MetaNet

A network of excellence for harmonising and synthesising the development of statistical metadata

Proceedings of the Final MetaNet Conference organised by the University of Athens

Samos, Greece

7th – 9th May 2003

Edited by
H. Papageorgiou
University of Athens, Greece

June 2003

The MetaNet Project

A network of excellence for harmonising and synthesising the development of statistical metadata

Coordinator:

Dr Joanne Lamb, CES, University of Edinburgh, St. John's Land, Edinburgh, EH8 8AQ, UK

Email: J.M.Lamb@ed.ac.uk

INFORMATION SOCIETY TECHNOLOGIES
(IST)
PROGRAMME

Contract for:

Thematic Network

Annex 1 - "Description of Work"

Project acronym:	**METANET**
Project full title:	**A network of excellence for harmonising and synthesising the development of statistical metadata**
Proposal/Contract no.:	**IST-1999-29093**

Project Partners

University of Edinburgh (UEDIN), UK

Statistics Netherlands, Netherlands

University of Athens (UoA), Greece

University of Vienna (Viu), Austria

Statistics Sweden (SCB), Sweden

Statistics Norway, Norway

Survey and Statistical Computing (SASC), UK

http://www.epros.ed.ac.uk/metanet/

This is the report of the MetaNet Final Conference, which was hosted by University of Athens and held in the Greek island of Samos, 7-9 May 2003.

This document, entitled "Proceedings of the Final MetaNet Conference" is the official deliverable D8 of WP7 of the project. It contains not only the presentations given, but also discussions and integration activities of the five work Groups that have been established during the project, as can be identified from the Conference Agenda that is provided.

The participants who gave a presentation were advised to provide an accompanying document in order to enable University of Athens (UoA) who edited this report, to cover all issues presented adequately. In cases where a supportive document has not been provided, we simply summarize the slides presented.

The structure and motivation of the MetaNet project is initially described.

The First Conference of the project was held at the Statistics Netherlands premises, Voorburg, 2-4 April 2001. The objective of that conference was to present the project to a wide range of people and organisations interested in statistical metadata, to recruit interested participants for the various workgroups, and to advance agreement on the detailed objectives for each group.

The Main Conference, on the other hand, focused on the integration of all activities undertaken during the projects life cycle and the possible exploitation of the network's results by organisations and/or other projects on metadata.

Forty five (45) delegates from sixteen (16) National Statistical Institutes (NSIs), six (6) Universities, eight (8) public national and international organisations and four (4) private companies have participated to this Final Conference.

Conference Agenda:

The agenda covered all Work Groups research topics in an effort to bring together the experience gained in all the project's work packages and achieve an integration of the results. Furthermore, a number of experiences of the NSIs participating were also presented.

Presentations were given mainly on the following topics:

- ✓ Overview of the Metanet project and overall achievements
- ✓ Metanet Work Group 1 achievements
- ✓ Metanet Work Group 2 achievements
- ✓ Metanet Work Group 3 achievements
- ✓ Metanet Work Group 4 achievements
- ✓ Metanet Work Group 5 achievements
- ✓ The COSMOS project

- ✓ The CODACMOS project
- ✓ The CODAM implementation experience
- ✓ METIS
- ✓ ONS experience in metadata
- ✓ Experience of the Latvian Statistical Office
- ✓ Digital government in the US
- ✓ Banca d' Italia experience
- ✓ SORS experience
- ✓ Implementing standards in official statistics and practical problems in their implementation
- ✓ Terminology problems
- ✓ Using models for metadata
- ✓ Eurostat indicators
- ✓ FP6 initiatives

Agenda - Time schedule of the Conference:

MetaNet Final Conference

7th – 9th May 2003-

Samos, Greece

Agenda

Wednesday 7th May 2003

09:30 Welcome by Professor H. Papageorgiou

Session 1:MetaNet and the EPROS programme

Chair: Joanne Lamb

10:00 Introduction: what is MetaNet, and what has it achieved? (Joanne Lamb, UEDIN, co-ordinator of Metanet)

11:00 Tea/Coffee

11:30 The COSMOS project (Joanne Lamb,UEDIN)

12:00 The Amrads project (Karen Barrie, University of Edinburgh)

12:30 Lunch offered by the host

14:00 The ONS experience in intoducing Metadata (Jan Thomas, ONS)

Session 2: Metanet Work Group 1 and associated themes

Chair: Jean-Pierre Kent

15:00 Jean-Pierre Kent: Report on Work group 1

15:30 Tea/coffee

16:00 Using models for metadata (Andrew Westlake, SASC)

16:30 Implementing standards (Simon Musgrave, UKDA)

17:00 Round table

17:30 Close for the day

19:30-10:30: Trip to Vathi *(exact pick-up time from each of the recommended hotels will be announced during the conference).*

Thursday 8th May 2003

Session 3 Metanet Work Group 2 and associated themes

Chair: Wilfried Grossmann

09:00 Wilfried Grossmann: Report on Work Group 2

09:30 The Banca d'Italia experience (Vincenzo Del Vecchio)

10:00 The Codacmos project (Alberto Sorce, ISTAT)

10:30 Round table

11:00 Tea/Coffee

Session 4: Metanet Work Group 3 and associated themes

Chair: Matthias Abelin

11:30 Matthias Abelin: Report on WG3

12:00 The SORS experience (Joza Klep, SORS)

12:30 Lunch offered by the host

14:00 Experiences in the Latvian Statistical office (Karlis Zeila)

14:15 Round table

Session 5: Metanet Work Group 4 and associated themes

Chair: Jan Magnar Byfuglien

14:45 Jan Magnar Byfuglien, Report on WG4

15:15 The CODAM implementation experience (Claude Macchi, SFSO)

15:35 Tea/coffee

16:00 Practical Problems of Implementing metadata standards In Official Statistics (Jozef Olenski; presented by Dusan Prazenka, Infostat)

16:30 Round table

Session 6: Metanet Work Group 5 and associated themes Terminology

Chair: Dusan Soltes

16:30 Dusan Soltes: Report on WG5

17:00 Terminology models (Reinhard Karge. Run software)

17:30 Close for the day

20:00 Dinner offered by the host in Arion hotel. Spouses and friends of the participants are welcomed.

Friday 9th May 2003

Session 6 continues

Chair: Dusan Soltes

09:00 Digital government in the US, (Carol Hert, Syracuse University)

09:15 METIS (Dusan Soltes, University of Bratislava)

09:30 Eurostat indicators (Zdenko Milonja, CSO, Croatia)

09:45 Round table

Session 7: Future developments and opportunities.

Chair: Joanne Lamb

10:15 Introduction (Joanne Lamb, UEDIN)

10:35 The e-NIPS project (John Charlton, ONS)

11:00 Tea/coffee

11:30 Helping People Understand Statistics Terms: The Statistical Interactive Glossary (SIG) and the GovStat Ontology (Carol Hert, Syracuse University)

12:00 Round table

12:30 Lunch offered by the host

Session 8: Conclusions

Chair: Andrew Westlake

14:00 Introduction (Andrew Westlake, SASC)

14:30 Open Forum

15:30 Close

Participants

The following persons participated to the Conference:

	Name		Organization
1	Reinhard	Karge	Run Software - Werkstatt GmbH, Germany
2	Jean - Pierre	Kent	Statistics Netherlands
3	Dusan	Soltes	Faculty of Management, Comenius University, Slovakia
4	Jana	Meliskova	AMRADS, Switzerland
6	Ilana	Yaacobi	Central Bureau of Statistics, Israel
7	Mary	Sweetland	NHS Scotland, ISD, UK
8	Andrew	Westlake	London, UK
9	Jenny	Linnerud	Statistics Norway, Norway
10	Karlis	Zeila	Central Statistical Bureau of Latvia
11	Aija	Zigure	Central Statistical Bureau of Latvia
12	Vincenzo	Del Vecchio	Banca d'Italia, Italy
13	Alberto	Sorce	ISTAT, Rome, Italy
14	Jan	Byfuglien	Statistics Norway, Norway
15	Joanne	Lamb	University of Edinburgh, UK
16	Karen	Barrie	University of Edinburgh, UK
17	Fabien	Perrot	Swiss Federal Statistical Office, Switzerland
18	Claude	Macchi	Swiss Federal Statistical Office, Switzerland
19	Karl Anton	Froeschl	University of Vienna, Dept. of Statistics, Austria
20	Witfried	Grossmann	University of Vienna, Dept. of Statistics, Austria
21	Jaakko	Ranta	Statistics Finland, Finland
22	Heikkl	Rouhuvlrta	Statistics Finland, Finland
23	Carol	Hert	Syracuse University, Seatle, USA
24	Matthias	Abelin	Statistics Sweden, Sweden
25	Peter	Pukli	Hungarian Central statistical Office, Hungary
26	Edmond - Lucian	Sinigaglia	Romanian National Institute of Statistics, Romania
27	Miroslava	Brchanova	Czech Statistical Office, Czech Republic.
28	Simon	Musgrave	Nesstar, Colchester, Essex, UK
29	Bryan	Scotney	University of Ulster, Northern Ireland
30	Jozica	Klep	Statistical Office of the Republic of Slovenia
31	Zdenko	Milonja	Croatian Bureau of Statistics, Croatia
32	Sally	McClean	University of Ulster, Northern Ireland
33	Laurent	Plancq	World Systems, Luxemburg
34	Marcia	Taylor	European Centre for Analysis in the Social Sciences, UK
35	Irena	Vipavc	Slovene Social Science Data Archive, Slovenia
36	Jan	Thomas	Office of National Statistics, Hampshire, UK
37	Bill	Bradley	Health Canada, Ontario, Canada

38	Robert	Mayo	Food and Agriculture Org. of the United Nations
39	Delka	Grozeva Koeva	National Statistical Institute, Sofia, Bulgaria
40	Dusan	Prazenka	INFOSTAT, Slovakia
41	John	Charlton	Office of National Statistics, UK
42	Daniela	Raffaele	ISTAT, Italy
43	Richard	Warnett	AVITAR, UK
44	Babis	Papageorgiou	University of Athens, Greece
45	Maria	Vardaki	University of Athens, Greece
46	Mike	Hatzopoulos	University of Athens, Greece

The main part of this report contains a synthesis of all presentations given at the Conference, dividing them into thematic categories. Five chapters and corresponding thematic categories have been selected for the discussion of the presentations of the Conference:

Chapter 1. MetaNet overview and its five Work Groups' (WGs) achievements

Chapter 2. Experiences of metadata related projects

Chapter 3. National experiences (NSIs experiences and other national organisations)

Chapter 4. International experiences and presentations of general interest (standards, models, terminology, etc)

Chapter 5. FP6 and future plans

A summary of each topic presented, discussions and points raised, is provided in the beginning of the corresponding chapter.

The next part of this section gives a brief overview of the MetaNet project.

