatory to use all of these concepts for every dataset.
- Coherence across datasets: As far as possible, metadata for "the same" data should be managed in one place and referenced from other datasets.

2. MetaStore is a tool which fully supports these principles. It has been implemented progressively since early 2005. It is not a mandatory tool but dataset owners may use their own tools as long as they provide consistent metadata to the common data warehouse OECD.Stat.

3. This paper discusses the degree of success of this implementation; the difficulties encountered and measures taken; the coverage and coherence of the resulting metadata; which drivers have proven successful in the implementation; and secondary effects on visibility of OECD statistics.

¹ Prepared by Russell Penlington (russell.penlington@oecd.org) and Lars Thygesen (lars.thygesen@oecd.org)

II. OECD'S STATISTICAL INFORMATION SYSTEM

4. OECD statistical activities are carried out in several Directorates by statisticians with different backgrounds and work experience. More than 100 statistical activities are carried out by the OECD, the vast majority of which are devoted to collecting, processing and releasing data. Statistics are used daily by a wide range of OECD analysts, whose primary needs are to access data and metadata in the most efficient way. Finally, both internal and external users increasingly need to navigate across databases to carry out multi-domain analyses and comparisons.
5. Although the decentralised model has advantages, there are also disadvantages. The main problem areas are related to the efficiency of individual statistical processes and to the overall quality (in particular, coherence and methodological transparency) of OECD statistics from the user's perspective. To address these problems, a corporate Statistical Information System has been built during the period 2003-2006. It contains tools for data and metadata collection, manipulation, storage, dissemination, discovery and retrieval, preserving the independence of data producers while making their data and metadata part of a coherent and seamless corporate system.
6. In setting up of the system, the OECD has made use of best practices among peers in order to integrate such practices and as far as possible, avoid duplication of work.
7. The overall architecture of the Statistical Information System consists of three layers:
 - The production layer for the collection, validation, processing and management of statistical data and metadata;
 - The storage layer where validated statistics and related metadata are stored;
 - The dissemination layer for producing statistical publications and online/offline interactive statistical products.
8. The system comprises five independent but inter-operating components, one of which is the metadata management system MetaStore.

III. THE METADATA PRINCIPLES

9. The metadata of international organisations basically aim to give the information necessary to understand if it is meaningful to compare macro data between countries, and to understand how much weight can be attached to such a comparison. It must therefore describe what the data mean (concepts), the overall quality of each of the data elements presented, as well as differences in quality and differences in the meaning of the data between different subject matter areas as well as between countries.
10. The metadata must primarily illuminate the following areas:
 - Concepts, definitions of concepts, including such phenomena as measurement units, transformations.
 - Delimitation of populations.
 - Dimensions of quality, related to the original production, such as sample sizes, standard errors on estimates, and other kinds of known sources of errors such as non-response.
11. In 2004, OECD adopted a set of corporate principles laid down in *Management of Statistical Metadata at the OECD*². The following principles are the most important ones.

² <http://www.oecd.org/dataoecd/26/33/33869551.pdf>

A. Consistency

12. Statistical metadata must be consistent. This means that:
- The same variable name, definition, and other description should be connected to the same statistics, no matter where it is and who is the “owner”;
 - The same variable name should not be used for statistics that are not identical;
 - Terms and concepts should be consistent throughout;
 - All OECD metadata, particularly reference metadata, should be made readily and freely available to external users.

B. Redundancy

13. Metadata on one element (a statistical collection or dataflow, a concept) should only exist as one instance; no matter how many times the same element is reused in different contexts. Ownership to the metadata should be clearly defined.

C. Common metadata items

14. A set of 41 metadata items are defined. All metadata from different subject-matter areas must be grouped under these headings. These are similar to the SDMX Cross-domain concepts and have been developed concurrently with those. The ambition is to have the closest possible consistency between the two sets.

D. Attachment levels

15. Metadata can be attached at any level of detail of the statistical data: at the global level (pertaining to all datasets), at the dataset level, at dimension level within a dataset, at dimension member level, time series, and individual observations. To ease understanding and avoid repetition of data, it is recommended to always attach metadata at the highest possible (or reasonable) level; exceptions will then have to be stored for those lower levels where they apply.

IV. METASTORE AND ITS FEATURES

16. MetaStore is a general toolkit for accessing and managing reference metadata for statistics. Statistical metadata can be separated into two categories:

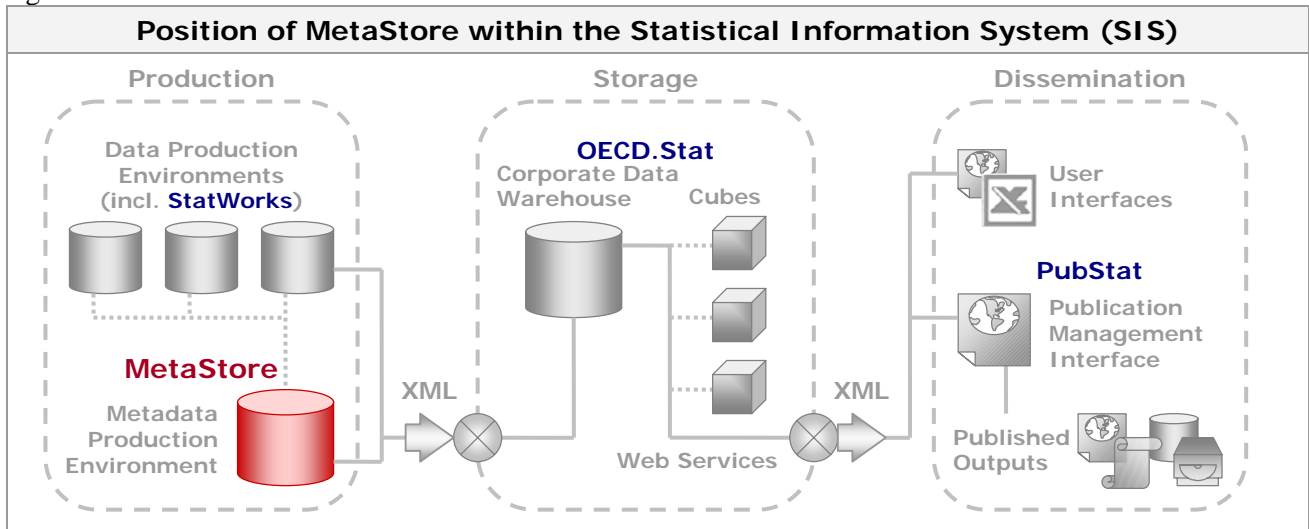
1. Structural Metadata describes the multi-dimensional structure of a dataset. Such metadata includes codes and names of dimensions and corresponding dimension members.
2. Reference Metadata is textual information documenting characteristics of the data within a dataset. Such metadata typically includes information concerning the collection, manipulation, purpose and quality of data at different levels of a dataset’s multi-dimensional structure.

17. MetaStore inherits Structural Metadata from the data production systems of the datasets. While MetaStore allows the storage of customised data attachment coordinate descriptions, it does not allow the management of Structural Metadata in the datasets’ production systems. MetaStore’s sole purpose is the management of Reference Metadata.

18. The ability of MetaStore to remain separated from the datasets’ data production systems allows it to be a common repository for managing Reference Metadata for datasets of different structures.

19. MetaStore is positioned in the production layer of the OECD Statistical Information System (SIS) for managing production reference metadata content. The following diagram outlines the position of each of components of the Statistical Information System.

Figure 1



A. Key Features

20. There are features and components of the MetaStore system for management of metadata both from within the web interface and from remote applications. The more notable features underpinning the core purposes of the system provide solutions to several common metadata management issues.

21. **Improving the quality of my statistical reference metadata:** While quality in statistics often refers to the data, the related metadata is an important vehicle to propel each of the dimensions of data quality. Furthermore, the same framework of quality dimensions can be equally applied to the metadata itself. Under the guidance of the Quality Framework and Guidelines for OECD Statistics (<http://www.oecd.org/statistics/qualityframework>), MetaStore aims to address issues for metadata accessibility, timeliness of metadata management, as well as coherence, interpretability, credibility, relevance and accuracy of metadata content.

- **Metadata Attachment** - MetaStore can attach metadata at any level of a dataset's structure. This is accomplished by allowing metadata to be attached at dataset and dimension levels and to all possible combinations of dimension members.
- **Rich Formatting** - Metadata can be entered, copied and edited with rich HTML formatting. The system's rich text formatting interface and integration with Microsoft Office allow existing content to be copied from all Office applications, as well as HTML sources such as web pages, with the formatting intact.
- **Standard Classification** – Metadata content must be classified into a set of common metadata item types which have been identified to improve the comparability and interpretability of metadata across datasets. Such types cover source information, collection, data characteristics, scope and coverage, statistical concepts, classifications, manipulation methods, and other aspects. Furthermore, several of the common metadata types must be chosen from a list of predefined texts or values rather than entered manually.

22. **Making metadata management more efficient:** Whilst the OECD does not regard cost-efficiency as a dimension of quality, it is a factor that must be taken into account in any analysis of quality as it can affect quality in all dimensions. If metadata can be produced and managed more efficiently with the same quality, then resources released can be used to improve quality in other areas.

- **Content Sharing** – Metadata content can be shared both within a dataset and between datasets. Ownership management prevents unauthorised update of content and the system automatically detects exact text matches to prevent duplication of metadata text within a single dataset.
- **Links & Flags** – Metadata can be further enriched by adding related hyperlinks with defined titles and links to the OECD Glossary of Statistical Terms (<http://stats.oecd.org/glossary/>). Items can also be flagged (on/off) for publication (whether shown in dissemination formats), archived (legacy versions), and private (draft versions for authoring user).

23. **Making metadata accessible to interested parties:** It is essential that reference metadata is accessible to both internal and external users to facilitate in the interpretability of the data for which is attached. The range of different users leads to such considerations as multiple dissemination formats and selective presentation of metadata.

- **Accessibility** – Metadata within the system can be accessed via a number of different methods. From remote applications, content can be directly accessed with the data coordinates loaded into the URL or extracted into the application via an ODBC query. Within the MetaStore web interface content can be accessed via broad search on coordinates and body text, dynamic filters for which criteria can be saved, and direct coordinate selection with the ability to find related coordinates with metadata attached.
- **Reporting & Exporting** – Metadata reports can be constructed for single data attachment coordinates or for a predefined number of levels below a specified data coordinate and output to HTML, Microsoft Word, Excel and XML for publishing. There also exists functionality to define, save and execute export criteria to feed XML formatted metadata into the OECD.Stat corporate data environment for dissemination.

24. **Integrating data production environments with MetaStore:** MetaStore is designed so that the functionality provided by the system uses components that can also be called from remote data production environments and applications. This means that the production databases can rely on MetaStore for metadata management. For detailed information on how to accomplish remote integration, please see the Remote Access in the MetaStore documentation available from the web interface.

- **Connectivity** - MetaStore is able to connect to and inherit the structure of any ODBC compatible database format (This includes databases such as SQL Server and Microsoft Access). Alternatively the system also has the possibility to read dimension information from text based files where a data production's dimension structure cannot be accessed via an ODBC connection.
- **Remote Interaction** – MetaStore is designed so that the functionality provided by the web interface uses components that can also be called from remote ODBC compliant applications. The web interface is also designed with dataset and dimension details (inherited from the data production system) embedded into key page URLs. This facilitates integration of remote applications with the actual web interface.
- **Bulk Processing** – MetaStore also allows for the bulk copying, moving and deleting of metadata content within a single dataset. Flexibility provided in bulk processing criteria includes wildcarding for both source and destination coordinates, and limiting to subsets of common metadata item types.

V. GOVERNANCE PRINCIPLES

25. The basic rules governing the management of metadata and the implementation of MetaStore stress the local ownership and responsibility for metadata. It is the unit responsible for collecting and managing the data – the data provider – who must also take care of metadata.

26. The use of MetaStore as the tool for managing the metadata is highly recommended in the metadata principles but is by no means mandatory. When populating the data warehouse OECD.Stat, the data provider

has to provide the metadata in the prescribed form (attachment levels, metadata items). The data provider may decide to transmit metadata directly from the proprietary production environment or from other documents to OECD.Stat. This means that in order to populate MetaStore, we are using carrots rather than sticks. Data providers must be persuaded to follow this route by attractive systems and good results. Therefore, demonstration of successful implementation of metadata for certain datasets is very important.

27. While it is evident that there is a considerable investment in populating MetaStore and organising the metadata in a new way, it must be demonstrated that, eventually, the management of the metadata can be more efficient with the facilities offered by the corporate system. Another important driver is the possibility of producing a better quality of metadata, leading to satisfied users and maybe a decrease in the efforts to support users who do not understand the data. One of the quality dimensions which can be greatly improved is the coherence of metadata between different databases. Finally, data providers should be interested in increased visibility on the Internet, see section VIII.C.

VI. POPULATING METASTORE

28. The MetaStore system is designed to facilitate populating and updating metadata content from production systems. In order to satisfy different user requirements, metadata can be either updated automatically from data production systems or entered manually through a web based interface.

29. The remote update of metadata through a parameter based procedure serves also to act as the agent for content entered manually through the interface. One of the more important parameters of this procedure for remote access is the sharing level. When metadata content is sent to the procedure one of three sharing levels can be initiated:

0 = No Sharing. There is no check to see if the exact text already exists in the system.

1 = Sharing Within Dataset. A check is made to see if the exact text already exists within the same dataset. The text is shared if a match is found.

2 = Sharing Across Datasets. A check is made to see if the exact text already exists within any dataset. The text is inherited if a match is found in another dataset and shared if found within the same dataset.