MetaNet - A network of excellence for harmonising and synthesising the development of statistical metadata (www.epros.ed.ac.uk/metanet)

Project Main Goals – brief description

The main objectives of MetaNet project have been the following objectives:

1. to develop proposals for standards in the methodology used for describing statistical metadata and statistical information systems
2. to develop proposals for recommendations on the metadata objects in a common conceptual model of statistical metadata

3. to disseminate these proposed standards to the relevant user communities and standards bodies
4. to interact with relevant FP5 projects on the development and agreement of these proposals, and to advise on methods of achieving coherence of approach in the field of metadata for statistical information systems
5. to integrate the different views of metadata into one model and bring together these different perspectives.

Establishment of the following Work Groups

Work Group 1: methodology and tools

Work group leader: Jelke Bethlehem, Statistics Netherlands

Duration: April - September 2001

Deliverable: Overview of technical aids to Metadata representation (D4)

The first Work Group is concerned with the technical aids to implementing metadata systems and exchanging metadata descriptions. Since the concept of metadata, and the demands that humans and machines will make on it is not static (witness the rapid impact of the Web), it is important to have structures and methodologies for utilising metadata in different representations. The first Work Group will therefore concentrate on ways and means of exchanging metadata between systems, identifying the possibilities and problems of different approaches (database, UML, XML, RDF, Schemas etc). The aim will be to alert future users of the possibilities that are emerging, and to give guidelines on how to keep abreast of these developments. The group will also survey the developments in customised software such as classification servers and repositories of classifications and definitions.

Work Group 2: harmonisation of metadata – structure and definitions

Workshop leader: Wilfried Grossman, University of Vienna

Duration: April 2001 - March 2002

Deliverable: The concept of metadata: A report on the nature of metadata and how these concepts can be used in practice (D5)

The second Work Group is concerned with the conceptual nature of metadata. It will consider all aspects of metadata relevant to statistical organisations, and identify the different strands of interest to different parts of the statistical organisation. A priori there are three such strands:

- Classifications – the ‘basic’ metadata, which give meaning to, coded numbers
- Process metadata which aids the efficiency of statistical systems, and transfers metadata from one part of the system to another, as well as using this

metadata in the production process. In particular there are metadata models held inside production and experimental systems which need to be explicated.

- Dissemination metadata, which is human oriented, and aids users in finding and interpreting the published data. This type of metadata is close to the library world, and very much web based.

Work Group 3: Best practice for migration

Workgroup leader: Bo Sundgren, Statistics Sweden

Duration: April 2002 - Sept 2002

Deliverable: Reference book for metadata standards and methodology. (D6)

The third Work Group is concerned with the practical implication of the findings of the first two Work Groups. It will look at examples of migration paths that have been achieved, and on the basis of this draw up strategies that can be developed for the needs of individual NSIs and other disseminators of Official Statistics. It will consider the different places within the production and dissemination systems where metadata can be valuable, and identify the role of the particular metadata in that environment. The report can be considered as a reference book for institutions wishing to adopt the recommended metadata standards and methodology.

Work Group 4: Adoption issues

Workgroup leader: Rune Gløersen, Statistics Norway

Duration: October 2002 - April 2003

Deliverable: A training manual for the adoption of metadata standards and systems (D7)

The final Work Group considers the implications of these findings for the institutions concerned. It will identify potential barriers to adoption – technical, structural and human, and consider how these might be overcome. Its output can be regarded as a training manual for the adoption of metadata standards and systems.

In addition, a **WG5: Terminology issues** was also established at the Metanet project meeting in Oslo, Norway, 9-11 October 2002.

According to the Annex 1, the main objective of the MetaNet Final Conference, was '*the presentation of the results of the Work Group by the contractors and engender a discussion of how the network, or a successor, can continue to contribute to the development of metadata standards and technology, and its take up'*.

METANET OVERVIEW AND THE FIVE WORK GROUPS' (WGS) ACHIEVEMENTS

In this section, the following presentations will be illustrated:

i) MetaNet overview by Dr Joanne Lamb

ii) WG1 Achievements by Mr Jean-Pierre Kent

iii) WG2 Achievements by Professor Grossmann and a follow up by Professor Karl Froeschl

iv) WG3 Achievements by Mr Matthias Abelin

v) WG4 Achievements by Mr Jan Magnar Byfuglien

vi) WG5 Achievements by Professor Dusan Soltes

vii) Metanet Reference Model by Mr Reinhard Karge

MetaNet overview

Joanne Lamb - Coordinator
CES, University of Edinburgh
J.M.Lamb@ed.ac.uk

The presentation mainly addressed the following topics:

- Presentation of the Partners and the members of the project
- Discussion about interested parties (associates) that have contributed with valuable experiences throughout the project framework
- Objectives and overall organisational approach of the network
- Activities planned and achievements according to the network's objectives
- Dissemination activities
- Expected results of the Main Conference and future perspectives

Several interesting points were raised in this presentation which are summarized as follows:

I. **A large number of members and associate members have expressed their interest in the network's activities** and have contributed with several papers during the project.

The ***members*** of the network were the following:

- Comenius University, Bratislava, Dusan Soltes
- Dimension EDI, Chris Nelson
- ISD NHS Scotland, Mary Sweetland
- Israel Central Bureau of Statistics, Ilana Yaacobi
- Mercator CS Ltd, Stephen Jenkins
- NESSTAR, Simon Musgrave
- Run Software-Werkstatt GmbH, Reinhard Karge
- Statistical Office of the Republic of Slovenia, Jozica Klep
- Statistics Finland, Jaakko Ranta
- Swiss Federal Statistical Office, Claude Macchi
- University of Ulster, Sally McClean
- World Systems, Laurent Planq

The ***associate members*** were the following:

- Banca d'Italia: Vincenzo Del Vecchio
- DESAN Research Solutions: Hans:Rutjes
- ECASS, ISER,University of Essex: Marcia Freed Taylor
- European Central Bank: Christos Androvitsaneas
- Health Canada: Bill Bradley
- Istituto Nazionale di Statistica - ITALY: Giovanna :D'Angiolini
- NTUA: Gina Panagopoulou
- ONS (UK): Catherine Ellis
- OpenSurvey: Ed Ross
- UKDA: Hilary:Beedham
- UN Economic Commission for Europe: Jean-Etienne Chappron
- U.S. Bureau of Labor Statistics: Dan Gillman
- Yale University: Ann Green

II. The **expected achievements and impact** of the project have been the following:

- a consensus of the core terminology and structure of statistical metadata models that are agreed
- techniques for developing the core models according to institution specific needs
- recommendations and advice for implementation of coherent metadata models
- a network of individuals committed to maintaining the momentum of developing statistical metadata within an agreed framework

III. The **activities undertaken** in order to achieve the goals:

The WGs co-operated for the integration of the individual results

Work Group 1: methodology and tools

technical aids to implementing metadata systems and exchanging metadata descriptions

Work Group 2: harmonisation of metadata structure and definitions

the conceptual nature of metadata

Work Group 3: Best practice for migration

practical implication of the findings of the first two Work Groups, examples of migration paths etc

Work Group 4: Adoption issues

implications of these findings for the institutions concerned

- Mailing list
  - Web page: http:// www.epros.ed.ac.uk/metanet and follow link to sign up
  - We have 122 subscribers
  - Email adam.taylor@ed.ac.uk for password to:
- Interested parties Website
  - Allows you to upload documents
  - And take part in discussion groups
- WG 1/2 meeting – Vienna October 2001
- All WGs meeting – Stockholm March 2002
- WG 3/4 meeting – Oslo October 2002
- WP 5 formedNot just technology
  - Impact on the organisation
  - Management of change, human issues etc
- Beyond the contract
  - Introduced a new Work Group

- Led by an Associate member

- Contribution to FP6 and other areas

- Forum for exchange of ideas
  - Web sites & meetings

- Interaction with other groups

- Other activities:
  - FP5 projects with metadata component
  - Clusters
  - Technology transfer
  - International initiatives
  - Interest in repositories
  - E-government & digital government
  - Use of administrative record
  - Planning for FP6**Results**
  - ✓ Draft of WG1 output on public website
    - Components split by space
    - Under resources
  - ✓ Draft of WG2 output
    - WP2-no9-Metanet.zip
  - ✓ Draft of WG3 output
    - Currently circulating
  - ✓ WG3/4 questionnaire
    - Report after this meeting**Dissemination activities**

- Eurostat workshop 2001
- Metis 2002
- CASI 2002
- IASSIST 2002
- Compstat 2002
- Open Forum on Metadata repositories, Santa Fe 2003
- Etc

WG1 Achievements

Jean Pierre Kent
Statistics Netherlands, WG1 Leader
jknt@rnd.vb.cbs.nl

The presentation stressed on the ambitions and the objectives of the WG1, the risks of overlap with other WGs, as well as the contributions by partners and members to the related WG1 deliverable.

The main objective of WG1 was to provide "*Technical aids to implementing metadata systems and exchanging metadata descriptions"* and not just to illustrate "Methodology and Tools" of metadata-based systems and projects.

The ambition of the WG1 Leader for the fulfilment of this task was as follows:

- Initial idea
  - Technical aids
    - Metadata standards
      - ISO 11179, Dublin Core, DDI
  - to implementing metadata systems
    - Modelling languages
      - UML
    - Metadata driven programs
      - Blaise
  - and exchanging metadata descriptions
    - Storage and communication languages
      - XML

The ambitions of WG1 as described in Annex 1 and then extended can be summarised as follows:

A representative sample of the population of metadata specialists

- National statistical offices
- National and supranational central banks
- Universities
- Data archives
- Commercial businesses
  - Statistical services
  - ICT services

A microcosm of metadata specialists

- working in different contexts
- with different kinds of data

- having different goals
- needing to represent metadata in different ways
- Eager to understand each other's point of view
- Willing to explain their point of view to each other
- Stop agreeing to disagree
- Give a picture of metadata that satisfies every one of us
- Present the differences in a perspective that turns them into different aspects of the same thing

The main danger was the overlap with mainly WG 2:

- Integrated model of metadata belongs to the task of WG2

WG1 vs WG2

Breadth	Depth
Inventory	System
Past	Future
Descriptive	Prescriptive

- We want to integrate our points of view on metadata
- What are the limits between the two?