30. The most prevalent sharing level chosen when updating metadata from remote applications is 1 (within the same dataset). This ensures sharing of metadata texts while maintaining ownership of the content within the dataset. Such functionality also serves to automatically “clean up” repetitions within legacy metadata systems when migrating metadata to the new system. An Excel based system has been created to facilitate migration from either an Excel worksheet or delimited text file extracted from the legacy metadata system.

31. The web interface has also been constructed using a system of “friendly URLs” with the dataset identifier and data attachment coordinates built in. This enables a straight forward approach to integrating data production applications with the web interface. In practice, as users select data coordinates within the data production system, a URL can simply be built using the same coordinates to direct the user to manage metadata content at the correct level within MetaStore’s web interface.

32. The input fields of MetaStore’s web interface consist of rich text WYSIWYG (What you see is what you get) editors. These allow a small set of formatting commands including bold, underline, italic, bullets, and tables. The editors utilise a comprehensive set of client side code to automatically transform formatted content into valid XHTML. Such a “clean up” transformation is also launched upon pasting content from external sources such as Microsoft Office documents and HTML sources. This enables content to be migrated efficiently from offline sources. XHTML was chosen as a rich storage format so that it can also be rendered into XML for publishing purposes.

VII. SHOWING METADATA TO USERS

33. The metadata is always presented, in whole or partly, along with the statistical data itself through all the different media used for dissemination. Thus, all publications and off-line electronic media are provided with some edited form of the metadata.

34. The way in which metadata are presented on-line is crucial to the usefulness of OECD statistics. There are basically two ways in which users inside and outside the organisation get access to the metadata: 1. stand-alone metadata that users may want to search in order to learn about potentially interesting data, and 2. along with the statistical data.

35. Stand-alone presentations of the full metadata of many datasets will be made available on the Statistics Portal on the Internet in the second quarter of 2006. This will make it searchable and increase visibility of the data behind them (see section VIII.C below). Already today, metadata for Main Economic Indicators have been made available through a special system on <http://stats.oecd.org/mei/default.asp?lang=e&subject=15>.

36. Together with the data, metadata is presented in the OECD.Stat Browser. Here the attachment levels will be reflected, so that metadata are shown at the level of detail where they belong. The following principles have been elaborated to present metadata in a way that will be immediately understandable to a wide audience.

37. In the OECD.Stat Browser, metadata availability is marked in the table view always with a red "i" icon ("i" stands for "information"), clicking this icon will reveal the metadata.

38. Besides metadata at the dataset and the dimension level, there exist three different metadata types:

1. Metadata for single dimension members (most often "definitions", e.g. Population coverage, Key statistical concepts used),
2. Metadata for particular observation values (for any complete combinations of dimension members; most often "exceptions") and
3. Metadata at higher-levels (for any incomplete combinations of dimension members; most often "exceptions").

39. These types will have the metadata availability mark in different places:

1. Metadata for single dimension members have a red "i" in the cell of the dimension member.
2. Metadata for particular observation values have a red "i" in the cell of the value
3. Metadata at higher levels will be marked in the following way (see figure 2 below): In order to avoid having a red "i" in all cells of a row when metadata pertain to all observations in that row, an extra column is introduced in the table view, containing a red "i" when there is a piece of metadata pertaining to all observations in the corresponding row. Reciprocally, an extra row is introduced, containing a red "i" when there is metadata pertaining to all observations in the corresponding column. These extra column and extra row will always be displayed independently on the concrete presence of such metadata.

40. An attribute "IsInheritable" to be set by the data provider is added at Dimension level. For hierarchical dimensions, when set to "true", the presence of metadata at parent level is indicated also at all child levels, and this applies for any of the above three types of metadata.

41. A red "i": will also be shown, when relevant, in the other windows of the OECD.Stat Browser: theme/dataset selector, dimension selector and dimension member selector.

Figure 2: Extra column and row indicating higher-level metadata

Dataset: 1--Gross domestic product

Country		United States				
Measure		C: National currency, current prices, millions				
Frequency		Annual				
Time		2000	2001	2002	2003	2004
Transaction						
B1G: Gross value added; total activity		i 9 100 200	9 402 600	9 710 400	10 200 000	..
B1G: Gross value added; total activity	B1GA_B: Agriculture; hunting and forestry; fishing	i 112 100	110 600	100 500	121 300	..
	B1GC_E: Industry; including energy	i 1 767 200	1 697 500	1 684 800	1 776 200	..
	B1GF: Construction	i 430 900	464 300	473 400	495 100	..
	B1GG_I: Wholesale and retail trade; repairs; hotels and restaurants; transport	i 1 795 400	1 851 300	1 924 700	1 998 300	..
	B1GJ_K: Financial intermediation; real estate; renting and business activities	i 2 879 100	3 023 600	3 121 700	3 268 500	..
	B1GL_P: Other service activities	i 2 115 500	2 255 400	2 405 300	2 540 500	..

VIII. LESSONS LEARNT

A. Drivers to acceptance of MetaStore

42. MetaStore was developed to be sufficiently flexible allowing users to manage their metadata in the same way as they would within their existing system. This was done to minimise the cost for users to migrate metadata and associated management to the new system. Coupled with a range of value added features, the architecture of MetaStore has resulted in a broad acceptance and willingness to use the system.

43. Given the decentralised nature of statistics at the OECD, it was important to offer and clearly communicate the efficiency gains and quality benefits for statistical activity owners to choose MetaStore as a metadata management system.

B. Coherence and Sufficiency of metadata in MetaStore and OECD.Stat

44. The extent to which users have utilised the new features of MetaStore has varied across datasets; however the quality of metadata has significantly improved overall for datasets that have moved to the new system. The most significant improvements, as recommended within the metadata management guidelines, fall into three main areas.

1. Increasing the volume of metadata to make it more comprehensive, sufficient and relevant for users of the data to which it is attached.
2. Splitting metadata content and attaching it to more accurate and relevant data coordinates that was not possible in previous metadata management systems.
3. Increasing coherency by standardising of metadata via sharing of content both with and across datasets as well as employing a common classification of metadata concepts.

45. MetaStore manages metadata for about 145 datasets out 240 within OECD.Stat.

C. Effects on visibility

46. The growing trend to freely disseminate statistical data online presupposes that it is sufficiently visible to be accessed by interested users. The accessibility of statistical data on the web by interested users depends largely on the ability of leading search engines to index relevant metadata associated with the data. The visibility of statistical data on the web is therefore highly dependent on the content, structure, and availability of the reference metadata that describes it.

47. There are a number of important factors that search engines take into account to rank web content in relation to search phrases entered by users. The MetaStore metadata management system has been developed and optimised for these factors so that the visibility of the resulting online data content is maximised. Such strategies required for optimising data visibility online also ensures that the content and structure of statistical reference metadata serves to enhance understanding of the data by end users.

IX. PLANS FOR THE FUTURE

48. As users continue to utilise the MetaStore system and seek ways to increase the quality of reference metadata, the value added features of the system will be employed more extensively.

49. So far, users have been able to use MetaStore to mimic, to some degree, the metadata management behaviour employed within older systems. Moreover, there is no official requirement for statistical activity owners to migrate metadata and management to the new system. As a critical mass of statistical activities migrate to MetaStore, more pressure will be put on remaining statistical activity owners to move to the new system. In parallel with this, more rigid rules will be put in place within the system to further enhance the standardisation and coherence of reference metadata across datasets. Implementation of more rigid metadata management rules in the future are likely to include:

- Enforcing the reuse of exact text matches within the same dataset (currently an option).
- Enforcing content to be filled for certain combinations of attachment levels and metadata concept (for higher levels of a dataset's structure in particular).
- Undertaking metadata quality reviews by dataset and proposing resulting actions to be taken by dataset owners.

**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

DEVELOPMENT OF A METADATA SYSTEM AT THE CROATIAN BUREAU OF STATISTICS

Supporting Paper

Submitted by Central Bureau of Statistics of the Republic of Croatia¹

I. INTRODUCTION

1. The project to develop a central metadata repository and public macro database, in the frame of Integrated Statistical Information System (ISIS), started at the Croatian Bureau of Statistics (CBS) in 2002. The primary target of ISIS is to manage a completely metadata-driven automated processing system, which requires a well-structured metadata system. The outcome of the project, which ended in 2005, was a new metadata model called CROMETA #2. The model is based on the METANET Reference ModelTM and customized for needs of CBS. The new model includes specifics of a previous version of metadata system at CBS, as well as concepts necessary to run PC-Axis as a main dissemination tool and to connect with BRIDGE^{NA} as a tool for maintaining classifications.

2. The model has been physically implemented on relational database management system and accompanied with an application for browsing and maintaining metadata. The development is ongoing, and both the metadata model and the maintenance tool are in the phase of testing. Physical implementation of the metadata model doesn't cover the whole model, just a part of it. The plan is to include all metadata needed for successful creation of so-called 'generators' – programs for building applications for data processing. The final metadata system solution should be presented in 2007.

II. HISTORY OF METADATA IN CENTRAL BUREAU OF STATISTICS

3. Central Bureau of Statistics of Republic of Croatia started to develop the metadata system on the client/server platform a few years ago. The start was actually in 2000, when the first version of metadatabase was designed and implemented. The main reason for the development was to create a completely metadata-driven automated processing system. Production of statistics in CBS was carried out on a mainframe and just a small part was developed for a client/server environment. Data processing in a mainframe environment was solved quite well with so-called 'generators', the applications for generating programs for data processing. This technique covers large part of all surveys that are carried out on mainframe, since 80% of them are very similar. The growing costs of maintenance of two platforms forced CBS to decide to transfer production to a

¹ Prepared by Maja Ledic Blazevic majalb@dzs.hr

client/server environment and use the same old idea of 'generators'. The precondition for realisation of that plan was to have well-structured metadata.

4. The other reason of same importance for the development of a metadata system was to harmonize statistical methods and data according to international standards, and to have compatible, or at least comparable data with international statistical organisations.

5. The first metadata repository, named *Inventory*, was developed based on a custom data model created in CBS. The implementation was carried out on a relational DBMS, MS SQL Server, as it was a pre-existing environment for database development at CBS. This repository was initially filled with collected metadata from a few selected surveys, mainly in the area of conceptual and operational metadata. There were problems in metadata collection because subject matter staff were not accustomed to thinking conceptually instead of considering contents of data. Other sources of metadata were also included in this inventory of metadata in CBS, for example, from the publishing department and Program of Statistical Activities for CBS that is approved by Parliament.

6. The idea of building 'generators' based on that first metadata system was not implemented because in the beginning of 2002 a joint project of CBS and Statistics Sweden (SCB) was launched. This project was sponsored by Swedish Agency for International Development Cooperation (SIDA) and supported by many consultants hired by Statistics Sweden. The primary goal of the project was to develop a central metadata repository and a public macro database. Early stages of the project were more dedicated to public macro database development. At that time, 2002-2003, a decision was made to use PC-Axis software as the dissemination tool at CBS. It was first offered as the dissemination tool with data from the Agriculture Census 2003 and was very well accepted by statisticians and external users.

7. *PC-Axis* has its own metadata with a specific model behind it. Data was collected from some statistical surveys and added to the new statistical warehouse. Cubes of aggregated data were defined based on that micro data and described by metadata. Then *PC-Axis* could be used to produce statistical tables based on a user's selection of themes, matrices and variables. The metadata model of *PC-Axis* covered different areas of metadata to *Inventory*'s model, and didn't fulfil the needs of CBS. Another solution was needed. Data and metadata prepared for *PC-Axis* was sufficient to develop a customized prototype of metadata-driven web solution for data dissemination in 2004.

8. In 2002 *BRIDGE^{NA}* was introduced at CBS as the metadata maintenance tool. After testing how much it covers CBS needs, only module *ClassE* was chosen to be used as a maintenance tool for classifications. It is still in use today. Behind *ClassE* lays the Terminology Model, a very developed model of metadata dealing with classifications.

9. In 2004, a decision was made to start development of a custom metadata model, based on Reference Model TM that was developed during the METANET project of Eurostat (2000-2003). Reference Model TM includes already the above-mentioned Terminology Model, as well other models of statistical metadata, and it is their common denominator.

10. In 2004-2005 the focus of development was set on building a central metadata repository rather than a public macro database. At that time the project was divided in two subprojects: metadata methodology development; and technical implementation. The final outcome of this stage, and the entire project, is the CROMETA #2 metadata model and its implementation.

III. METADATA METHODOLOGY

11. As noted before, a new metadata model was based on Reference Model TM, but existing metadata consumers, *Inventory*, *PC-Axis* and *BRIDGE^{NA}*, also had to be included. There was no need to work on *BRIDGE^{NA}*, but other metadata had to be mapped to the Reference Model TM. Thorough analysis provided specifications on how to integrate specifics of *Inventory* and *PC-Axis* Macrometa into the new model.

12. A group responsible for metadata methodology, consisting of both statisticians and information technology experts, worked in a technical implementation group on specifications for how to build metadata. These specifications were created on the basis of Reference Model TM analysis. A specification consists of a concept, its characteristics and references to other concepts. The extensions of Reference Model TM were created in the form of additional characteristics of existing concepts when possible, and additional new concepts and their references when needed. It is important to stress that not many differences were found between metadata models, and not many extensions were added.

13. The Reference Model TM was hard to understand due its complexity, especially for non-methodologists in the technical implementation group. Therefore, a document was created about the whole model and metadata process, important in many ways. The main idea of the document was to recognize basic metadata concepts and their relationships. It provided simple vocabulary for non-experts and examples to understand the concepts. It contains procedures how to process metadata and in which order. The rules and behaviour of concepts and characteristics were also described for that purpose.

Examples from the document about metadata process:

*All objects that are observed at CBS are defined as **Statistical object types**:*

- *Person*
- *Household*
- *Enterprise*
- *Commodity*

A statistical object type may further be divided hierarchically into sub-types, as far as needed.

- *Person:*
 - *Employed person:*
 - *Full-time employed person*
 - *Part-time employed person*
 - *Unemployed person*

*All objects that are generally measured at CBS are defined as **Global variables**:*

- *Income*
- *Age*
- *Activity*
- *Occupation*

*When defining what characteristics to observe for a certain object, statistical object types are combined with global variables to form **Object variables**.*

- *Income of person*
- *Age of person*
- *Income of enterprise*
- *Activity of enterprise*

14. When developing a methodology for CROMETA #2, metadata concepts were categorized into nine metadata sections or areas according to the logical relationships between them. These sections can be described as follows:

a. Studies and questionnaires

This section contains metadata regarding the various studies/surveys carried out by the statistical office, for example, all metadata regarding studies, versions of them, general method for performing them, etc. Examples of metadata concepts in this section are Study, Study Version, Questionnaire, Question, Interview method, Population, Coverage type, etc.

b. Variables and measurements

This section contains metadata about variables collected within the frame of the statistical activity, as well as the methods and ways of measuring them. In this section variables are described from a

more general point of view, regardless of use and implementation in different studies. Examples of metadata concepts belonging to this section are Global variable, Object variable, Measure unit, Basic measure unit, etc.

c. Processing and validation rules

This metadata section could be described as a combination of variables and measurements versus studies and questionnaire sections, meaning that variables and measurements are put in the context of studies/surveys. Here are metadata regarding processing of studies, including validation, production processes, registers, cubes and tables created. Examples of metadata belonging to this section are Context data element, Data collection, Derivation rule, Register, Cube, and Table, etc.

d. Classifications

This section contains all metadata concerning classifications. Examples of metadata concepts in this section are Classification family, Classification, Classification version, Classification item, Correspondence table, etc.

e. Publications

This section contains metadata concerning publication and dissemination of statistics. Examples of metadata concepts belonging to this section are Publication series, Publication, Edition, etc.

f. Organisational structure

This section contains all metadata concerning the organisation, related organisations, persons working within the organisation and the responsibility of the latter, etc. Examples of metadata concepts belonging to this section are Organisation, Organisation scheme, Person, Organisation level, etc.

g. History and version handling

Metadata objects as instances of metadata concepts may exist in an indefinite number of versions. Version management is extremely important in order to keep consistency in metadata and data over time. The main part of the history and version handling is implemented through methods applied on metadata objects. Every metadata object has several versions valid in some period of time and is described with a certain status. An example of metadata belonging to this section is Update information, used for logging all changes made to a metadata object over time.

h. General metadata concepts

All metadata contains some general characteristics, and they are sometimes metadata concepts themselves. All these general metadata concepts are kept within this section. Also, there are metadata concepts used in all, or at least several sections, and they too are placed in this section. Examples of metadata belonging to this section are Language, Keyword, Footnote, Theme, Statistical object type, Status, etc.

i. Access and authorization rules

This section contains metadata about how metadata and data could be accessed by internal and external users and the authorisation rules applied to metadata and data. Examples of metadata belonging to this section are User group, Privilege, Access condition, Access form, etc.

15. It is important to stress that metadata concepts in CROMETA #2 can also be grouped by any other type of categorisation, e.g. according to the approach of the OECD, it is possible to categorize metadata by usage (1).

16. The priorities in implementation are also defined from the document about metadata process. It was not intended to implement everything from the CROMETA #2 model, but the concepts that are first on the priority list. It was decided to work first on general metadata, history and version handling, access and authorization rules, organisational structure and variables and measurements. After those sections, focus went on studies and questionnaires and publications. Since classifications are handled by BRIDGE^{NA}, they are just referenced from CROMETA #2 for value domain definitions. Processing and validation rules section is planned for

development in 2006 because of its importance for subsequent project of 'generators', which are tightly connected to this section.

17. The document about metadata process provided identification of starting points of metadata collection. It was concluded that metadata can be independently collected from different sources, and collection was carried out in that way. Metadata was collected from *Inventory*, Program of Statistical Activities of CBS, *PC-Axis* Macrometa and few appointed surveys. Finally they were successfully transferred to CROMETA #2.

18. Metadata methodology development was a very important part of the work done, because it is a necessary precondition for successful work on technical implementation. Knowledge about the metadata model must be at a high level to create visible results. All participants working with metadata systems must have clear understanding of metadata and related concepts. The experience from *Inventory* shows that statisticians who tried to enter metadata had problems in that area, for example statistical objects and variables were mixed up.

IV. SOME SPECIFIC METADATA CONCEPTS OF CROMETA #2

19. As mentioned before, there were some extensions of the Reference Model TM in CROMETA #2. Some of them will be described below. They are defined as totally new concepts resulting from mapping work, e.g. Edition of publication or Geographic collection level. Alternatively, as new characteristics of existing concepts from Reference Model TM, e.g. sequence number of a theme needed for specific predefined order of themes presentation, different from alphabetical order.

A. Hierarchical structures

20. The organisational structure of a statistical office can be changed over time. New departments can be established, or moved to another statistical division. Sometimes two departments are merged. A new concept of Organisation scheme (or structure) is therefore added to the metadata model. It represents the hierarchy of the organisations in some period of time. Relationships like parent and child organisations exist in the frame of this object and they are not directly connected to the organisation.

21. In the same way, the themes of statistics that are processed in surveys also create a hierarchy. This is a separate metadata object in CROMETA #2 called Theme structure. Even more, it is possible to define two different theme structures at the same time. In real life, for example, the theme structure used at CBS differs from the theme structure at EU, and some studies carried out in the EU are still not yet established in CBS. Naturally there are themes then without studies belonging in this so-called 'EU structure' but all the studies are connected to a theme belonging to 'CBS structure'.

22. Similar to these ideas, questions in questionnaires are handled in the hierarchy separately from the concepts of question and questionnaire. This differs to Reference Model TM where 'question' is connected to a set of 'sub-questions' in the concept itself. In CROMETA #2 model the hierarchy of 'main' and 'sub-questions' is moved from the 'question' concept. This allows reuse and unification of questions and could improve the layout of questionnaires because very often the questions of same meaning are asked in many different ways.

B. Status and version handling

23. It is very important to explain history and version handling. Each metadata concept can have several versions that are valid in exclusive periods of time. At any one moment, one version is currently valid. This is determined with the Status of the metadata object. It is defined that there are six possible states of metadata. They are:

- a. Under development
- b. Released
- c. Authorized
- d. Archived
- e. Frozen
- f. Deleted

24. By default status is set to 'Under development', but it can be immediately set to 'Released' or even to 'Authorized' if the person who adds the object has sufficient privileges. There is one and only one version at a time that has status 'Authorized'. This is the currently valid version. 'Released' means only that work on metadata object is done and waits for approval. Once when metadata object gets status 'Authorized', it enters the queue of versions that have top (or first) version and a strong relation between themselves, meaning that they have a known predecessor and successor version. After a version with 'Authorized' status is replaced with new version with 'Authorized' status, it automatically gets its status changed to 'Archived'. It then becomes the predecessor version of the new one, and the new one becomes its successor version.

25. Status of an archived version can be changed to 'Frozen' and this protects it from any changes in the future. Status 'Deleted' is for keeping information about metadata that are deleted from some reason. Metadata in deleted state can be restored back by an administrator of the system. Metadata can be physically deleted only when they have status 'Under development'. Whenever some significant change on metadata must be applied, it is recommended to create a new version. If changes are small, for example, spelling corrections, they can be done without defining new version. Such changes are kept in the Update information, a general property of all metadata concepts in CROMETA #2.

C. General properties

26. General properties of metadata concepts are defined in a specific way in CROMETA #2. They are either generated from relationships with general metadata concepts, or they are attributes of the object. General metadata concepts, such as Language, Keyword, Footnote, Synonym, Milestone, Status, etc., can be applied to all metadata. Some of them are mandatory, like Status, but the majority are not. They can be freely used whenever needed for a better description of some object. Some relationships of metadata objects and general objects can be of type 1: n, like Keyword, Footnote, Document, Registered users, Contact persons, User groups, etc.

27. Each metadata object can have title and description in an unlimited number of languages. For any new language it is enough to add it to metadata system and then add language dependent characteristics in the newly added language. There could also be several types of title for each metadata concept, for example official, short, alternative, etc. This gives the opportunity to present metadata in different forms, for example different length of column and row headers in tables.

D. Meta-metadata

28. In the Access and authorization rules section, User group types are defined and they are connected to Privileges defined in the same section. User group of particular type inherits privileges from User group type. In practice, anyone has the right to browse or read data, expert user groups can add and edit their metadata, while only the administrators have the right to delete them as well.

29. Privilege is an example of many meta-metadata defined in CROMETA #2. Meta-metadata provides common vocabulary and whenever some property of metadata concept could be categorized by some definition, it was implemented as meta-metadata. Properties described by meta-metadata can be validated during the capture of metadata and this makes the overall solution more generic. Meta-metadata can also be

maintained through the common interface. Examples of meta-metadata in CROMETA #2 are Document type (legal base, methodology guidelines etc.), File format, Organisation level, Graduation, Publishing type, Variable type, Interview method, Sample method, Matrix type, etc.

E. Metadata sources

30. Regarding classifications, CBS is using the ClassE module from BRIDGE^{NA} software as a tool for classification maintenance. When classifications development started in CBS, thorough studies of the Neuchâtel terminology model was initiated. This model was selected since it is both a terminology and conceptual model, and that makes it generally applicable and independent from the software platform. Successful tests of including classifications in CROMETA #2 have been carried out. The integration of the classifications section will enable full advantage to be taken of all the general solutions relating to classifications. For the moment, reference is made from CROMETA #2 to BRIDGE^{NA} classification database when value domains are defined based on classifications.

31. As mentioned before, metadata has been collected from several sources and this process continues. There were five appointed surveys chosen as the metadata providers, but at the moment not all of them are used. Whatever could be used from the previous metadata system and other sources, has been already transferred to CROMETA #2. After the presentation in November 2005, it was decided to add more concepts necessary to compare CBS statistics with Eurostat, through the Compliance Database. The previously mentioned theme structure allows creation of another structure according to the Statistical Requirement Compendium, and measurement of the compatibility of variables used in CBS's studies to the ones defined in EU statistics. That proves that the model is flexible enough to be adapted when necessary.

V. TECHNICAL IMPLEMENTATION

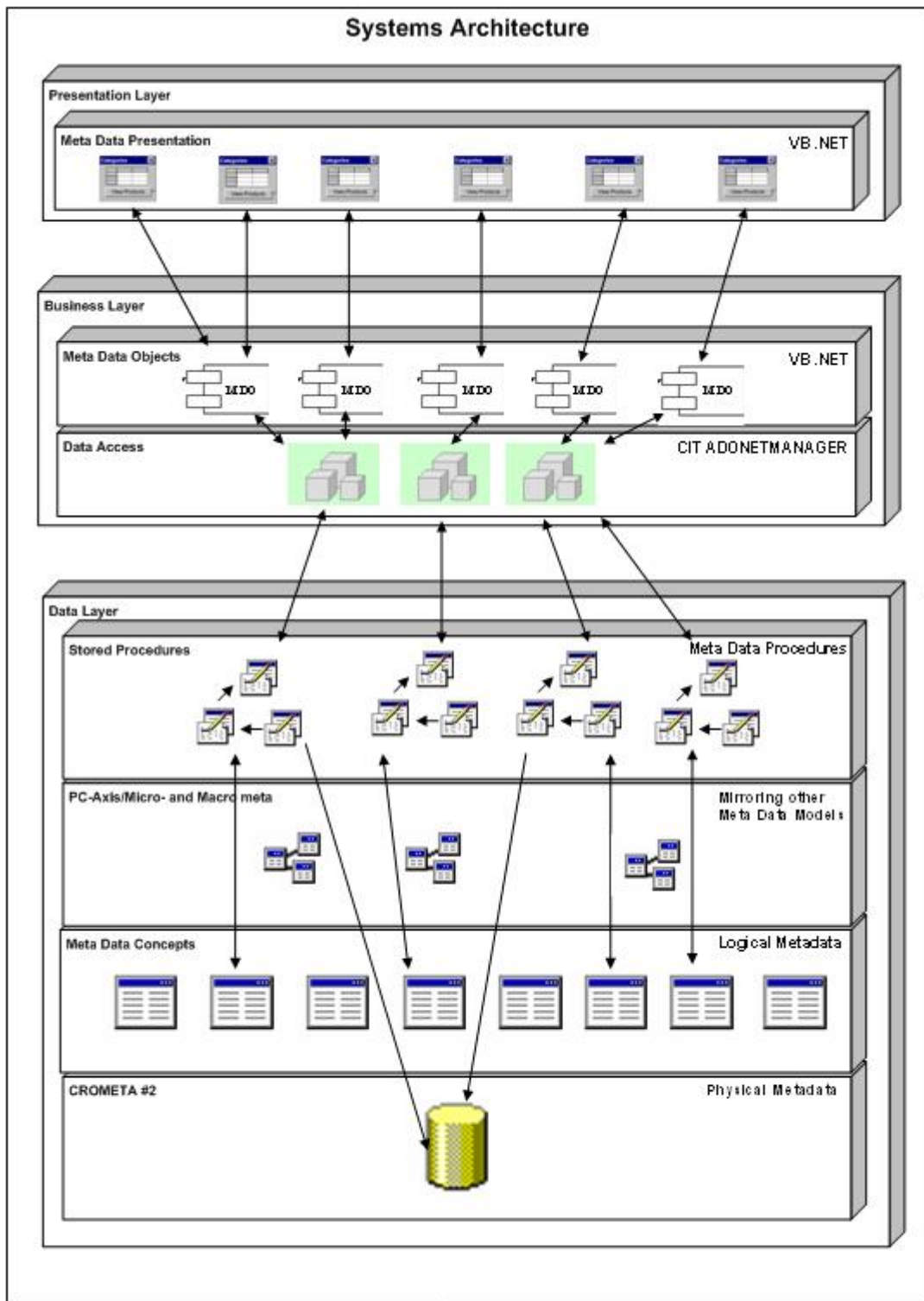
32. This section is about technical implementation because it not much is visible of the metadata model without physical implementation and an appropriate tool for browsing and maintaining metadata. The development tools used for conceptual design of metadatabase was Sybase Power Designer 9.1. The same software was used to create the physical data model that was then generated on the Microsoft SQL Server 2000 RDBMS1. All database development was carried out on the Microsoft SQL Server platform2, while CROMETA #2 maintenance tool has been developed in VB .NET3. Applications were running on Windows 2003 servers on presentation in Zagreb in November 2005. Actually, since development is still ongoing, the solution is already moved to VB .NET5 and tested on Microsoft SQL Server 2005.