The Deliverable of WG1 consisted of a large number of papers on tools, methods, national and internation perspectives. The main chapters were the following:

- Overview (J. Lamb)
- Dimensions of Metadata (ed. J.-P. Kent)
    - Structures of Metadata (W. Grossmann)
    - The Metadata Life Cycle (A. Green & J.-P. Kent)
    - Metadata Usage Level (E. v. Bracht)
    - Metadata Usage Types (J. Lamb)
    - User Functionality (J. Lamb)
- Modelling (ed. A. Westlake)

    - a presentation of aids to modelling and communicating metadata
    - main focus: UML (modelling) and XML (storage and communication)

- Metadata models (ed. J. Bethlehem & J.-P. Kent)
- A catalogue of models:
    - 22 models are presented
    - Not exhaustive
    - Intended to be progressively completed
- General standards (ISO, Dublin Core, DDI) and models (abgt, CWM)
- Tools based on a general model (Cristal, Tadeq)
- Tools based on a specific model (Blaise)

WG2 Achievements: Using the UMAS Proposal as Reference Model - Mapping the DDI to UMAS

Wilfried Grossmann
University of Vienna, WG2 Leader
Wilfried.Grossmann@univie.ac.at

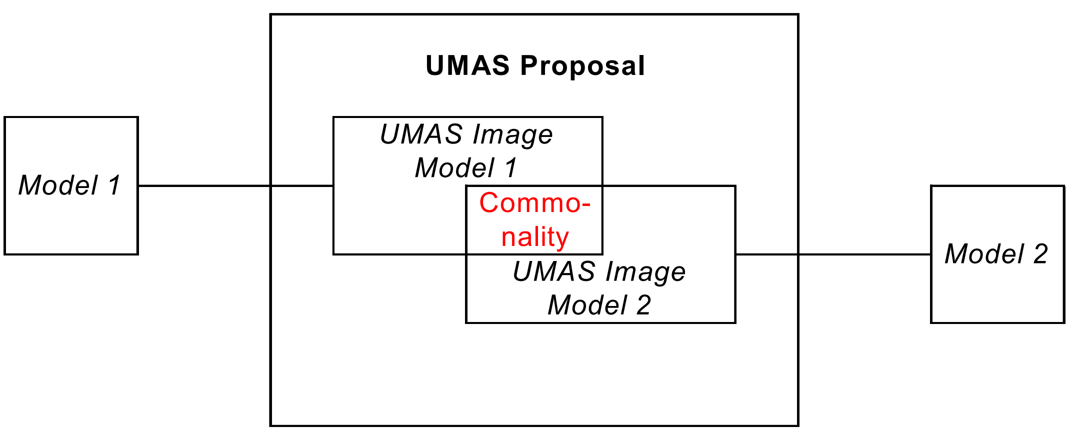
Overview

1. Summary of the UMAS Proposal

Unified **M**etadata **A**rchitecture for **S**tatistics

Goal:

Define a framework to understand ***commonalities and differences*** *of Data / Metadata Models from a* ***statistical point of view,*** *irrespectively of terminology and goal of the specific models* ***This goal can be achieved by mapping each model to the UMAS proposal***



The commonalities of the models are the rational base for harmonisation or reuse of information in *Model 1* in connection with information in *Model 2*

Advantage:

We need not pairs of mappings between models, but only one mapping for each model.

Due to changing requirements on statistics and development of new methodology the UMAS should rely on statistical concepts, which are (up to now) invariant

Basic features of UMAS proposal

All important entities occurring in connection with statistics are considered as instances of categories

All categories are modelled in according to a unified model

All categories occur in a twofold way:

- *As Category Models (CM) either in*
- ⇒ Extensional Format (Metadata)
- ⇒ Intensional Format (Metatext)
- *As Category Instance Data (CI-Data)*

The Category Model describes categories according to four different views:

- *Conceptual Category View* (Conceptual model)
- *Statistical View* (Role of the category within the statistical ontology)
- *Data Management View* (Access and Manipulation of Category Instance Data)
- *Administration View* (Management and bookkeeping of the structures)

Dynamics of categories is characterised by the main steps:

- Definition and design
- Production
- Processing
- Dissemination and Exchange

Information about the dynamics with respect to the different views is given by:

- Category Model (CM)
- Production Blueprint
- Processing Blueprint
- An appropriate view onto: category model, production and processing blueprint

A concrete model can be mapped into the UMAS proposal according to the following steps:

i) Determination of scope of statistical reconstruction of reality

ii) Identification of statistical categories

iii) Specification of category views

iv) Mapping of terminology inside category views

Implicit requirement:

Knowledge of both models in detail

Is this requirement realistic?

See results of the survey of WG 4

2. Details of a Mapping of DDI to UMAS

Note: These tables are a preliminary proposal for a mapping and subject to discussion and approval by DDI authorities

2.1 Scope of Reconstruction

What part of empirical reality is reconstructed in the model?

Table 1: Scope of reconstruction

UMAS-Terminology	DDI-Terminology
Discernible Units	Unit of Analysis
Collectives	Universe
Characteristics	Variable
Characteristic Value	Value / Category
Set of Characteristic Values	Range
Grouped characteristic values	Category Group
Scales	*not explicit*
Measurement Unit (Family)	*not explicit*
Statistical Dataset	Data
Statistical Domain	*not explicit*
Statistical Information System	*not explicit*

Note the special role of Domain and Statistical Information System:

These categories bind together all things objects occurring in connection with a specific area.

Within DDI Section 1 (Document Description) is domain Information

2.2 Identification of Categories

Which categories occur in the model?

In which way are they handled within the model?

Table 2: Identification of Categories

UMAS-Terminology	**DDI-Terminology**	**Usage (main/supplementary)**
Statistical Unit	Unit of Analysis (Analysis Unit)	supplementary
Statistical Population	Universe	supplementary
Statistical Variable	Variable	supplementary/main
Modality		none
Statistical Values	Data Values / Category	supplementary/main
Set of Statistical Values (Range)	Range	supplementary
Scales		none
Measurement Unit (Family)		none
Grouping Level	Category Group	supplementary
Nomenclature		none
Classification	Standard Category (?)	supplementary
Statistical Dataset	Data(file)	main
Statistical Domain		supplementary
Statistical Information System		none

Follow-up presentation for WG2 – Why are metadata important?

Karl Anton Froeschl
University of Vienna, Dept of Statistics
Karl.froeschl@ec3.at

Question: Why are metadata so important?
Simple answer: Because data are important

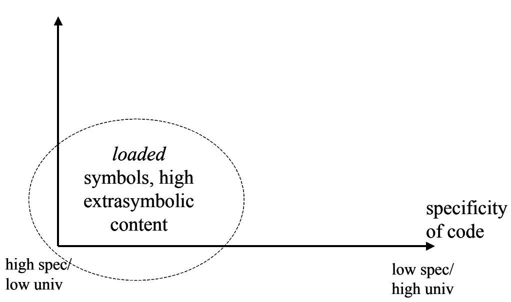
Prelude

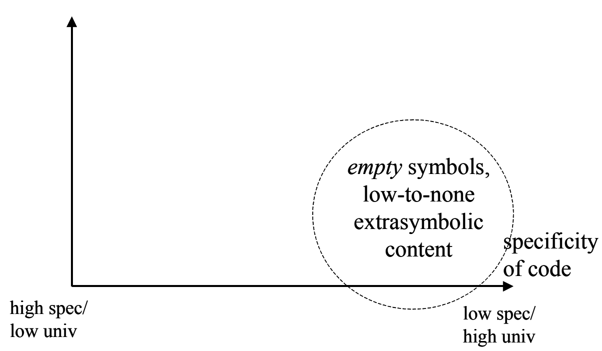
- There is a need for communication, increasingly so!
- **Symbols** are (not only) good for communication!
- We do not use **symbols** arbitrarily, nor do we use arbitrary symbols!
- **Data** is **symbols**!
- **Metadata**, by name, is **data**!

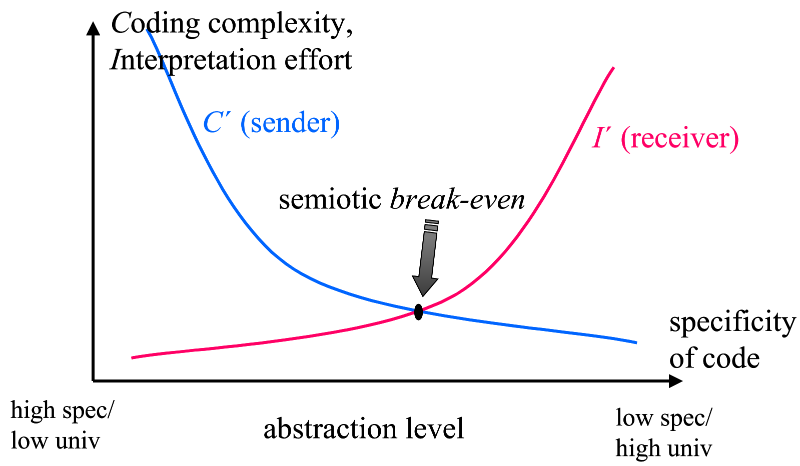
The Economics of Symbol Use

- We talk about communication (meta-communi-cation)
- Here & now, we talk about meta-communication
  - conditions for the interpretation of (symbol-coded) utterances in a –possibly-remote context
  - interpretation in the receiver's context (pragmatics)
  - context ... "self-contained" linguistic domain/world frame to be shared
- *If we don't understand the data, how then can we hope to understand the metadata?*

Make it efficient!







Implications (for us?)

- Efficient coding ... strike a balance between
  - maximally universal language (syntax) ... universal alphabets
  - maximally domain-specific semantics ... meaning preservation
- Optimum ... depending on shared background context of communicators
  - the less context is shared the more the I′-curve shifts to the right
  - the more abstract the symbols the greater the effort in meaning reconstruction

Some Advice?

- Find break-even semiotics, but how?
- Heuristics: structure candidates (empirism?), methodology (normative!)
  - the means: discourse invariants ... the grammar of contexts: cognitive efficiency, by data compression
  - the ends: operation invariants ... operational efficiency: symbols good for both cognitive and physical trans-formations
- Isomorphism: liberal use of any coding systems provided they structurally match

WG3 Achievements

Matthias Abelin
Statistics Sweden
Matthias.Abelin@scb.se

The initial focus of WG3 was strategies for migration from legacy systems to a more integrated statistical metadata system.

The results of the work of WG3 have been a comprehensive manual, which describes the development and implementation of statistical metadata systems, the underlying theories, users needs, processes and best practices.

The presentation mainly focused on ***"Developing and implementing statistical metadata systems",*** covering the following topics:

- Basic concepts
- Users and usages of statistical metadata
- What should statistical metadata inform about?
- How can statistical metadata be obtained?
- Architectures and infrastructure
- Development and implementation strategies

Basic Concepts:

Under the first heading basic concepts such as information, data, metadata, statistics, statistical systems and statistical processes are discussed.

- Information, data, and metadata
- Statistical data and "statistics":
    - Statistics are estimated values of statistical characteristics
    - A statistical characteristic is
        - a statistical measure applied on
        - the values of one or more variables
        - for the objects in a population
    - True values vs estimated values: errors, quality
- Statistical systems
- Statistical processes

Users of statistical metadata

Users are grouped in six main categories:

- Users of statistical outputs
- Producers of statistical outputs
- Software products
- Respondents
- Managers and ”owners”
- Researchers and methodologists

What should statistical metadata inform about?

It is important to understand the complexity, variety and, sometimes, conflicting demands from different users of statistical metadata systems. Software products can also be an important user of metadata, perhaps this is sometimes forgotten in discussions about metadata.