33. The technical architecture follows common techniques for modern system development, fronting a multi-tiered, scalable solution, customized for a multi-user environment. It consists of three main layers (see Picture 1)

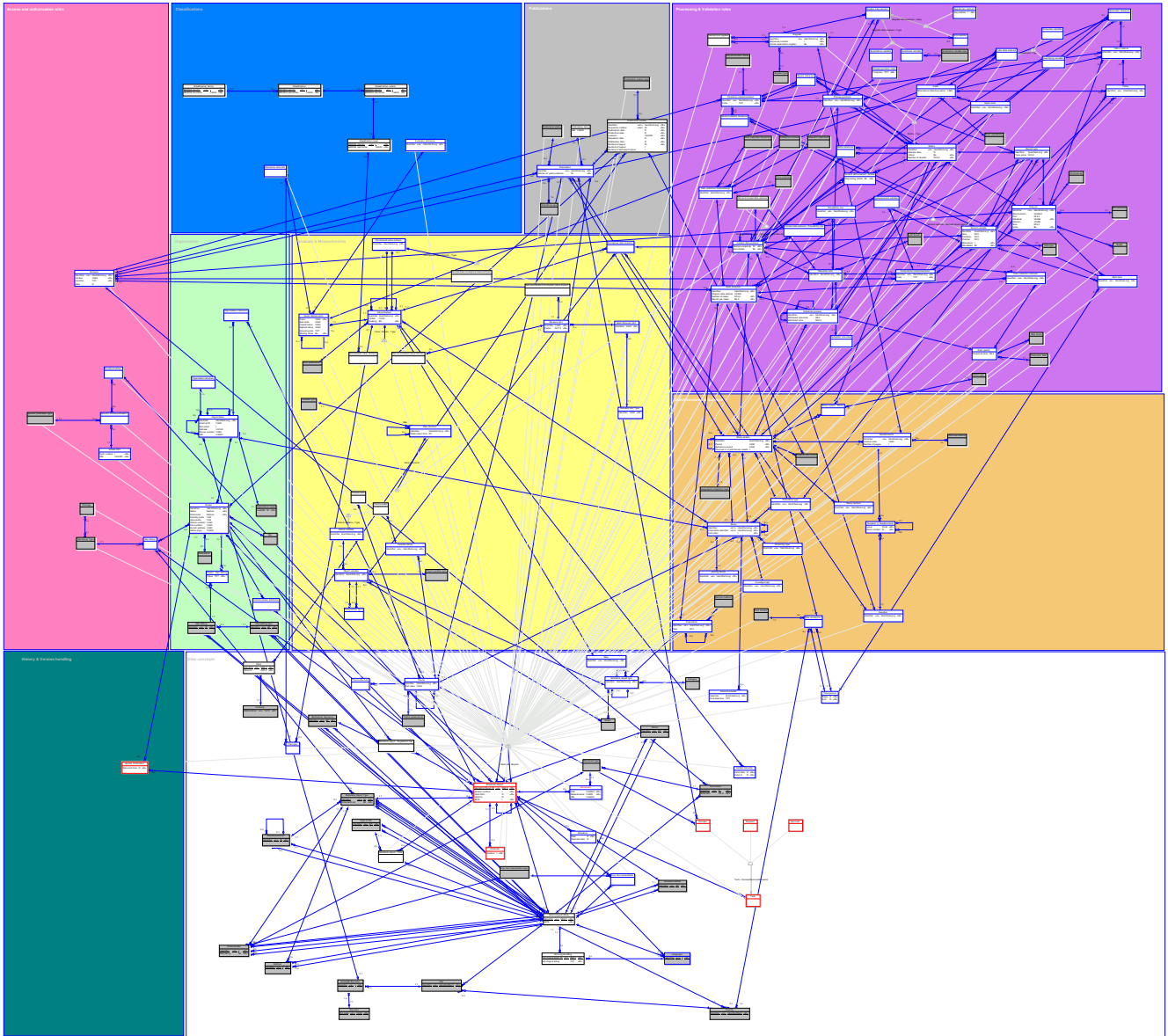
- a. Data layer
- b. Business layer
- c. Presentation layer

34. The Data layer consists of four tiers. The physical storage tier is a normalised data model in 5th normal form (see Picture 2). It has more than 170 concepts when looking at the conceptual model and more than 230 tables in the physical database, connected through numerous relations. For faster retrieval of data, to avoid multiple joins, the logical exposition tier is defined and it is actually a denormalised data model. There is also an alternative view tier, used for mirroring additional models from other metadata consumers (like PC-Axis Macrometa), and at the end the access tier in the form of stored procedures used for managing data.

35. The Business layer is managing the business logic of the metadata solution and consists of three tiers. The lowest part is the data access tier, managing all query execution. On top of that is data communication tier responsible for receiving, formatting and forwarding requests from the top-situated business tier. The latter logically reflects the conceptual metadata model and implements metadata concepts through classes according to an object-oriented approach. This tier provides possibility of accessing metadata from other applications and it is the integration point for metadata consumers of the ISIS.



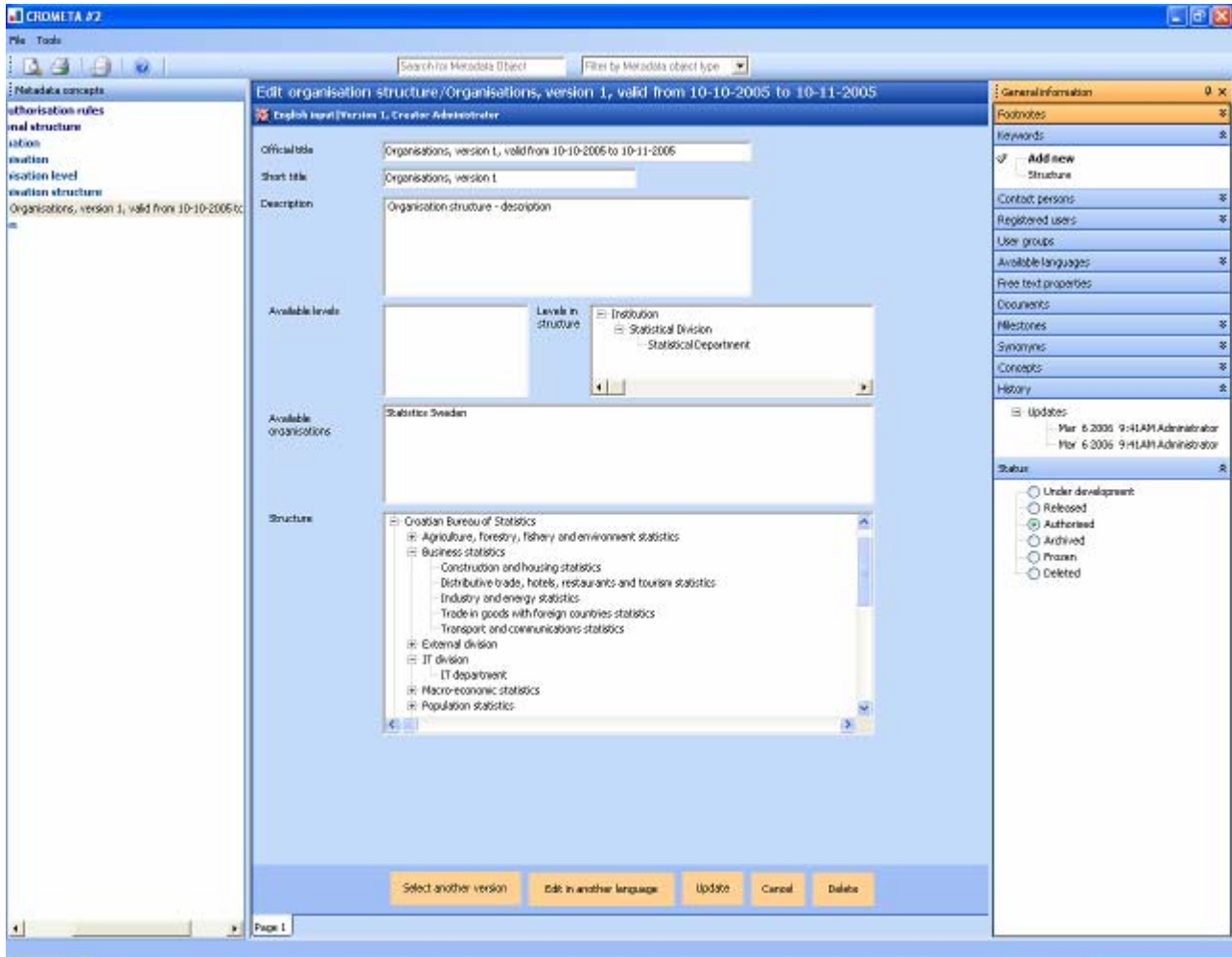
Picture 1. Systems architecture of implementation of CROMETA #2



Picture 2. Physical model of CROMETA #2

36. The presentation layer is actually the CROMETA #2 administrative and maintenance tool. It is a single user interface for all users to enter, administer or maintain metadata, when browsing or searching the metadata repository. This is a thin client solution. All processing and execution is carried out on the application and database servers.

37. The maintenance tool has been developed as a user-friendly interface to the central metadata repository. The development of the tool is ongoing and whatever could help users to navigate better through metadata is either integrated or planned to be. The application presents metadata grouped in sections and listed alphabetically. At the moment alternative views and wizards are developed for non-experts, to guide them through metadata definitions in the right order according to the metadata process. All metadata concepts can be added in many languages and the user interface itself is multilingual; default language is set according to user's choice. General methods for adding, editing and deleting data are established for all metadata concepts. All general properties are easily retrieved and changed. It is possible to search based on titles of metadata objects and by types of metadata objects, but this quick search will be replaced with the advanced one in the next versions. Users have the possibility to subscribe to metadata objects, for example, to select an object and specify the frequency of e-mail notifications about any changes on metadata object they have chosen.



Picture 3. Maintenance tool of CROMETA #2

VI. CONCLUSION

38. The development of a metadata system is a long-term process. The experience of CBS is that a metadata methodology must be strictly defined and well developed before going into technical implementation, but some questions can be raised afterwards that cause adaptations in the methodology. System development is an iterative process and many tests remain to be done before it can be put in regular production. The attempt of implementing the Reference Model TM resulted in establishment of the first versions of a central metadata repository in CROMETA #2 and respective maintenance tool. On the basis of the feedback from statisticians who will provide metadata, the maintenance tool will be adjusted and further developed.

39. It is the explicit target of CBS to have the first surveys processed in the automated system in 2007. That means a lot of work on the metadata system and 'generators' during 2006. The prototype of the new website integrating macro and metadata is already tested and now the bulk of data from the mainframe must be transferred to the statistical warehouse. In the future statisticians should eventually have all the tools they need for complete metadata and data maintenance and dissemination. Before that it will be necessary to include and motivate statisticians even more to accept the metadata concept and pass all their knowledge to CROMETA #2. Any misunderstanding of metadata concepts could lead to meaningless statistical data.

VII. SOURCES

- (1) CBS/SCB-SIDA Project, Presentation Zagreb, Croatia, November 30th 2005, Development of a Central Metadata Repository and a Public Macro Database at the Central Bureau of Statistics of Croatia

**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

Using SDMX standards for rapid dissemination of short-term indicators on the European economy

Supporting Paper

Submitted by Eurostat¹

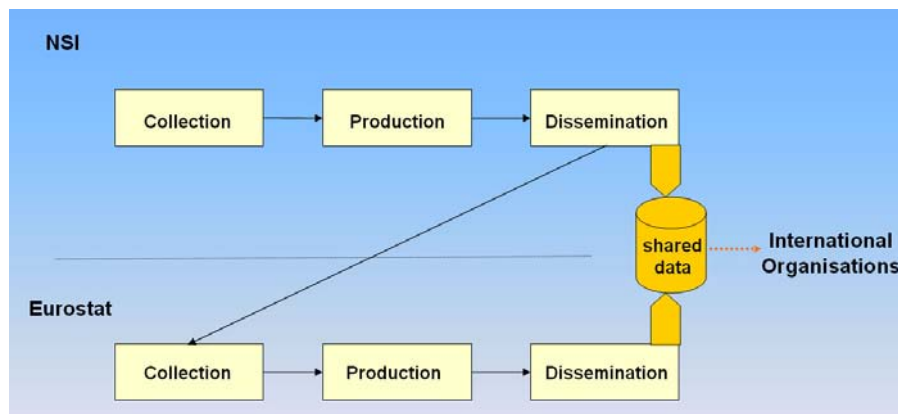
I. BACKGROUND INFORMATION

A. The Original Ideas behind SODI (SDMX Open Data Interchange)

1. The SODI project focuses on the interoperability of statistics for collecting and disseminating short-term statistics, especially in the domains of the Principal European Economic Indicators (PEEI), with the overall objective of increasing timeliness and accessibility. SODI is an SDMX implementation project, which means that it is one of the official proofs of concept for SDMX. It is comparable in scope to the "National Accounts World Wide Exchange" project (NAWWE) undertaken by the OECD.
2. The main benefits expected from SODI are:
 - improved quality and timeliness of statistics;
 - reduced reporting burden, through the use of common formats for data exchange and data sharing of statistical information on web sites to complement or replace direct reporting;
 - more user-friendly access to data and related metadata, for business users and citizens, when publishing international and national statistics on the web;
 - reduced human resources needed to process the data in the Competent National Authorities (CNA) and Eurostat.
3. SODI is a data sharing project in the European Statistical System. The Data Sharing Model is a mechanism by which data (e.g. statistics) are made available to users in a common environment (the Internet) in a common technical format and with agreed common codes and metadata. In this model, users locate and retrieve the data relevant to their needs using a registry made available to them by those partners participating in the data sharing exercise. The model, currently being tested for the dissemination of the Principal European Economic Indicators, is being tested on other domains.

¹ Prepared by B. A. Lindblad, L. Maqua, M. Pellegrino and G. Sindoni - Eurostat.

4. The picture below shows the basic idea of data sharing. In the SODI project, the shared data are maintained by Eurostat.



B. The SODI Pilot Project

5. Before SODI was launched, a pilot project has been conducted. A Task Force with Eurostat and the national statistical institutes of Germany, France, the Netherlands, Sweden and the UK performed trials, testing different ways of transmission (Push – the traditional method of sending data to Eurostat, and Pull – the data sharing approach via web services) and different data formats (SDMX-ML – the XML version of the SDMX data format, and SDMX-EDI, the GESMES-TS compatible version of SDMX). The results of the trials were successful:

Country	Format	Method	Status
DE	SDMX-ML	Pull	successful transmission
FR	SDMX-EDI	Push	successful transmission
NL	SDMX-ML	Pull	successful transmission
SE	SDMX-EDI	Push	successful transmission
UK	SDMX-ML	Push	successful transmission

6. The SODI pilots had two main deliverables: a report on the issues encountered, which was delivered to the FROCH² group in June 2005, and a proof of concept for the data sharing approach, which was delivered, together with live demonstrations, in November 2005 at the FROCH group and the SPC³.

7. The conclusions of the SODI pilots exercise have been drawn as follows:

- The approach of SODI, based on SDMX standards, is technically feasible: work should continue towards the objective of opening SODI to public access in the second half of 2006.
- The SODI pilots have enabled the identification of the issues that must be taken in order to pass to an operational implementation of the SODI concept in terms of widening the range of indicators. In principle, SODI aims to cover all the PEEI. However, experience has shown that for the data-sharing approach to work, certain criteria concerning the choice of indicators have to be fulfilled, such as the availability of a key family fully compatible with the rules used in

² Friends of the Chair, a high level group which acts as a think-tank to the Statistical Programme Committee

³ Statistical Programme Committee, comprising the presidents of the National Statistical Institutes of the European Statistical System.

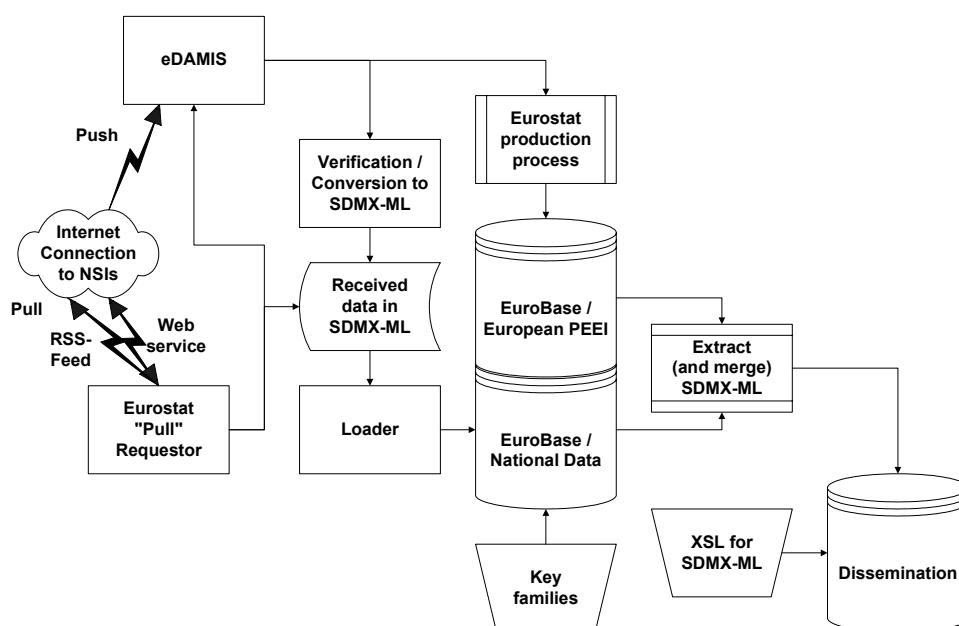
GESMES/TS and SDMX-ML; an acceptable level of data quality; and an acceptable level of harmonisation in the national versions of the indicators.

- The SODI pilots have enabled the identification of the steps that must be taken into account for covering more countries. The pilots have enabled countries to identify the nature of the work required and hence to determine whether costs are manageable. In general it appears that this should be the case.
- The SODI TF should continue work as its input will be needed to deal with several of the issues identified in the Issue Report.

II. THE CONCEPTS AND IMPLEMENTATION OF THE SODI PROJECT

A. The SODI Approach towards Data Sharing and SDMX Implementation

8. The SODI process accepts both SDMX-EDI and SDMX-ML as an input, and can receive data both via eDAMIS⁴ and from a web service set up by the competent national authority on its web site. In the further processing up to dissemination, SODI only uses the SDMX-ML format. So, despite being some type of shortcut to Eurostat's normal data processing cycle, SODI fits into the data life cycle of Eurostat. The processing of data in SODI is synthesised by the following picture:



B. Project Organisation and Budget

9. The project is financed as an action of the X-DIS project (XML for Data Interoperability in Statistics) by the Commission programme IDABC (Interoperable Delivery of European eGovernment Services to public Administrations, Businesses and Citizens) for the budget years 2005 to 2008.

10. The project requires a close co-operation with the Member States, which is coordinated by the SODI task force described in the following paragraph.

⁴ eDAMIS is a system aimed at implementing Eurostat's concept of a "single entry point" for statistical data. That is the hub where data sets should be sent by national competent authorities and delivered to competent Eurostat's production units.

C. The SODI Task Force

11. The SODI task force has been enlarged since the pilots. Now, the National Statistical Institutes of the following countries participate in the task force (with ECB and OECD as observers): Denmark, Germany, France, Italy, the Netherlands, Norway, Slovenia, Sweden and the UK.

12. The members of the SODI task force

- support Eurostat in the SODI implementation;
- are consulted on the SODI work plan and other important documents on SODI and SDMX;
- give advice on SODI issues;
- send data to Eurostat to be used in the SODI process;
- receive technical support by Eurostat and its consultants on SDMX and the implementation of SODI.

III. ISSUES TACKLED BY SODI

13. This paragraph summarises the issues encountered or identified during the SODI pilots. It marks points for decision, threats and opportunities, and gives recommendations and lessons learned, tackling both technical and non-technical issues.

A. Technical issues

SDMX-EDI and SDMX-ML

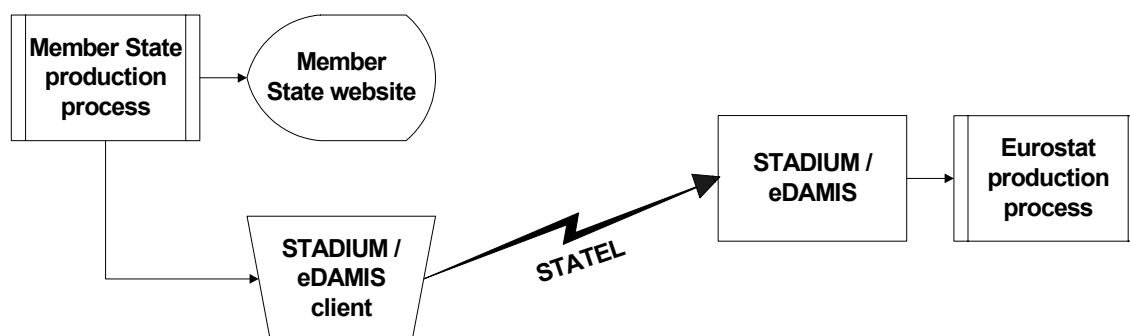
14. One of the main results of the pilots is that the existence of two different data formats does not cause any real problem. The conversion from SDMX-EDI to SDMX-ML within the pilots is being done with a simple "home-made" tool, which shall be replaced in the future by a more appropriate one, based on open source software and including also structural definition maintenance and conversion to other formats.

SDMX-ML for dissemination

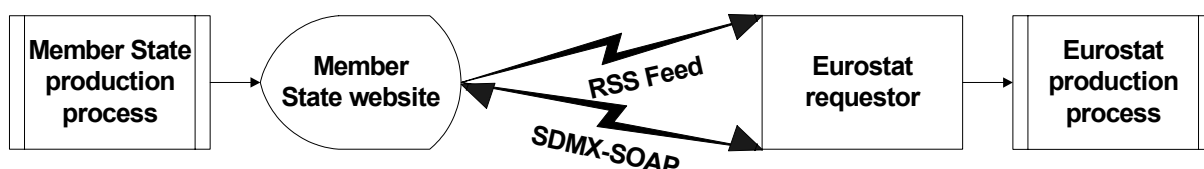
15. The SODI project will test as well the use of SDMX-ML for dissemination. For this, formatting information (in XSL, the eXtensible Stylesheet Language, or CSS, Cascading Style Sheets) will be added to the SDMX-ML data, so that they can be visualised in any modern browser.

Push and Pull

16. The Push method in SODI means that the Member States send their data to Eurostat using the Single Entry Point (SEP) currently implemented using a system named STADIUM or the new eDAMIS system:



17. The Pull method is characterised by the fact, that a Eurostat application (called “requestor”) fetches the data from the web site of a National Authority:



18. The requestor is triggered by an RSS feed. RSS stands for “Really Simple Syndication”, and is the web standard for news feeds. The implementation of the Pull method is currently under construction in both Eurostat and two National Statistical Institutes (Netherlands and Germany).

19. The Pull method is more difficult to implement both at Eurostat and at the NSIs; because of this, it has been the subject of in-depth technical investigations within the SODI pilots. However, this method will finally fit better into a general SDMX-ML dissemination environment in the Member States, so that a seamless integration into the production process is automatically achieved. In addition, the Pull uses only standardised internet compatible methods. The Pull approach guarantees that data are available at the same time for Eurostat as they are published nationally, while it requires that the National Statistical Institutes publish data conforming to the European concepts.

B. Standardisation issues

Statistical terminology

20. The development of more efficient processes for sharing data requires the adoption of a standard terminology for describing the statistics being exchanged. The Metadata Common Vocabulary (MCV) elaborated under the SDMX initiative provides the common set of terms (and related definitions) to be used for the sake of terminological consistency. Agreement on such a standard implies a continuous update to reflect core concepts used within SDMX and with national institutes. SODI, therefore, is one of those initiatives which can provide a valuable “reality check”, through the description of data structures and the attachment of a set of reference metadata documenting the data. Existing ambiguities in the use of a term, or the fact that not all terms have been identified in the MCV yet, call for a parallel expansion of the MCV during 2006.

Structural metadata for “Data Structure Definitions” (Key Families)

21. It is a prerequisite for the full SODI implementation that all key families are SDMX-compliant. This was not the case for the existing structural definition on GDP, which was not compatible with GESMES/TS (and hence with SDMX). GDP data are collected in the European System of Accounts (ESA95) framework. The ESA95 structural definition was recently revised using a SDMX-compliant standard (GESMES/TS which can be regarded as equivalent to SDMX-EDI). A thorough analysis of existing GESMES structural definitions of the PEEI, currently being performed, will allow to define priority domains for SDMX implementation and requirements to migrate from GESMES to SDMX.

Reference metadata

22. In accordance with the general principle that no data should be made available without an acceptable coverage of metadata, Eurostat conducted a review of metadata available from national, European or international web sites. The main issues were: a) the availability of sufficient metadata coverage in multiple languages; b) the conceptual overlap between formats used within national and international web sites; c) issues concerning the updating and dissemination process of the different metadata items.

23. The metadata coverage associated to the first test domains is quite good, although the same kind of information is available from different providers (national web sites, Eurostat, IMF) and not all of the metadata are still available in English from all countries. The need of advancing towards a more coordinated framework which links national and EU metadata has been stressed by several participants in the SODI task-force.

24. Eurostat, in coordination with member States, intends to make use of the latest standards for associating data with a consistent and standardised set of metadata items. In this context, the use of SDMX standards would allow the harmonisation of presentation styles and at the same time the standardisation of data and metadata descriptions, so that these can be exchanged, read and processed by computers without manual intervention.

25. Through the use of standard concepts, applied to the exchange of data sets (on the basis of the formal definition of the data structure) as well as to metadata sets (on the basis of the definition of the metadata structure), there is the concrete possibility of setting the requirements for a European concept family of reference metadata to be exchanged and shared by using web services to navigate, find and process the information.