Statistical metadata can be obtained from several different sources:

- Metadata about a specific statistical system
    - metadata about a statistical system’s goals and performance
    - metadata about the contents of a statistical system
    - metadata about processes and tools
- Metadata providing overviews and general knowledge
    - analyses of goals and goal fulfilment
    - overviews of available statistical outputs
    - documented methodological knowledge and experiences

How can statistical metadata be obtained

- Endogeneous sources of statistical metadata
    - design and construction processes
    - operation and monitoring processes
    - usage processes
    - evaluation processes
- Exogeneous sources of statistcal metadata
    - research and development processes
    - evaluation processes

    - compilation processes

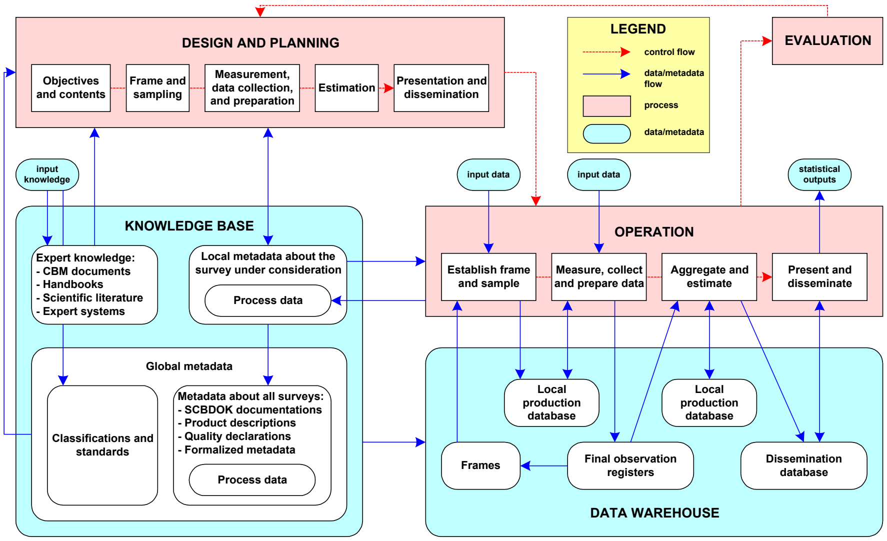
- Autonomous sources of statistical metadata

Architectures and infrastructure

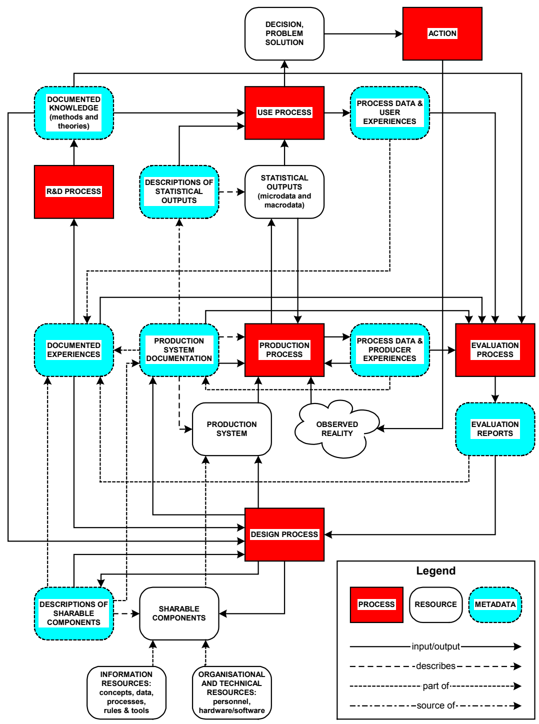
- Major metadata components
- Developing a metadata architecture
- Maintaining the metadata infrastructure

The importance of the design face is stressed in the results of WG3. This face is both an important producer of and consumer of metadata, especially the use of standards. Monitoring and evaluation are also stressed as important sources for improvement of quality and effectiveness. A tighter relation can be created between new methods and the evaluation of these methods through the common use of a metadata system.

The production process perspective is important but also has limitations. The concept of autonomous metadata sources is introduced. These are metadata sources that have to be created and maintained by special efforts dedicated to these tasks, and which are thus not automatically generated as side-effects or by-products of other processes. For example, a classification database.



Metadata flows: sources, usages, feed-back loops



Development and implementation strategy

- Benchmarking – “business intelligence”: study what others are doing.
- Perform the “Who? Why? What? How?” analysis.
- Work out a shared vision of an “ideal” statistical metadata infrastructure.
- Discuss the road towards the vision and set priorities.
- Establish project plans for reaching the goals.

Golden Rules:

Do not miss the golden rules for designers of metadata systems, project coordinators and top management.

If you are a designer...

- Make metadata-related work an integrated part of the business processes of the organisation
- Capture metadata at their natural sources, preferably as by-products of other processes
- Never capture the same metadata twice
- Avoid un-coordinated capturing of similar metadata – build value chains instead (Michael E. Porter)
- Whenever a new metadata need occurs, try to satisfy it by using and transforming existing metadata, possibly enriched by some additional, non-redundant metadata input
- Transform data and accompanying metadata in synchronised, parallel processes, fully automated whenever possible
- Do not forget that metadata have to be updated and maintained, and that old versions may often have to be preserved

If you are the project co-ordinator

- Make sure that there are clearly identified "customers" for all metadata processes, and that all metadata capturing will create value for stakeholders
- Form coalitions around metadata projects
- Make sure that top management is committed; most metadata projects are dependent on constructive co-operation from all parts of the organisation
- Organise the metadata project in such a way that it brings about concrete and useful results at regular and frequent intervals

If you are the top manager...

- Make sure that your organisation 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

- If a metadata project should go wrong – cancel it; don't throw good money after bad money

- When a metadata project fails, make a diagnosis, learn from the mistakes, and do it better next time

- Make sure that your organisation also learns from failures and successes in other statistical organisations

- Make systematic use of metadata systems for capturing and organising tacit knowledge of individual persons in order to make it available to the organisation as a whole and to external users of statistics

WG4 Achievements

Jan Byfuglien
Statistics Norway, WG4 Leader
jan.magnar.byfuglien@ssb.no

Introduction

This introduction was sent to UoA by Jenny Linnerud and serves as an overview of the presentation:

Work group 4 of the MetaNet project should address issues related to adoption of metadata standards and solutions. The idea originally was that this work would build on the output of the preceding work groups of the project, which should bring forward proposals for common solutions and standards. WG4 should identify any barriers of technical, organisational or human nature to the adoption of such common solutions and standards.

As the preceding work groups did not succeed in providing common solutions or standards, but rather accumulated a large amount of information on metadata issues, models and standards, the objectives of work group 4 were reformulated as follows:

- to address issues related to the harmonisation of metadata models and concepts between statistical organisations and between subject matter specialists and other experts
- to seek to identify the major issues in relation to the acceptance and adoption of common data/metadata standards within statistical organisations
- to identify priority areas for further harmonisation and development of common solutions/standards
- to produce a report on these findings
- to produce a training manual

In order to address these issues it was found useful to perform a user survey among a number of statistical organisations, addressing both the organisation as such and a selection of persons within the organisation.

Some of the conclusions of this survey are:

It is documented that metadata is considered in the wide sense and performs different functions within the statistical production process. The NSIs attach high priority to the development of user oriented metadata. The respondents within the NSIs, representing a wide variety of experiences and tasks, seem to support this view, giving user oriented metadata relatively high importance. A conclusion can be that one should accept a wide definition of metadata and focus more on the different functions metadata performs in the statistical production process.

Some solutions for handling metadata are getting rather common in most national statistical organisations (classification database, facts about the statistics), but there is still some work to be done in clarifying the application of different tools and not least to improve the exchange of solutions and experiences.

The limited survey on terminology related to statistical micro data indicates that a fairly large degree of consensus can be reached on key terms such as statistical units/observation unit, (statistical) variable and (statistical) value. Relatively few favour other terms, but there are still some strong opinions on alternative terms and possibly, definitions.

Regarding human issues, the importance afforded metadata creation seems to be low which means that this activity will inevitably suffer at the expense of traditional work aspects. Over 25% of the respondents think that management underestimates resources needed for metadata capture. Thus, if the human 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.

There is a large consensus that the most important requirements at the international level are the need to develop agreed international standards for data/metadata concepts and terms and to agree on common models for handling data/metadata.

Relatively many are aware of and use general tools such as SAS, SPSS, HTML and XML. However, the awareness of many projects/tools/models more specialised for metadata is not widespread among those taking part in this survey - which should comprise a substantial number of persons that are to some extent in the Metanet project. A general problem is that there are so many different approaches/tools with some relationship to metadata that makes it difficult for those other than specialists to get an overview and make use of the possibilities. Thus it is a challenge to focus on some main strategies and solutions and to disseminate these.

In short some conclusions based on the survey and a review of the preceding inputs from the MetaNet project:

- There is no clear consensus on a definition of statistical metadata or a common understanding of the borderlines. Perhaps the concept is becoming too broad and it might be more useful to focus on specific functions.

- There is a gap between some specific metadata models and tools development and the integrated and practical solutions asked for by subject matter specialists. A stronger involvement of the latter group in the development of metadata solutions should be encouraged.

Main presentation

The main points raised in the presentation by Jan Byfuglien about WG4 achievements give an analytical view of the introductory remarks and are as follows:

Original description of tasks of WG 4:

- Consider the implications of these findings (i.e. the findings of the 3 other groups) for the institutions concerned
- Identify potential barriers to adoption – technical, structural and human and consider how these might be overcome
- Its output can be regarded as a training manual for adoption of metadata standards and system

Some main challenges for WG 4:

- The results of the other WGs came relatively late – and there were no agreed metadata models and standards emerging
- A vast amount of documentation provided – but not consistent regarding models, concepts and terms
- The task of trying to build on and synthesise the work of the other WGs for dissemination and adoption not possible given resources and time available!

The objectives and tasks reassessed:

- To address issues related to the harmonisation of metadata models and concepts between organisations and experts
- To seek to identify major issues in relation to acceptance and adoption of common data/metadata standards
- To identify major benefits from adopting common standards and solutions
- To identify priority areas for further harmonisation and development of common solutions/standards
- To produce a training manual

A user survey performed in order to ensure a broader understanding

- A common understanding of what metadata is?
- How well-known are metadata solutions?
- Agreement on basic concepts and terms?
- Problems facing the adoption of metadata solutions?
- Priorities for further tasks for harmonisation and development?

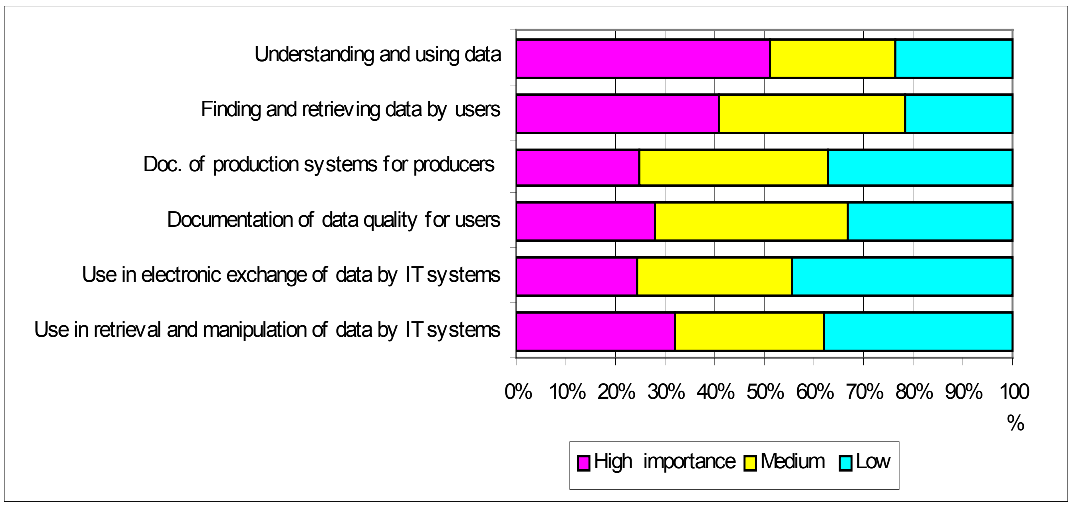
What is metadata and the functions it performs

- Within the MetaNet project there has been different definitions and specifications of what metadata is – from rather narrow to somewhat more general.

- The survey confirms that metadata, for instance within NSIs, is perceived in a broad sense, having different functions – that it is important to focus and clarify.

- Some weight given to user-oriented metadata

The most important functions of metadata: NSIs



Metadata – availability of systems and solutions.

Most NSIs in the survey were well situated concerning:

- classification databases

- facts about statistics for end users

- database of questionnaires and questions

Situation less satisfactory concerning:

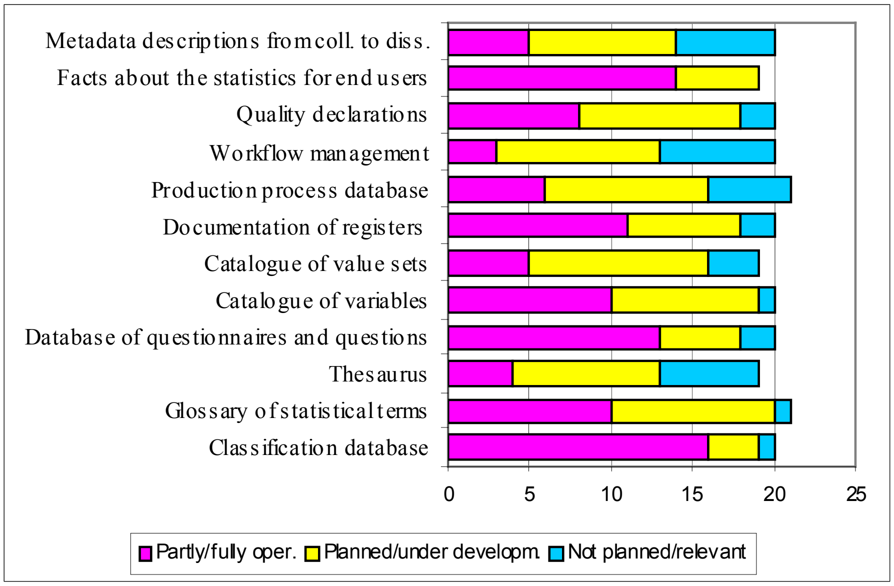
- thesauri

- workflow management

- production process database

Concrete follow up could be to provide better specifications of different systems and solutions – and to improve exchange of experiences and practises

Metadata – availability of systems and solutions. NSIs



USAGE & AWARENESS - Languages & Standards

	In use in my organisation	I am not aware of this
UML	17/35	114/217
XML	34/35	43/217
ISO11179	10/35	127/217
Dublin Core	12/35	112/217

Models and tools – awareness and usage?

Relatively wide usage and well known, general tools:

- SAS - SPSS
- HTML – XML

Relatively widespread among (participating) NSIs:

- Blaise
- PC-Axis
- GESMES/CB

Not well known:

- many specific metadata models and tools

Too many 'buzzwords' related to metadata – improve training on the most important solutions?

Terminology problems

Agreement on some basic concepts and terms?

The following definitions were presented:

- Statistical units are the entities for which information is sought and for which statistics are ultimately compiled. Statistical units may be real world objects (e.g. person, enterprise) or abstract objects (e.g. accidents, transactions).
- Statistical variables are the defined characteristics for the statistical units that are used in the measurement process, e.g. income, sex, age or production volume.
- Statistical values are the concrete result of the measurement process for each statistical variable and statistical unit, e.g. 112 000 Euro, female, 37 years, 3,5 mill. tons. Values can be determined on different measurement levels or modalities and by using a classificatory procedure.Do the definitions related to statistical microdata cover your needs?
    - Statistical unit: 194 yes, 15 no
    - Statistical variables: 189 yes, 11 no
    - Statistical value: 188 yes, 14 no
- Apparently a relatively large consensus in this limited area, but:
    - a certain number are questioning the terms and the definitions
    - the problem of common understanding increases when approaching 'models' for other parts or the 'totality of the statistical production cycle'.Approach harmonisation both from the output side (SDMX?) and from the input side?

Problems facing the adoption of metadata solutions

- Organisational and human issues are considered to pose the greatest challenges in relation to documentation of quality and for the support of retrieval and usage.
- Organisational issues appear to be important to achieve internal documentation of production systems
- No clear opinion on which aspect of 'human issues' are prominent, but many think that priority of time between data and metadata is a major problem
- The confusion on terminology and lack of information on relevant solutions – might be important factors...

Priorities for further internal metadata development

Contact persons of NSIS:

- Highest priority: documentation for users to find, retrieve, understand and use statistical information
- Lower priority: IT related metadata and documentation of production systems for producers

The respondents within all organisations:

- Highest priority: A common strategy for handling data/metadata and a common 'model' from input to output
- Lower priority: improved documentation on micro level and documentation of statistics being distributed

Priorities for development in the international context - NSIs

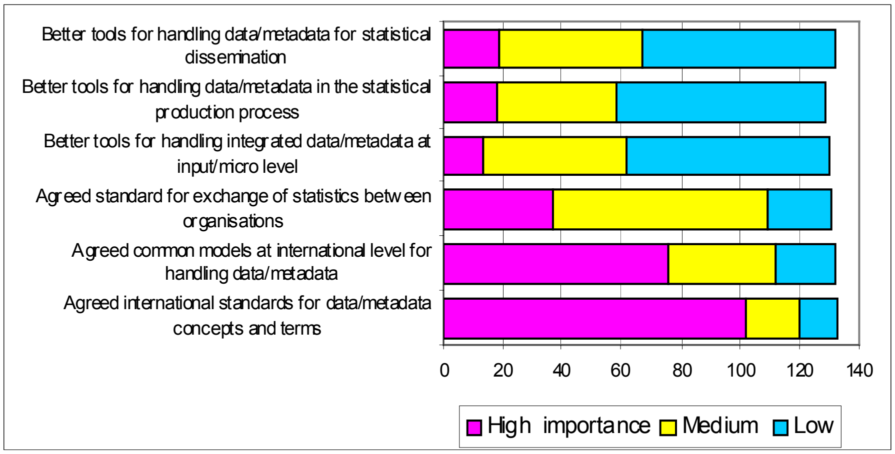
- Top priority:
  - agreed international standards for data/metadata concepts and terms,followed by
  - agreed common models at international level
  - agreed standard for exchange of statistics between organisations
- Lower priority to tools for handling:
  - integrated data/metadata at input/micro-level
  - data/metadata in the statistical production process
  - data/metadata for statistical dissemination

The most important improvements in the international context. NSIs

Some comments from respondents

- Be user oriented and start with simple, practical solutions
- Some metadata projects too abstract – seeking to solve non-existent problems
- Harmonisation and standardisation at international level crucial
- Harmonisation not only for metadata but for the field as a whole

- Common understanding of basic models and structures necessary before seeking to standardise terminology
- Active usage of metadata to drive statistical processing important



Some concluding remarks

- Metadata has for many a diffuse and wide meaning. To be accepted, and rather focus on specifying the different functions it plays in the production process?
- A gap between some theoretical metadata work – seeking to solve meta-problems – and practical statistical production processes?
- Much metadata jargon and specific metadata projects and tools are not widely known and understood
- Still some way to go to approach a common understanding of models/concepts/terms related to data/metadata?
- Many NSIs are actively working on solutions to improve the handling of metadata: Improve specifications of different solutions and the exchange of experiences
- Is a common model, covering all areas of statistics and all parts of the production process feasible? A stepwise approach starting from the dissemination side – and from the input side?
- The development of user oriented metadata important within organisations
- International standardisation and harmonisation expected for data/metadata models, concepts and terms

Finalisation of the report of WG 4

- Take account of the very few comments received
- Take account of any final input from the other WGs and the final conference
- Extend the chapters on technical issues, organisational and human issues in relation to adoption
- Develop the part on training manual as part of preparing the TES training session

How to proceed after MetaNet?

- Improve the understanding of user needs and set clearer priorities for what to achieve
- More focus on data and information processing – from input to output – and less focus on metadata models separated from data
- More targeted actions to achieve specific objectives in focussed working groups
- Networking in order to provide a wide exchange of experience
- More focus on how to achieve consolidated proposals and agreed solutions

WG5 Achievements - METANET TERMINOLOGY

Dusan Soltes
Faculty of Management, Comenius University, Bratislava, Slovakia
dusan.soltes@fm.uniba.sk

Introduction:

- Metadata, metainformation still rather new concepts although their origins go back to Sundgren's research in 1970s
- Classical area of metadata/metainformation: statistics
- First international attempt for a kind of international codification in 1980s under the ECE/UNDP/SCP 1981-84: METIS project

Main Results of the METIS project:

- Guide to metainformation systems in statistics
- Selected chapters on designing metis for statistics
- Metis terminology with an intention to define some basic key meta-related concepts from the area of state statistics

Key Concepts/Recent Developments:

- Data is the physical representation of information reflecting the real world
- Information is the semantic content of data and informs about the phenomena and processes of the real world
- Metadata is the physical representation of metainformation and is a description of data
- Metainformation is the semantic content of metadata and informs about data/information
- In 1990s came a new revival and a real "boom" in the development of metadata and metainformation
- Main forces: internet, www, datawarehouses, data mining …
- Metadata & metainformation became absolutely inevitable for any efficient utilization of all above new problem areas

Metadata models, tools, methods, structures, systems

- ISO/IEC 11179	- SIDI/SDOSIS	- INFOLOGIC	- DATALOGIC
- UML	- METIS	- MOM	- MIMAMED
- MDA	- GESMES	- ISTAT	- MAMEOB
- XML/XMI	- NESSTAR	- IQML	- CRISTAL
- DDI	- CLAMOUR	- IDARESA	- FASTER
- CWM	- METANET	- CMR	- DUBLIN CORE
- ESPLORIS	- MOF	- GESMES	- TRIPLE S
- SDSS-GDDS (IMF)	- NEUCHATEL GROUP		
- CORBA	- SGML/XML	- CODAM	, etc

All the above models, tools, methods, structures and systems have among various other aspects brought also the whole range of terminologically related issues and mainly differentiation.