26. The implementation of such a system implies that institutions preparing metadata in a standard format on their websites will make it possible for Eurostat (and other organisations) to access this information from the web rather than putting in place ad hoc transmissions in various formats. The progress made on the identification of commonalities in the existing metadata systems and on the standardisation of the terminology and concepts used (see the SDMX work on cross-domain concepts and on the “Metadata Common Vocabulary”) will help reducing the metadata reporting burden of national institutes and, at the same time, will improve the quality and consistency of metadata descriptions across countries.

27. While working on the technical infrastructure, Eurostat is currently improving the granularity of the reference metadata format, with the aim of extending the conceptual coverage of the format and in particular for incorporating more elements on quality assessment, according to the criteria identified by the European statistical code of practice. The modular list of concepts (described in Annex 2) is built on the current format used within Eurostat, with some limited extensions on quality elements which are going to be further detailed by the end of this year. The current list is going to be used for testing the possibility of disseminating a good selection of reference metadata with regard to PEEI data.

C. Statistical issues

Concepts

28. As mentioned in the previous paragraph, the harmonisation of concepts is indispensable for sharing data. Fortunately, for most PEEI and National accounts data (ESA95) this is already the case. For other data flows, for instance in the area of social statistics, this has to be checked case by case before integrating them into SODI. Especially when the Pull method is used, it is indispensable that Member States and Eurostat use the same concepts for publishing data. This also includes coordination on the possibility of disseminating seasonally adjusted figures for infra-annual data, on the method used and on the provision of relevant methodological explanations to the user.

29. Eurostat, under the “Data Life Cycle initiative” (CVD, Cycle de Vie des Données) plans to reduce the number of code lists in use for the same concept. SODI – by requiring a unique concept for reception, production and dissemination – contributes to this effort.

Statistical confidentiality

30. At the moment, we are not planning to cover data flows where all or part of the data are confidential. However, this has to be checked for all data flows before they are considered by SODI; and in case of (partial) confidentiality this has to be treated correctly by the dissemination modules.

Validation

31. SODI assumes that national data are for publication. However, even when Eurostat is taking the national value “as it is”, a minimum amount of technical verification has to be done to prevent from human errors or technical problems in the transmission process leading to application failure or inconsistent data. So, at least a formal verification (correctness of XML syntax, adherence to SDMX standard, compliance with code lists) has to be done. More statistical validation should only be

performed, at this stage, when fully automated. In addition, the response to an erroneous message has to be defined.

Footnote Treatment

32. Footnotes are part of the SDMX data model, so the processing of footnotes in any environment conforming to SDMX is neither a problem for the standard nor a technical issue. However, in addition to footnotes received with the SDMX message containing the data, there may be additional footnotes at different levels (footnotes might apply to a single observation, a country, a period or some other dimension of the data) which have to be treated correctly. First, in the output national data have to be correctly tagged as “national”; second, footnotes added by a Eurostat production unit have to be applied correctly; third, there may be standard footnotes for certain dimensions, which are defined by Eurostat, but have to be applied as well (or exclusively) to national data. This could be the case when data do not include the whole country (for instance, German data before October 1990) or a country uses a slightly different concept for one dimension (e.g. a different method of seasonal adjustment).

D. Political Issues

The role of Eurostat in a data sharing environment

33. In a data sharing environment, the role of Eurostat has to be redefined. Will Eurostat only become a coordination body, responsible for the harmonisation of concepts and methods, and maybe with a stronger role in quality assessment, or will Eurostat do more than just compile the national data? Which data treatment will still remain to Eurostat? How to manage shared responsibility with other organisations with respect to the maintenance and update of statistical structures (structural definitions, code lists, etc.)?

Aggregates

34. Especially delicate is the question of aggregates (like EU25, EU15, Euro-zone), which are calculated by Eurostat. Will these aggregates only be calculated for data treated by Eurostat, or will it just be the aggregate of the data available. It might have to be explained a situation where a European aggregate does not correspond to the data in a given table.

Releases

35. This regards the treatment of different releases of the same data: at the moment, the basic idea is to mark, in the Eurostat tables, the incoming data as “national”, before they have been processed by Eurostat and replaced by “European” data. This becomes a problem in the (frequent) case of successive releases. When releases come, will the “national” updates substitute the (older) “European” ones, or will they stay invisible, until they have been processed by Eurostat? And what about the aggregates in this case (there might even be a different aggregate policy for original and updated data).

Embargo Policies

36. Short term statistical data are often published under an embargo policy, i.e., data may not be published before a certain date and time, however, this often differs in the European Statistical System (especially concerning dates and handling of delays). It has to be clarified whether SODI will require a harmonisation. A special question arises when Eurostat’s embargo date is later than the national ones. Should in this case the “national” data already be published, of course without a European aggregate?

Ownership of the Data

37. At the moment, it is planned that Eurostat marks the incoming data in a footnote as “national” and removes this footnote after validation by the production unit. So Eurostat would distinguish between data, for which it takes the ownership and responsibility, and national data, where the National Statistical Institutes are the responsible owners.

E. Organisational Issues

Coexistence of Standards and Methods

38. As Member States might progress with different speeds or set different priorities for their dissemination systems, we might experience a very long period with both Push and Pull method and both SDMX-ML and SDMX-EDI used in parallel. Eurostat's data reception environment has to care for an automatic transparent integration and conversion process, unless there is an explicit expressed will in the ESS to agree on a single method and/or a single data format.

Embargo Treatment

39. Currently, embargos are handled by the production unit processing the data. In the future, if Eurostat ceases the manual treatment of data for certain data flows, the Eurostat embargos have to be handled by the dissemination environment, while the national embargos have to be handled by the reception environment (this applies only for the "Push" approach, with "Pull" we expect data not be available to Eurostat before the national publication). A special case arises when the national and the European data shall be published simultaneously. In this case, the "Pull" as currently designed will not achieve precisely synchronous publication, as the necessary processing by Eurostat will cause a delay for the publication of the European data. It might be necessary to redefine slightly the objective of "simultaneous publication".

Integration of SODI into Eurostat's Data Life Cycle

40. In its basic idea, the publication of national data on Eurostat's web site is opposite to the data life cycle project of Eurostat, as neither the reception nor the production environment is used in the case of the "Pull" model. On the other hand, the idea of a single entry point and a reduction of the number of production systems are vital for the correct functioning of Eurostat's IT. So, the Pull method will be integrated into the single entry point. In particular, the requestor will be integrated into the eDAMIS system. In addition, eDAMIS will learn to handle SDMX-ML as a generic format (like it handles GESMES today) not requiring a separate envelope for the metadata. In the production environment, a special "SODI" process will be created. Although this is at the moment an additional process, finally, with further harmonisation in the ESS, several currently different processes could be given up in sake of this single process.

F. Legal Issues

"Pull" and the Obligations of Member States

41. Before using the Pull method in production for data flows covered by Commission or Council Regulations, it has to be clarified if this method fulfils any obligations of the Member States to deliver data to Eurostat. Legally, there can be a difference between the obligation to deliver data (without being explicitly asked) or to provide the data on request (as in the Pull method).

SODI and SDMX in Legal Acts

42. Normally, a general reference to a standardised format in legal acts is preferred, rather than a specification of the data format. This is wise as the lifecycle of legal acts is normally longer than the lifecycle of data transmission methods. Although we expect XML formats like SDMX-ML to have a very long lifetime, technical progress on transmission protocols or XML related standards (like IPv6, XML security, standardisation of web services, and so on) will influence the SDMX standard and its implementations. On the other hand, we expect SDMX-ML – with the standardisation by ISO and the commitment of the stakeholders – to become a widely accepted standard, so it will be perfectly covered by most existing legal acts.

Principal European Economic Indicators List

Set 1: Price Indicators

- 1.1. Harmonised Consumer Price Index: MUICP flash estimate: release end of reference month
- 1.2. Harmonised Consumer Price Index: actual indices: release 2,5 weeks after reference month

Set 2: National Accounts Indicators

- 2.1. Quarterly National Accounts: flash GDP: release t+45
- 2.2. Quarterly National Accounts: first GDP release with breakdowns: t+60
- 2.3. Quarterly National Accounts: Sector Accounts: release t+90
- 2.4. Quarterly Government Finance Statistics: release t+90

Set 3: Business Indicators

- 3.1 Industrial production index: release t+30
- 3.2 Industrial output price index for domestic markets: release t+35
- 3.3 Industrial new orders index: release t+50
- 3.4 Industrial import price index: release t+30
- 3.5 1. Production in construction: quarterly: release t+45
2. Monthly: release t+30
- 3.6 Turnover index for retail trade and repair: release t+30
- 3.7 Turnover index for other services: release t+60
- 3.8 Corporate output price index for services: release t+60

Set 4: Labour Market Indicators

- 4.1. Unemployment rate: release t+30
- 4.2. 1. Job vacancy rate: quarterly
2. monthly: release t+30
- 4.3. 1. Employment: monthly release t+30
2. quarterly: release t+45
- 4.4. Labour cost index (US: Employment cost index) release t+60

Set 5: Foreign Trade Indicators

- 5.1. External trade balance:
intra- and extra-MU; intra- and extra-EU: release t+46

EUROSTAT – SDMX CROSS-DOMAIN CONCEPTS MAPPING

EUROSTAT DISSEMINATION METADATA CONCEPTS		MAPPING TO CURRENT DRAFT OF SDMX CROSS- DOMAIN CONCEPTS
Top level	Child level	
Metadata Update	Last certified without update	Date of update
	Last update of content	Date of update
Contact	Organisation	Contact
	Address	Contact
	Contact name or service	Contact
	e-mail address	Contact
Data coverage	Short description of data domain	Data presentation
	Data breakdown and main variables	Data presentation
	Units of measure	Data presentation
Periodicity	Periodicity of compilation	Frequency and periodicity
	Database frequency	Frequency and periodicity
Timeliness and punctuality	Timeliness	Timeliness and punctuality
	Punctuality	Timeliness and punctuality
Transparency of practices	Legal acts, reporting requirements	Institutional framework
	Rules on confidentiality	Institutional framework
	Internal access	Transparency
	Commentary on the occasion of release	Transparency
	Notification of changes in methodology	Transparency
Accessibility	Release calendar	Release calendar
	Simultaneous release	Simultaneous release
	Dissemination formats	Dissemination formats
	Documentation on methodology	Accessibility of documentation
Quality cross-checks	Related data and quality cross-checks	[No direct concordance]
	References to quality reports	[No direct concordance]
Accuracy and reliability	Overall accuracy assessment	Accuracy
	Quality checks before release	Accuracy
Comparability and coherence	Comparability over time	Comparability and coherence
	Comparability over space	Comparability and coherence
	Comparability with related sources	Comparability and coherence
	Comparability between datasets	Comparability and coherence
	Breaks in time series	Comparability and coherence
Relevance	Rate of available statistics (user needs)	Relevance
	Intended audience and purpose	Relevance
	Supplementary data	Supplementary data
Statistical concepts and classifications	Statistical concept	Statistical concept
	Definition of indicators	Statistical concept
	Classification system	Classification systems
	Conformity with official standards	Classification systems
	Classification coverage	Classification systems
Scope of the data	Reference area / geopolitical entity	Scope/coverage
	Time coverage	Scope/coverage
	Statistical unit	Scope/coverage
	Statistical population	Scope/coverage
Accounting conventions	Reference period	Accounting conventions
	Base period	Accounting conventions
	Basis for recording	Accounting conventions
Nature of basic data	Data source used	Source data
	Type of survey	Source data
	Methods of data collection	Source data
Compilation practices	Compilation	Statistical processing
	Adjustments and weights	Statistical processing
	Data validation	Statistical processing
	Revision policy and practice	Revision policy and practice
Other	Warnings on re-use and limitations	[No direct concordance]

**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): Using metadata for searching and finding statistical data in websites and portals

STATISTICAL METADATA IN STATISTICS NORWAY
Contributed paper

Submitted by Statistics Norway¹

I. INTRODUCTION

1. Development of functionality and services providing users easy access to and use of the metadata systems is central in Statistics Norway's (SSB) metadata strategy. This paper will focus on the content and development of a dedicated area for statistical metadata on SSB's website. The overall aim of this development is to make the contents of all our metadata systems more accessible and easier to use. The aim for 2006 is to test the statistical metadata area with users inside SSB. The contents of our variables, classifications and file descriptions servers will be displayed in this area. The design will be flexible so that the contents of other metadata systems can be added when appropriate.

II. METADATA STRATEGY

2. Statistics Norway has developed many different metadata systems to serve different purposes and different user groups. In the last years, there has been a strong focus on the need to link existing systems and a requirement that new metadata systems should not be built in isolation. To facilitate this Statistics Norway has developed a metadata strategy, which was approved early in 2005, together with a metadata plan for the next few years [1].

3. The strategy focuses on establishing a common understanding through establishment of documentation and concepts linked to metadata, clear roles and responsibilities, and a stepwise development of content, integration and linkage of master systems for metadata. Our aim is that metadata should be updated in one place and accessible everywhere. Our metadata systems should also serve as tools for the harmonisation and standardisation of our documentation.