SOME EXAMPLES:
ISO/IEC 11179- PARTS 2, 3, 4, 5, 6

- FOR DATA ELEMENTS AS A MINIMUM IS NEEDED:

- OBJECT CLASS	- ATTRIBUTES
- PROPERTY	- QUALIFIERS
- REPRESENTATION	- PROPERTIES
- CONCEPT	- BASIC ELEMENTS
- NAME	- SYNONYMS/HOMONYMS
- IDENTIFIER	- REGISTRATION
- DEFINITION	- DOMAINS
- SCHEMES	- RELATIONSHIPS

DUBLIN CORE

- DUBLIN CORE IS A SET OF 16 ELEMENTS (6 OF THEM COMPLY WITH ISO/IEC 11179-3 1994) :

- COMMON ATTRIBUTE	- TYPE
- TITLE	- FORMAT
- CREATOR	- IDENTIFIER
- SUBJECT	- SOURCE
- DESCRIPTION	- LANGUAGE
- PUBLISHER	- RELATION
- CONTRIBUTOR	- COVERAGE
- DATE	- RIGHTS

Metadata terminology: current situation

In different glossaries also such terms as:

- account
- accounting conventions
- census
- contact
- industry
- organizations
- version

NEUCHATEL TERMINOLOGY ON CLASSIFICATIONS

- CLASSIFICATION
- CLASS. FAMILY
- CLASS. VERSION
- CLASS. LEVEL
- CLASS. ITEM
- EXCLUSION REF.
- CASE LAW
- CLASS. INDEX

CLASS. INDX ENTRY
ITEM HISTORY ENTRY
CORRESPONDENCE TABLE
CORRESPONDENCE ITEM
GROUPING
FOOTNOTE

Different meaning in different sources (iso, un/ece, eurostat, un/classif., imf, ecb, …):

- data
- data dictionary
- data element
- definition
- observation unit
- register
- survey
- term

What is needed? Types of metadata

- Document oriented
- Object oriented

- Macro
- Micro
- Local
- Global
- Classification
- Conceptual
- Operational
- Organizational
- Semantical
- Technical
- Infological
- Datalogical
- Discovery
- Control
- Process
- Contextual
- “Deep” metadata [“shallow”]
- Enhanced
- Survey
- e-metadata

Metadata Holdings

- catalogue
- directory
- glossary
- thesaurus
- register
- classification
- code-list
- nomenclature
- metadata base

METADATA OBJECTS (STATISTICS)

- INDICATORS/ITEMS
- FORMS/SURVEYS
- PUBLICATIONS
- FILES
- OUTPUT TABLES
- TIME-SERIES
- ALGORITHMS
- STATISTICAL POPULATIONS
- STATISTICAL UNITS
- CLASSIFICATIONS
- NOMENCLATURES
- CODE-LISTS
- PHENOMENAS

METADATA FUNCTIONS

- INFORMATION
- DOCUMENTATION
- INTERPRETATION
- NAVIGATION
- SERCHING
- ORIENTATION
- RETRIEVAL
- PROCESSING
- PRESENTATION
- UTILIZATION
- INTEGRATION
- COMPARISON
- RATIONALIZATION
- “MINING”

About WG5 and its achievements:

The more WG 1-4 have been proceeding in their work areas the more it was clear that terminology becomes one of most important issues:

For Metanet internal communication

For future users

For further research and development

- new cross-sectional interest group established at Oslo meeting October 2002
- open group
- two approaches to terminology:

Authors responsibility to attach a list of used terms and their definitions on a standard template and their judging by the WG5

Self-searching by the WG5

Goals:

- Standardization and unification of Metanet terminology
- terminological annex for every final product of the Metanet
- a separate summary final dictionary/glossary/thesaurus of Metanet terminology with definitions and source references of all key meta-related terms

METANET TERMINOLOGY TEMPLATE

- - IDENTIFIER OF A TERM
- - NAME
- - DEFINITION
- - SOURCE/REFERENCES
- - AUTHOR/CONTACT
- - DATE
- - COMMENTS (HISTORY, SPECIFICATIONS, …)

WHAT HAS BEEN COMPLETED:

- - TEMPLATE
- - INVENTORY OF EXISTING META-TERMS
- - ANALYSIS OF EXISTING TERMS
- - INITIAL SCOPE OF TERMINOLOGY WILL BE WITHIN 4 WG FINAL MANUALS

METANET TERMINOLOGY: EXPECTED RESULTS

- - A BASIC THESAURUS/GLOSSARY OF TERMS
- THAT ENCOMPASSES BROAD TERMS SUCH AS IN ISO/IEC 11179
- - LINKS TO MORE SPECIALIZED STATISTICAL METADATA
- - UPDATES OF THE UN/ECE STATISTICAL UN/ECE METADATA TERMINOLOGY
- - IMPLEMENTED AS A PRINTED GLOSARY AND METANET WEB FILE

WHAT IS NEEDED:

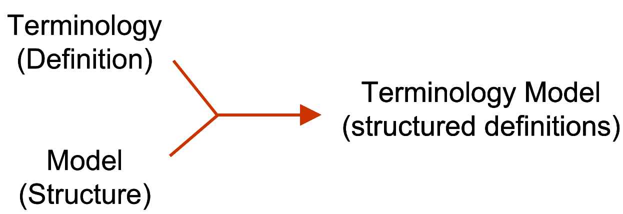
- - MORE TIME
- - SOME FRAMEWORK AFTER METANET WILL BE COMPLETED ON 31 JULY 2003
- - MORE ACTIVE COOPERATION FROM 4 WGs
- - AS E-WORK NOT ENOUGH WE NEED ONE EXTRA SPECIAL MEETING ON TERMINOLOGY
- - FINAL VERSIONS OF 4 FINAL WGs MANUALS
- - SOME FUNDING INCLUDING PUBLISHING COSTS

MetaNet Reference Model

Reinhard Karge
Run Software - Werkstatt GmbH, Germany
reinhard.karge@run-software.com

The Reference model developed in the Metanet project was presented, as well as the overall concept of the structure and harmonisation of metadata considered when developing the terminology model. In addition, parts of the model were presented (conceptual and xml, html format). In the proceedings we mainly concentrate on the overall concept of the structure and harmonisation of metadata considered through the framework of the project, since the analysis and documentation of the Reference Model is a separate deliverable of Metanet. Only a brief overview will be given here.

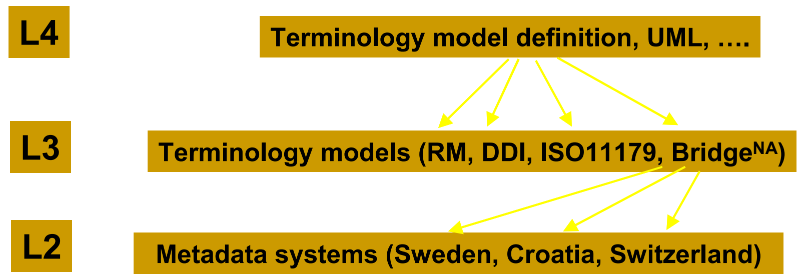
This presentation would be better considered in relation to the presentation of WG2 achievements by Prof Grossmann as well by the presentation by Mr del Vecchio about a model's levels given in chapter 3.



Core level terminology model (L4)

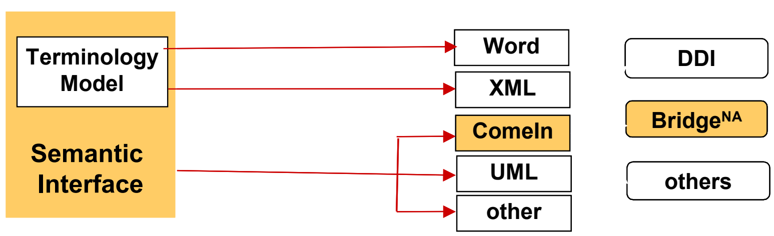
- **Concept** The concept defines a term that refers to a basic idea in a subject matter area. The names for concepts and synonyms must be unique in a terminology model. The concept is defined by:
  - **Name** The name is a single word or group of words that identifies the concept.
  - **Description** A description or definition of the named concept
  - **Characteristic** List of characteristics that describe the details of a concept. Characteristics are defined as <u>Characteristic</u>.
  - **Synonym** List of synonyms that can be used instead of the concept name.
- **Characteristic** The characteristic of a concept defines a relevant detail (attribute) of a concept. The names and synonyms for characteristics must be unique within a concept definition. The characteristic is defined by:
  - **Name** Single word or group of words that identifies the characteristic
  - **Description** The description or definition of the named characteristic.
  - **Related concept** If the characteristic is not simply defined as text but refers to another <u>concept</u> the referenced concept is mentioned here. In documents referenced concepts are visualized as underlined terms.
  - **Synonym** List of synonyms that can be used for the characteristic.
- Extended TM for describing active metadata (metadata behavior)

Reusability



Model Transformation

- Agreements required for interfaces
    - Technical agreement (XML, COM, ...)
    - Semantic agreement (semantic standard)
- Using the same semantic standard for different technical interfaces makes communication easier
- Metadata exchange is one example for a passive communication interface



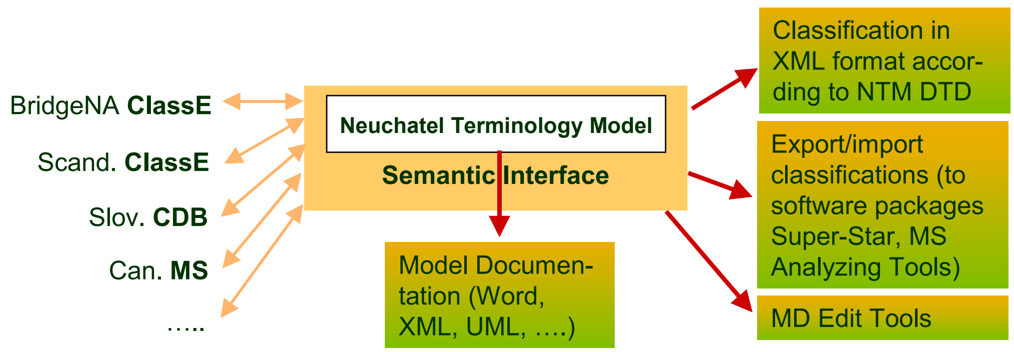
Neuchatel Experience

Neuchatel terminology model for classifications

- Developed with the purpose of providing a conceptual definition of concepts (objects) and characteristics (properties)
- Developed with the purpose of building a classification server based on the agreed terminology

Applications build based on the Neuchatel Terminology Model

- Bridge metadata system
- Danish/Norwegian classification server
- Slovenian classification database
- Canadian metadata system
- (Portuguese Classification server)



- ONS Metadata system

Developing a terminology model

<u>What do we need?</u>

–A number of subject matter experts

–Two or three model experts without subject matter knowledge

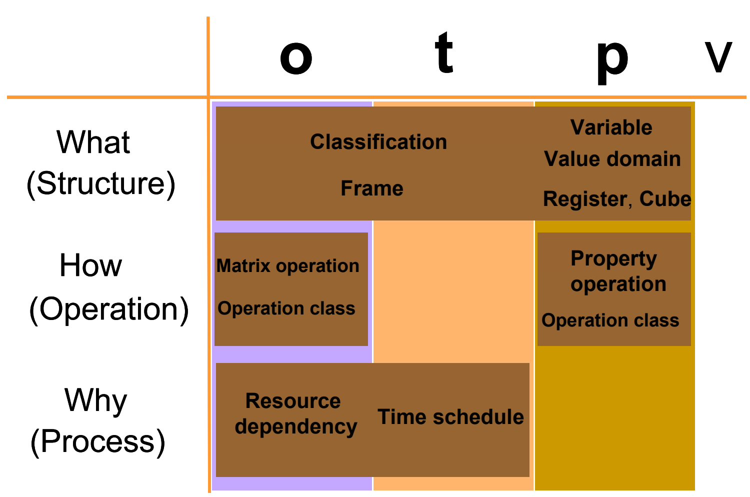
<u>How do we work?</u>

1)Agreeing upon a key concept (e.g. classification)