4. The metadata plan comprises almost all Statistics Norway's metadata and contains several project proposals. Some of these projects started before the strategy was finalised, but they are now taking the strategy into account in further development:

- Definitions of key concepts linked to metadata

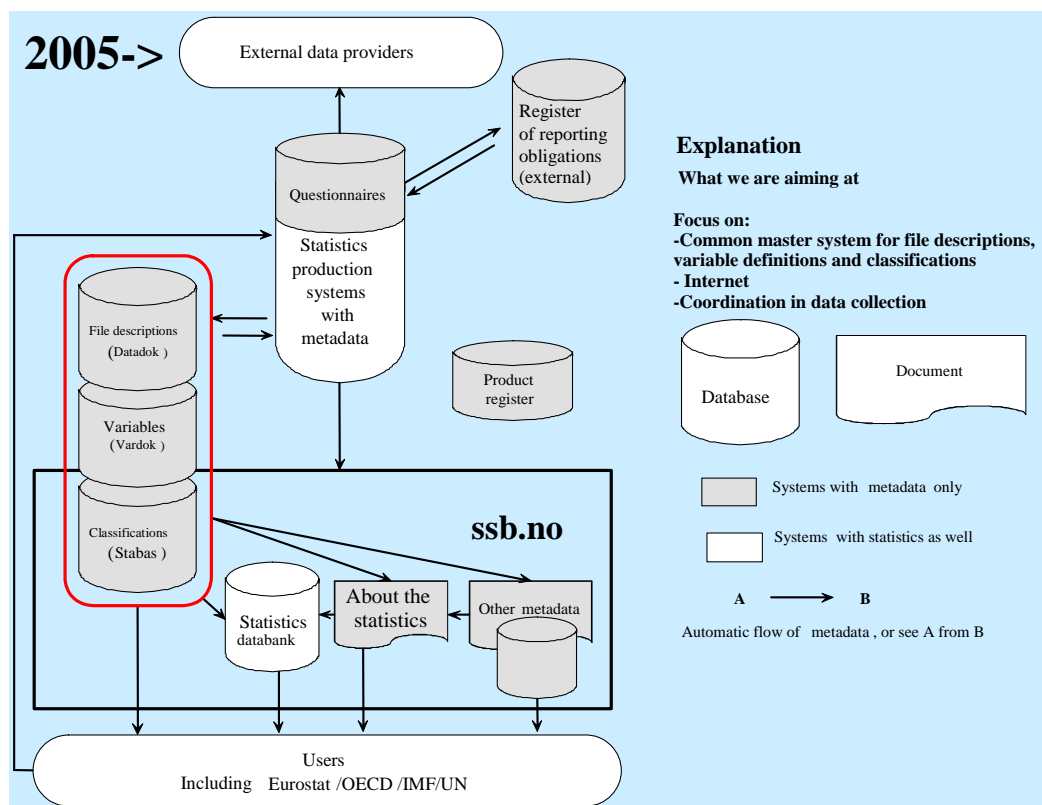
¹ Prepared by Anne Gro Hustoft (agt@ssb.no) and Jenny Linnerud (jal@ssb.no)

- Development of master systems
 - Further development of functionality and content of the master systems for standard classifications (Stabas), variables definitions (Vardok) and file descriptions (Datadok)
 - Coordination and linkage of the master systems and other data and metadata systems
- Other metadata systems and activities
 - About the statistics
 - Information about micro-data
 - Data collection metadata systems
 - For more proposals see [1].

5. A small group was set up to define key quality and metadata concepts. The set of key concepts was limited, but included the concepts quality, quality (dimensions) in statistics, statistics, register, variable, indicator, indicator, classification, code list, statistical unit (observation, reporting and analysis unit), measurement unit and population. It has already been experienced that even if it might be easy to agree on some basic definitions such as those given by the SDMX Metadata Common Vocabulary and existing Norwegian concepts related to these, the challenge is to apply these in practice, in a statistical table or on a micro level. The establishment of concepts linked to metadata is crucial both for statisticians and others to know what, how and where to put metadata in the different systems.

6. Figure 1 shows important elements of Statistics Norway's metadata systems that we are moving towards. Focus has been put on the linkage of master systems for classifications, variables definitions and file descriptions, that more systems and metadata will be available on the external web and on coordination of data collection.

Figure 1. Development of metadata systems in Statistics Norway



7. Efficient exchange of data and metadata across different systems and platforms is a major issue. In 2005 Statistics Norway prioritised a service oriented architecture. 9 services for variables, 3 services for file descriptions and 3 services for classifications were prototyped. The introduction of these services has been so well taken up by the organisation that the demand is currently exceeding the supply. Further development of services is planned to continue until 2008. We aim to link all our metadata systems by services thereby improving the availability of updated information.

III. STATISTICAL METADATA WEB PAGE

8. The overall purpose of the Statistical metadata web page is to make Statistics Norway's metadata systems more accessible and easier to use. Both internal and external users will get easier access to the metadata by displaying the contents of these systems in a common web page. Our work within this area has been inspired by the corresponding web pages of Statistics Canada (www.statcan.ca/english/concepts) and Statistics New Zealand (www.stats.govt.nz/statistical-methods).

9. The main purpose of the web page is to give access to information stored in the metadata systems and delivered by web services but the page will also contain links to other relevant metadata. This year we will establish links to the metadata connected to each published statistics ("About the statistics"), metadata about our data collections ("About the data collection"), questionnaires and other relevant documentation. The contents of the web page will be extended as other relevant metadata becomes available.

10. The figure below gives a picture of the links we intend to include on our web page. All these metadata exist at present, but they are stored in different systems and different places on the web. They are difficult to find for the inexperienced user.

Figure 2. Links in the statistical metadata web page

About the statistics	Standards
• Listed by division	• <u>Classifications</u> (Stabas)
• <u>Listed</u> by subjects	• Nomenclature and documents
• Advance release calendar	
	File descriptions and registers
<u>About</u> the data collection	• File descriptions (Datadok)
	• Registers (Datadok)
Definitions	
• Statistical units	Statistical methods
• Variables (Vardok)	• Articles
• Terminology	• Reports
<u>Questionnaires</u>	International links
	• Definitions
	• Standard classifications
	• Statistical methods

11. A project group consisting of three IT-specialists and one standards/metadata specialist is in charge of developing the web page. In addition specialists within different subject matter areas will test the system during the development process. The project group will also forward a plan for organisation and quality assurance connected to the operation of the underlying metadata systems.

12. Our aim is to make the statistical metadata web page accessible to users within Statistics Norway in 2006, and to make a version for external users in 2007. The external version of the statistical metadata web page (2007) will also be available in English. Planned resources for the development of an internal statistical metadata web page (2006) is 1300 man-hours for the IT-specialists and 300 man-hours for the standards/metadata specialist who is also the project manager.

IV. STATISTICAL METADATA

13. In this section we give some more information about the underlying metadata systems and other relevant metadata that will be made available through our web page.

A. Documentation of variables (Vardok)

14. At the METIS-meeting in 2004 Statistics Norway presented a system for documentation of variables (Vardok) [2]. This is a central system for documenting variables (e.g. definition, validity periods, classifications used) and also a tool for harmonisation of names and definitions of variables.

15. Since 2004 the Vardok system has been further developed by introducing the possibility of multilingual functionality (English and two different versions of Norwegian), of describing variables by equations and of linking different variables (if variable A is the sum of variable B and variable C, you don't have to define B and C when you define A, you just link to the definitions of B and C).

16. This year Vardok will be linked to our event-history database and About the statistics. Last year it was linked to our system for dissemination of statistics (Statbank). The screenshot below shows the current situation for Vardok regarding information fields. Only four information fields are available for translation to English. The rest should be automatically translated or are deemed unnecessary.

Figure 3. Information fields in Vardok

The screenshot shows the 'Variable details' window in the Vardok system. The window title is 'TESTVERSJON - VARDOK - Documentation av variables in Statistics Norway 07-MAR-2006 14:26:53'. The interface is in English. The 'Primary language' is set to 'Bokmål'. The variable being edited has the ID '1294' and the name 'Farm type of agricultural holdings'. The definition is: 'The type of farming of a holding is determined by the relative contribution of the different crop and livestock enterprises to its total agricultural production. Standard gross margin is applied as common measurement of the various enterprises (crop and livestock). The classification of farm types has 4 levels.' The variable is owned by '130', has a sensitivity of 'Ordinær', and a contact of 'pro'. It is valid from '31.07.1999' to an empty date field. The 'Stat/Obs unit' is 'Holding'. There are fields for 'Internal document', 'External document', and 'Beregning'. The 'Internal comments' field is empty, with an 'Ext. comm.' button. The 'Standard' field has radio buttons for 'Yes' (selected) and 'No', with a value of 'Klassifisering av jordbruksbedrifter etter driftsform 1999'. The 'Codelist' field has radio buttons for 'Yes' (selected) and 'No', with a value of 'Agriculture'. The 'Subject area' is '10.04.10'. The 'Statistic' field is empty. The 'SSB source' is 'Landbruksstatistisk system'. The 'External source' is empty. There are checkboxes for 'Files in Datadok' (Yes/No), 'Defn. approved for Internet' (checked), 'Internal use' (checked), and 'Linked' (Yes/No). At the bottom, there are buttons for 'Export to Word', 'Save', 'OK', and 'Cancel'. The 'Created' date is '21.04.2005' by 'pro', and the 'Edited' date is '07.03.2006' by 'jal'. The 'Copy Id' field is empty. The window is part of a larger application with a menu bar (Exit, Edit, Search, Export, Reports, Help, News, Window) and a status bar (Post: 1/1, <DSK> <FSK>).

17. We are also about to finish a web-version of the variables documentation system.

18. An important part of our work the last two years has been documentation and quality assurance of new variables. At present (March 2006) 1200 variables are documented in Vardok. 1104 of these are approved for dissemination within Statistics Norway, 421 are approved for dissemination outside Statistics Norway (the figure below shows the development in the number of variables documented). The variables that are not approved for dissemination are still being discussed within the subject matter divisions that are responsible for them. As soon as the variables are approved for dissemination outside Statistics Norway, they can be displayed on the Internet through Statbank, About the statistics and About the data collection.

Figure 4. Progress in Vardok

Year	Number of variables documented per year	Number of divisions with access per year
2002	157	5
2003	352	5
2004	323	6
2005	284	11
2006	84 so far	12
Total	1200	-

19. The number of divisions with access to Vardok has been restricted (12 of maximum 18 subject matter divisions in 2006) but as soon as the division flags their variables as approved for internal use, all other divisions can give feedback on the definitions.

20. Within the Vardok-project we have made a special effort to start harmonising the accounts statistics variables. The different subject matter divisions in many cases use the same names for their accounts statistics variables, but define them a bit differently, because they are subject to different laws and regulations. The treatment of these variables therefore requires a tighter cooperation between the involved divisions than the documentation of other variables.

21. Parallel to the development of the variables documentation system, SSB participates in the Neuchâtel group² where terminology and models connected to variables are discussed.

22. 2006 is the last year in the development phase for the Vardok-project. Next year all 18 divisions will be able to document their variables definitions in Vardok.

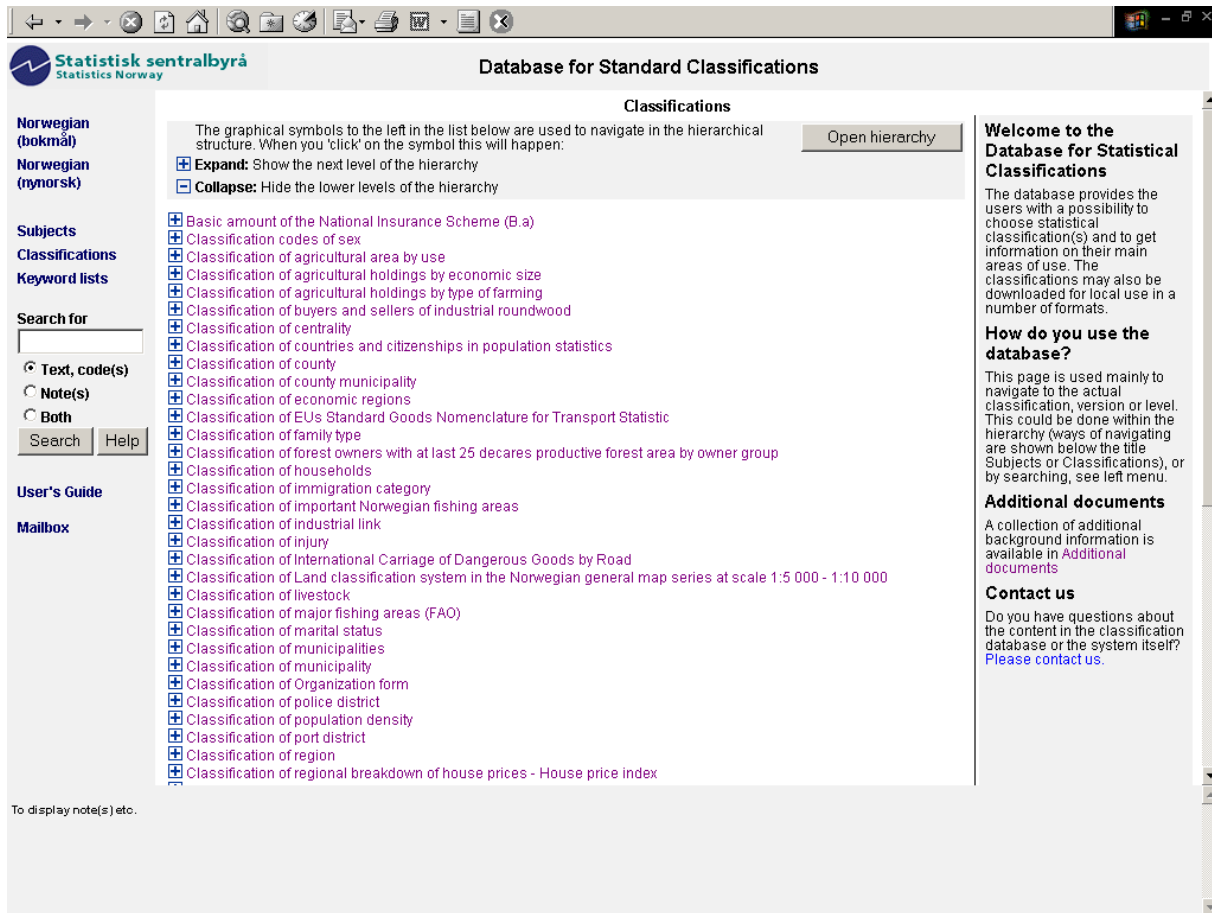
B. Classifications (Stabas)

23. Statistics Denmark and Statistics Norway started a joint project in 2001 to implement the Neuchâtel model in a multilingual standard relational database. The project implemented Version 2.0 of the Neuchâtel terminology for Classification database object types and their attributes with a few attributes being left out. Production started in Statistics Norway in 2002.

24. The model was implemented using Microsoft Access Database for demonstration and testing purposes and Oracle for production purposes. An application in Visual Basic was developed for maintaining the database (creating and updating classifications including variants, indexes and correspondence tables). Import and export functions have been defined to a preliminary XML format. In addition a web-based browser was developed and is used for the dissemination of selected classifications on the Internet.

² The Neuchâtel group working with terminology models for classification databases was established in 1999 and consisted of Statistics Denmark, Statistics Sweden, Statistics Switzerland, Statistics Norway and run Software-Werkstatt. Statistics Netherlands and the Bureau of Labor Statistics (USA) have joined the group for the work on variables.

Figure 5. Database for Standard Classifications (www.ssb.no/english/stabas).



25. At present (March 2006) we have 51 current classification versions on our Intranet, 92 older versions and variants on our Intranet and 49 current classification versions on SSB's website. With very few exceptions these classifications are available in three languages (English and two different versions of Norwegian). A one-way link from Vardok to Stabas was established in 2003 and from Statbank to Stabas in 2005.

26. The statistical metadata web page will link up directly to our web-based version of Stabas and provide some administrative reports to be used in keeping the system updated and the quality of the information high.

C. File descriptions (Datadok)

27. We document all permanent data files in our file documentation database Datadok. The database was built in 1998 but wasn't mandatory until 2002. At present (March 2006) we have over 6 500 file descriptions stored there in Norwegian. There is a two-way link between Vardok and Datadok. We are currently running 3 prototype services for Datadok to provide the statistical metadata web page with more easily available and searchable file descriptions.

D. About the statistics

28. About the statistics is metadata that describes each statistics that is published by Statistics Norway. It contains administrative information, information about statistics production, variables, concepts, sources of errors and uncertainty, comparability, coherence and availability.

29. Last year we started implementing About the statistics using an CMS (Content Management system)-platform. There were two important reasons for doing this. The first was that use of CMS will make general updating of About the statistics more easy for the subject matter divisions. The other reason is that CMS will make it possible to link About the statistics to Vardok and Stabas. Then we will no longer define variables in

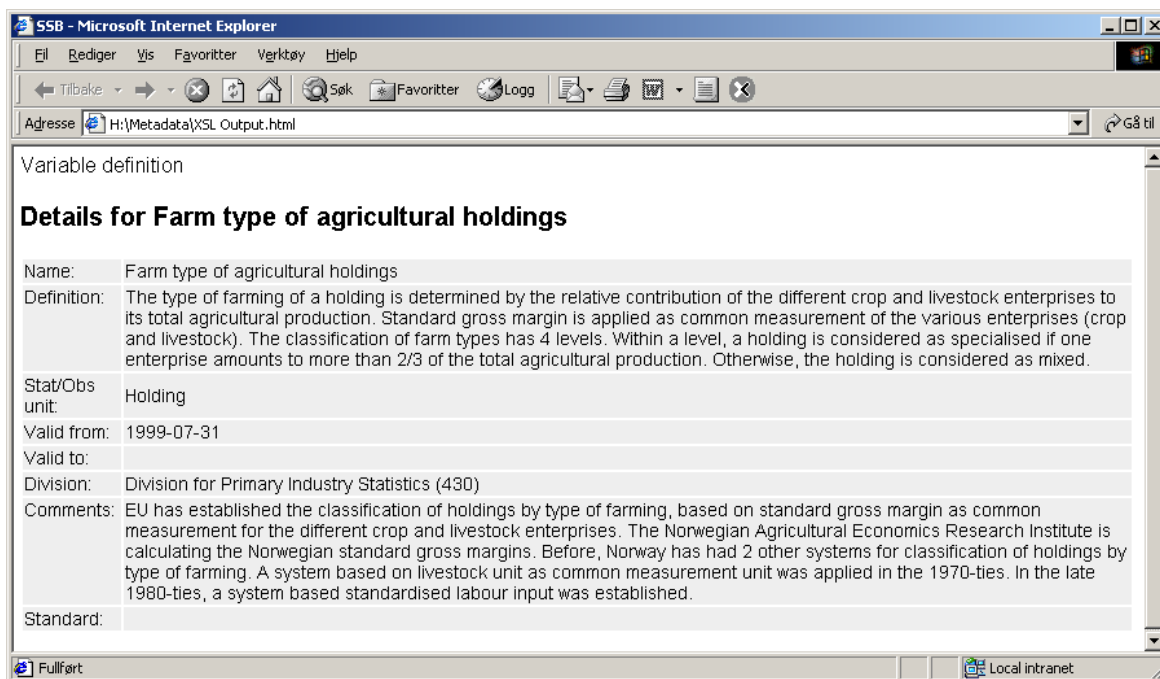
About the statistics, but link to the relevant variables in Vardok. In the same way we will link to the relevant classifications in Stabas. This will bring us closer to the goal that all definitions of variables should be documented and updated in one place (Vardok) and linked to other places where they are needed.

E. About the data collection

30. Researchers frequently use data collections from Statistics Norway for their research. However, the process from finding out what you need, to actually getting the data, may be long and troublesome, especially for inexperienced researchers. Statistics Norway has therefore (with support from the Research Council of Norway) developed a website (soon to be translated into English) to make information about this process more easily available. Among other things, this page provides the users with documentation of several data collections. Each data collection has a general description e.g. of data quality, and it also contains a list of relevant variables, including variable documentation from Vardok.

31. The next figure shows how the variable documentation is displayed through About the data collection on the Internet. If you compare it to fig. 3, you see that About the data collection has chosen to collect only those Vardok-fields that were found relevant for this specific group of users (researchers). At present Statbank only displays the definition information field. This is all that Statbank has found relevant for their specific group of users (mainly members of the public).

Figure 6. Variables documentation in About the data collection



F. Questionnaires

32. In SSB we have developed one online data collection process for businesses (IDUN) and one offline data collection process for reporting from local to central government (KOSTRA). In 2006 we are developing one metadatabase for questionnaires to support data collection from both businesses and local government. We are also re-examining the tools and processes with which we make questionnaires (paper, electronic, computer assisted interviews etc). In 2006 our ambition level is only to make all our questionnaires available on the statistical metadata web page in pdf format. In the future we would like to connect here to a question and questionnaire bank.

G. Other documentation

33. Other documentation will consist of links to international websites (e.g. SDMX Metadata Common Vocabulary, Eurostat's standard classifications in Ramon and Eurostat's definitions in Coded), publications related to statistical methods and other metadata documents (e.g. terminology).

IV. METADATA MANAGEMENT

34. Our plans for organising metadata management are still developing, but we have identified some points that we think will be important to achieve a successful future for the statistical metadata web page:

- All subject matter departments should appoint one person responsible for the topic of metadata. These metadata managers must have thorough knowledge of the metadata systems linked to the page, and should act as advisers to other people in their department. The responsible persons should also work together to solve metadata questions involving cross-departmental subjects.
- All subject matter divisions should have a contact person for Vardok, Stabas and Datadok to provide training and support to new users.
- The Department of dissemination should be involved and secure a unified presentation of the metadata disseminated to external users.

35. This preliminary list of points will be extended during 2006, and will result in a concrete plan for metadata management.

REFERENCES

[1] Metadata strategy in Statistics Norway. Hans Viggo Sæbø. Eurostat Metadata Working Group meeting, Luxembourg, June 6-7 2005.

[2] Variables documentation system in Statistics Norway. Anne Gro Hustoft and Jenny Linnerud. Contributed paper to the Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS), Geneva, February 9-11 2004

[3] Neuchâtel v2.1 <http://europa.eu.int/comm/eurostat/ramon/miscellaneous/index.cfm?TargetUrl=D>

**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)

UNECE TASK FORCE ON ELECTRONIC DATA REPORTING FOR PRIMARY DATA COLLECTION

Note prepared by the UNECE Secretariat

SUMMARY

1. The purpose of this paper is to inform the participants at the April 2006 Work Session on Statistical Metadata (METIS) about the establishment of the Task Force on Electronic Data Reporting, its programme of work, and to solicit a broader interest and collaboration among the national statistical offices.

BACKGROUND

2. Electronic data reporting has already long tradition in the UNECE's statistical work programme. Since early 90's the UNECE was involved in the development, lead by Eurostat, of general standards for the electronic data interchange in statistics. This work is continued by a number of international organizations.

3. In response to the needs of national statistical offices, a work session on statistical data reporting was organized in February 2002 (<http://www.unece.org/stats/documents/2002.02.edr.htm>). The first work session recommended following up to this activity and continuing the international exchange of experience. This activity was undertaken jointly with Eurostat

4. In 2004 the Conference of European Statisticians has selected the topic of electronic data reporting as a theme of its seminar at its 2005 plenary session. In concluding the seminar the Conference decided to continue the international cooperation in electronic data reporting with the aim to share experience and solutions, make recommendation on specific issues (including security), and explore possibilities for joint development of open source EDR. The Bureau of the Conference has approved the Terms of reference of the Task Force in October 2005 and the Work Plan in February 2006, and recommended that the participants at the METIS work session are informed about this initiative. The Bureau of the Conference also agreed that the UNECE would seek a possible cooperation with Eurostat and OECD.

THE MANDATE (TERMS OF REFERENCE)

5. The mandate of the Task Force on Electronic Data Reporting for primary data collection is the following¹:

- a) share solutions for the efficient integration of EDR in the context of multi-mode surveys with EDR as an option;
- b) share experiences in:
 - meta-data driven generalised system for EDR;
 - promoting take-up by respondents of the EDR response option;
 - creating and promoting common data taxonomies, including their relevance, community of interest take-up and the wider applicability in linking business reporting and statistical data collection;
 - participating in a wider Whole of Government approach to Electronic Data Reporting;
- c) make recommendations regarding special issues involving electronic questionnaire design;
- d) outline what is known and not known about possible mode effects;
- e) document the current state of knowledge regarding XBRL, SDMX and other “pull” based collection systems;
- f) make recommendations regarding security arrangements and their public communication;
- g) explore the potential for joint development of open-source EDR and case management solutions;
- h) document special issues affecting Internet collection of population census information including the elements of a business case for electronic collection as an option, the case for outsourcing part of the development, integration with other collection modes.
- i) other related issues that the Task Force wishes to recommend.

6. In its work, the Task Force should take into account the different challenges regarding the EDR process for the individuals/households and businesses/other institutional respondents and the differences in national infrastructure (e.g. legislation, security rules).

7. The Task Force should organize an exchange of experience through e-mail and Internet and possible organization of a Work Session in 2006 (plan: October-November 2006). The Bureau of the Conference requests a report of the Task Force at its February 2007 meeting, for further discussion at the June 2007 plenary session of the Conference.

MEMBERSHIP OF THE TASK FORCE ON EDR

8. The members of the Task Force will be experts in national and international statistical agencies dealing with methodological and technical issues associated with electronic data reporting. The Task Force would be able to use the networks created through past and present events organized by the Conference and some national statistical offices in related areas, namely:

- Joint UNECE/Eurostat Work Session on Electronic Data Reporting;
- Joint UNECE/Eurostat/OECD Work Sessions on Statistical Metadata (METIS);
- Joint UNECE/Eurostat/OECD Meetings on Management of Statistical Information Systems;
- joint activities on EDR undertaken by statistical offices of Australia, New Zealand, Canada and the US Bureau of the Census.

METHODS OF WORK, TENTATIVE TIMETABLE AND FORM OF FINAL OUTPUTS²

April 2006 The Task Force will be launched (METIS meeting). The output from this phase will be the membership list.

April – June 2006 The Task Force members will agree on a detailed schedule and will divide between themselves responsibilities for particular issues listed in the TOR. This phase will be carried out using e-mail

¹ See ECE/CES/BUR/2005/11/Rev.1

² See ECE/CES/BUR/2006/16

and teleconferencing consultations. The output from this phase will be an outline of the final output with responsibilities for its individual components.

June 2006-October 2006 The Task Force members will make a first collection and selection of material on good practices, with a view to presenting the material at the Work Session on EDR.

October or November 2006 The UNECE will organize a work session, possibly with Eurostat and/or OECD, on electronic data interchange. The aim of the work session is to submit the material already collected and prepared for a broader discussion, and to solicit more experiences and good practices from countries not directly involved in the Task Force.

February 2007 The Task Force will submit a progress report to the CES Bureau.

December 2006 – April 2007 The Task Force will work, using e-mail and teleconferencing, in progressing final outputs on individual issues listed in the work plan and prepare:

- a report for the June 2007 Plenary Session of the Conference;
- a website on good practices and national experiences in the field of EDR (that will become accessible on the public site after the approval of the report).

June 2007 – October 2007 The Task Force will incorporate the comments received into the report and the website and agree on a mechanism for future updating of the website. The public version of the report and the website will be made available in the course of this phase.

Late 2007 - Spring 2008 The UNECE secretariat may possibly organize a follow-up Work Session (in cooperation with Eurostat and OECD).

- - - - -