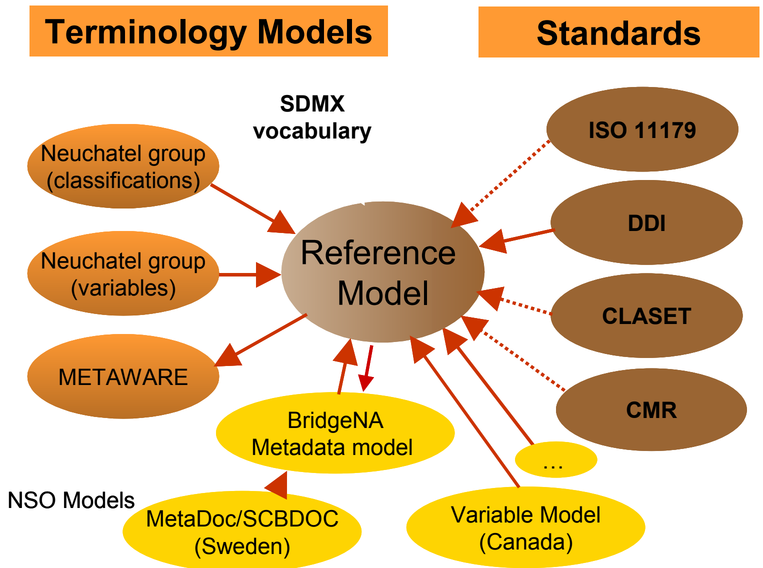
2)Model experts ask subject matter experts for a definition of the concept

3)Model experts ask subject matter experts for specific characteristics describing the concept

4)Model experts try to figure out, whether a characteristic refers to another concept that requires a definition, i.e. to a concept that has characteristics. If yes, continue with step 2. If no, the characteristic can be described by a number or plain text.



Reference Model



What has been done

- A first version of the METANET reference model has been provided
  - A method has been provided for transforming a model (metadata model) into a semantic model (1:1 mapping)
  - A method and tools for mapping semantic models to the reference model has been provided
- The reference model has been provided as Word document, XML glossary and HTML documents
- About 10 of metadata models and standards have been investigated
  - About 100 terminology models with more than 1000 characteristics have been defined
  - About 90% of the concepts and 80% of the characteristics are of common interest and part of the reference model
  - More than 50% of the missing concepts and characteristics are of common interest and should be added to the reference model
- The reference model covers all metadata areas according to different metadata classifications discussed in the project
  - Nevertheless, the reference model cannot be considered as completed, since no of the areas is complete and other classifications might be introduced
- The reference model provides a tool for exchanging and harmonizing metadata

What is left to be done The present mappings provided with the METANET reference model can only be considered as draft

- Just a few models (METAWARE, DDI, ComeIn, ISO 11197) have been mapped partially
- DDI and ISO11179 refer to old versions

- Updating the reference model and providing mappings for the most important standards and models is a continuous task, that requires an authority and resources

- Only a number of interested metadata providers is able to guarantee consistency and quality of the reference model in future

- Having the mapping completed provides a good base for metadata exchange
  - Tools must be provided for metadata exchange
  - Improved tools are necessary for documentation and presentation of the reference model

▶ We are still in the collection phase

▶ We consider a restructuring phase in 2005

▶ After a second collection phase we expect another restructuring of the RM about 2007

METANET Reference Model (TM) – Overview

(as presented in the first relates web page)

Select concept for more details:

- Access Conditions
- Access Package
- Account Process
- Activity Family
- Alternative Item Title
- Basic Measure Unit
- Case Law
- Citation
- Classification
- Classification Family
- Classification Index
- Classification Index Entry
- Classification Item
- Classification Item Search Word
- Classification Level
- Classification Variant
- Classification Version
- Classifying Variable
- Computation Implementation
- Concept
- Conceptual Data Element
- Conceptual Value Domain

- Contact Person
- Contact Person Reference
- Context Data Element
- Correspondence Item
- Correspondence Table
- Coverage Type
- Cube
- Cube Data Element
- Data Collection
- Data Collection Element
- Data Collection Methodology
- Data Description
- Data Element
- Data Element Reference
- Data Item
- Data Quality
- Data Set
- Derivation Rule
- Derivation Rule Class
- Dimension Reference
- Exclusion Reference
- Final Observation Register
- Financing Type
- Footnote
- Global Variable
- Hyper Cube
- Identifier
- Item History Entry
- Keyword
- Label
- Language
- Language Reference
- Link Variable
- Maintenance Unit
- Matrix
- Matrix Operation
- Matrix Operation Class
- Matrix Reference
- Measure Unit
- Measure Unit Type
- Metadata Object
- Milestone
- Object Variable
- Operation Implementation
- Organisation
- Organisation Reference
- Periodicity Type
- Person
- Plausibility Rule
- Population
- Process Implementation
- Publication
- Qualitative Variable
- Quantifying Variable
- Quantitative Variable
- Question

- Questionnaire
- Record Reference
- Record Type
- Record Variable
- Register
- Register Data Element
- Software
- Statistical Object Type
- Statistical Process
- Study
- Study Method
- Study Version
- Subject Area
- Synonym
- Table
- Table Instance
- Table Reference
- Term
- Theme
- Time Period
- Title Type
- Update Information
- User Group
- Value Domain
- Value Domain Item
- Variable Family
- Variable Reference
- Version List Entry
- Word
- Description Template
- Dynamic Description

EXPERIENCES OF METADATA RELATED PROJECTS

The COSMOS project

(Cluster Of Systems of Metadata for Official Statistics, IST-2000-26050)

Joanne Lamb,
University of Edinburgh, Coordinator of the COSMOS project
J.M.Lamb@ed.ac.uk

The main points addressed were:

- Clustering projects and partners
- Objectives and work plan
- Key events

The overview of the presentation is as follows:

COSMOS is an accompanying measure (Cluster) of five projects of the European Union Framework 5 research programme. Its objectives are fourfold:

- to build better metadata repositories by exchanging ideas and experiences in using metadata systems for the individual projects
- to identify a common set of metadata objects, with agreed definitions, attributes and methods
- to implement a demonstration subset of these objects to show interoperability of the developed systems
- to define a methodology for further developing this interoperability.

To achieve these objectives, the project aims to:

i) maximise the interchange between the key developers of the metadata models in each project

ii) give an opportunity for all participating institutions to meet and exchange experiences

iii) demonstrate the interoperability of the project outputs, while respecting the exigencies of each project plan

The projects involved and the partners are illustrated in the following two tables:

Project	Full title	Ref number
FASTER	Flexible Access to Statistics, Tables and Electronic Resources	IST-1999-11791
IPIS	Integration of Public Information Systems and Statistical Services	IST-1999-12272
IQML	A Software Suite and Extended Markup Language (XML) Standard for Intelligent Questionnaires	IST-1999-10338
METAWARE	Statistical Metadata Support for Data Warehouses	IST-1999-12583
MISSION	Multi-agent Integration of Shared Statistical Information over the (inter)Net	IST-1999-10655

COSMOS Cluster Projects

Participant name	Country	Projects
University of Edinburgh	UK	IQML, Mission
University of Essex	UK	Faster
University of Athens	EL	IPIS, Mission
Desan Marktonderzoek BV	NL	IQML, Mission
Statistics Sweden	S	Metaware
University of Ulster	UK	Mission
World Systems (Europe) Limited	L	Metaware
Dimension EDI	UK	IQML

Participants according to clustering projects involved in COSMOS

The project's **workplan** is highlighted as follows:

- There are five elements of the Cluster's Workplan:
  - » Bring together the key designers of each project
    - develop a common understanding of a common framework
  - » Have a meeting at which all partners of all projects are present
    - identify which elements of the core model will be implemented
  - » Two Workpackages proceed in parallel
    - developing the agreed APIs
    - defining the trial scenarios
  - » Final meeting
    - demonstration of the trial scenario
    - evaluation report of the trial and impact of the Cluster's activities on the projects
  - » Interaction with other networking and standardising activities

The **key evens** to date are the following:

- Initial meeting November 2001
- Main conference 1-3 May 2002
- Establishment of technical workgroup
- Exchange of knowledge and expertise on metadata
- Establishment of strategic workgroup
- FINAL CONFERENCE: 17-19TH JUNE 2003

Key technical group developments:

- Working towards correspondence of definitions in different metadata systems...
- leading towards development of an object-oriented common conceptual model for statistics...
- and subsequently to a demonstration of interoperability using a subset of the model
- progress towards a metadata registry to serve as a single entry point that potentially allows searches of all the metadata in all the Cosmos projectsdevelopment of a common technical architecture for the publishing of statistics on the Internet

Other progress:

- exchange and transfer of knowledge and expertise in the field of metadata
- raised awareness of the different needs of different users for a variety of outputs
- Establishment of a strategy group to:

- identify, document and discuss issues that need further consideration and resources
- suggest areas for future work
- produce a formal report with recommendations

Issues already identified:

- requirement for more detailed information about user needs
- further development of the common model to include those areas that will not be developed as part of the Cosmos demonstration
- assurance of data quality in technical systems
- data availability - access control & SDC
- organisational issues
- knowledge transfer

Potential use

- Potential gain for all the projects in considering and, where appropriate, adopting, innovation & solutions from the cluster
- Faster model is being revised in line with with the Cosmos common model
- We anticipate widespread application of the common model and further developments in interoperability as the work of the group reaches completion and is disseminated to external organisations

The CODACMOS project

(Cluster of data collection integration and metadata systems for official statistics, IST - 2001 – 38636)

Alberto Sorce, ISTAT, Coordinator of CODACMOS project

sorce@istat.it

The main points discussed were the following:

- Objectives and aims
- Consortium description
- Expectations
- Main results

The overview of the presentation is as follows:

General description:

CODACMOS is a cluster of RTD projects that directly or indirectly deal with data collection methods and models. As a starting point will be the projects: TELER, DATAMED, IQML and IPIS. Other current R&D projects DEALING WITH DATA COLLECTION AND METADATA EXCHANGE will be taken into consideration

CODACMOS will review, integrate, adapt and enhance the findings of the projects with the aim to improve the efficiency of data collection by NSIs, Administrations and other institutes, i.e. to increase the speed and ease and decrease the costs for both data collectors and providers while raising the quality of the resulting data.

By bringing together the work and the key researchers from the cluster projects, CODACMOS will add value to their concepts, models, tools and solutions for data collection, by studying and experimenting in the field the solutions for the enterprises and the households.

Main objectives:

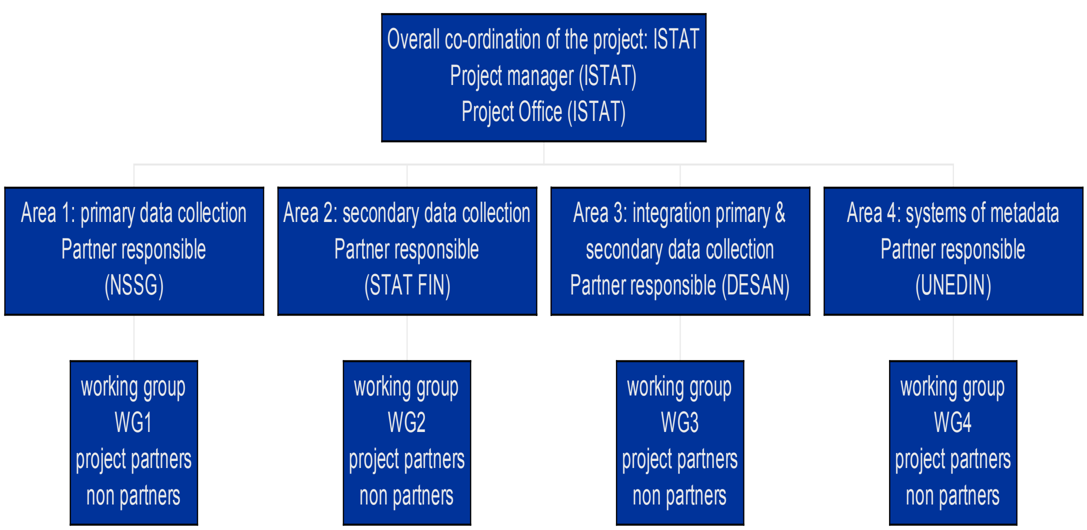
- To review or to rationalise the state of art and the development of current solutions given by a cluster of relevant projects that facilitate the data collection and exchange from the respondents;
- To specify EU key issues for the standardisation/harmonisation of data collection models and methods for the description of metadata standard
- To optimise the use of existing archives/registers or other administrative data for statistical purposes based in a close co-operation between institutions and administrations by exchanging experience and ideas in using the integration of different data sources for the cluster of projects;

- To identify the "experimental field" areas and to implement demonstrations of the solutions (or models) proposed

- The integration of primary and secondary EDI by:
    - Review and analyse the underlying models and the existing situation
    - Develop a core model, based on the input of the previous tasks
    - Prepare best guidelines and examples in the field of primary and secondary data integration
    - Make proposals for the development of new techniques for the optimisation of primary and secondary data integration

- CODACMOS will play a major role in
    - Strengthening the European Statistical System via:
    - The 1st asset of the project:
    - Importance and relevance of the subject
    - The 2nd asset of the project:
    - Quality and width of the consortium
    - The 3rd asset of the project:
    - Possibility of preparing a valuable FP6 project

Consortium description



Structure of the Consortium:

- Statistical Offices (5)
- Universities (3)
- Non-statistical data collectors (2)
- Commercial companies (2)
- Research institutes (1)

Trials reported for each clustering project on data collection issues:

DATAMED

	Trial 1	Trial 2	Trial 3
Used methods	CAWI, CATI, CASI	CAWI, Web-TV	CAWI, CAPI (mobile)
Type of data collection	Primary EDI	Primary EDI, Secondary EDI, Combination	Primary EDI, Secondary EDI
Used technologies	SPSS MR Software Suite	SPSS MR Software Suite, Web-TV interviewing software	SPSS MR Software Suite, Mobile telephone interviewing
Statistical domain covered	Business statistics, Social Statistics	Business statistics, Social Statistics	Business statistics, Social Statistics
Type of respondents	Households, Very small enterprises	Households, Very small enterprises	Households, Very small enterprises
Countries covered	GR	IT	PT

TELER

	Trial 1
Used methods	Internet based data transfer
Type of data collection	Primary EDI
Used technologies	Data extraction software package (automatic mapping)
Statistical domain covered	Economic area (financial and accounting information used for statistical purposes-balance sheets data)
Type of respondents	Enterprises and intermediaries (accounting firms)
Countries covered	IT, NL, FI, etc.

IQML

	Trial 1	Trial 2	Trial 3	Trial 4
Used methods	Internet based Survey	Internet based Survey	Internet based Survey	Internet based Survey
Type of data collection	Primary EDI	Primary EDI	Primary EDI	primary
Used technologies	PxQ (precursor to QPT)	QDT / QPT / SAT	Repository / QPT / DIT	Repository / QDT / QPT / SAT
Statistical domain covered	N/A	Economic, Social Statistics, Education	Economic, Social Statistics	Business statistics
Type of respondents	Internal users	Employees, Enterprises, Students	Internal users, Enterprises	Internal users, Enterprises
Countries covered	NO	NO, IE, GR	NO, IE	NO, IE, GR

IPIS

	Trial 1	Trial 2	Trial 3	Trial 4
Used methods	Survey data	Survey data and data from administrative sources	Administrative source	Survey data
Type of data collection	Secondary EDI	Secondary EDI	Secondary EDI	Secondary EDI
Used technologies	Internet	Internet	Internet	Internet
Statistical domain covered	N/A	Economic, Social Statistics, Education	Economic, Social Statistics	Business statistics
Type of respondents	Internal users	Employees, Enterprises, Students	Internal users, Enterprises	Internal users, Enterprises
Countries covered	NO	NO, IE, GR	NO, IE	NO, IE, GR

Categorisation According to Type of Respondents

Type of Respondents	Relative project	Solution Experimented	EDI Category
Very Small Enterprises	DATAMED, TELER, E-QUEST	CAWI, CATI, CASI, Mobile CAPI, Web-TV	Primary
Small and Medium Enterprises	IQML, TELER, E-QUEST	Web, E-mail survey	Primary
Large Enterprises	IQML, TELER, E-QUEST	Web, E-mail survey	Primary
Intermediaries and Accounting Offices	TELER, E-QUEST	Internet based data collection	Primary
Households	DATAMED	CAWI, CATI, CASI, Mobile CAPI, Web-TV	Primary
Individuals	DATAMED, IQML	CAWI, CATI, CASI, Mobile CAPI, Web-TV, e-mail survey	Primary
National Professional Associations	TELER	Internet based data collection	Primary, Secondary, integration
Local and National Public Administrations	IQML, IPIS, TYVI	Internet based data collection	Primary, Secondary, Integration

Criteria for the selection of experimental fields and the design of demonstrations

- Value
- Coherence
- Viability
- Precision
- Spread
- Coverage by clustered projects' trials
- Other criteria

The data collection and reporting concepts and methods- The concepts of direct extraction of data

- Automatic mapping
- The information stored somewhere
- Direct data extraction of files from administration or enterprises information system
- Data extraction package
- The computer assisted interviewing concept (like CAI methods)

- Data collection and exchange via Internet
- Data interchange for importing and exporting data and metadata
- The concept of Government e-link
- The concept of integration of various data sources in order to collect data for statistical purposes
- Scanning and automated data entry.

Considerations

- In a regular electronic format, the use of electronic questionnaire becomes more important only for data collectors
- The use of web based data collection becomes an important objective to avoid eventual double reporting and to reduce burden as a whole
- From respondent point of view the web forms and tools seem to be easy acceptable (no software installation is required), but some respondents don't trust internet applications and dial-up connections are too expensive
- From data collectors point of view 2 main advantages:
- The simple deployment of electronic questionnaires
- The possibility of design of the questionnaires by subject matter statisticians

E-Quest: a metadata based system for electronic raw dates collection.
Main feature.
•electronic questionnaires must benefit the respondents, or else the respondents will have no incentive to use them;
•e-Quest it has to be usable for different survey, including the more complicated economic surveys;
•it must to be possible to change an electronic questionnaire without modifying the source code of software;
•.e-quest has to be able of to be wide;
•it must be gives high priority to the safety of the data;
•for the surveys in the economic field the software has to be practicable for enterprises of all the dimensions as for intermediaries (id est. accountant firms) who are empowered to fill in the questionnaire to the place of their clients;
•present and future foreseeable IT trends should be taken into account;
•e-Quest it has to be usable not only to the respondents but also to the institute of statistic;

The AMRADS project
(Accompanying measure for R&D in Statistics)

Karen Barrie, University of Edinburgh
karen.barrie@ed.ac.uk

The AMRADS project is an accompanying measure to CPA 8 of the Information Society Programme. It is designed to create the conditions for facilitating technology and know-how transfer of the results of research projects of the European Programme of Research in Official Statistics.

The presentation mainly focused on the following topics:

- Describing the background to the project as a whole
- Detailing the activities of the Metadata Theme

E-Service (TYVI)

the first data collector using internet has been implemented using the model TYVI (of dates flow from enterprises to authorities). in which an external operator is responsible of the service of capture and validation of the questionnaires.
The supplier of the data capture the questionnaire to be compiled through an web application. The captured data are loaded in a database of ownership of the operator. The interested administration can receive the information through a protocol of exchange (as the Ftp). The project was born as a cooperative plan in which the financial supervision authority, the bank of Finland and Statistics Finland have merged into know-called Virati collection. It deals all the data that it captures from the credit institutions and investment service enterprises so that every authority can extract the data it requires from the combined bulk of data. Statistics of Finland has co-ordinated the data content, and designed and produced the workbooks to be filled in.

- Summarising the key findings
- Drawing conclusions for the future

Objectives of the project

- To support the transfer of technology and know-how
- From research in FP5 projects and elsewhere

- To normal working practice within statistical agencies

Participants were presented according to the topic of their involvement in the project:

- Statistical Disclosure Control - Anco Hundepool, Statistics Netherlands
- Quality - John Charlton, Office for National Statistics, UK
- Business Registers - Seppo Laaksonen, Statistics Finland
- Automated Data Capture (CAI, EDI) - Alberto Sorce, ISTAT, Italy
- Timeseries - Senn Lanfranco, University of Bocconi, Italy
- Metadata - Joanne Lamb, University of Edinburgh

Work Plan:

- Partners study the field of their theme
- The NTTS/ETK conference in Crete June 2001
- Partners form virtual working groups
- A feasibility questionnaire is sent to NSIs
- Partners hold discussion workshops on their selected theme
- Partners visit selected NSIs to discuss their TTK needs
- A training session is held for each theme
- A final conference, to be held in Rome in September 2003

Workshop training program

Session 1 **What is Metadata?** 10:00 Introduction 10:15 Group Discussions 12:00 Feedback 12:45 Lunch	**Session 2** **The Vision** 14:00 The OECD Vision 14:20 The SORS Vision 14:40 Open Discussion 15:20 Summing Up 15:30 Coffee	**Session 3** **Planning & Persuading** 16:00 Introduction 16:20 The SORS Story 16:40 Open Discussion 17:20 Summing Up 17:30 Close for the day
Session 4 **From Plans to Projects** 09:00 Introduction 09:15 The UK ONS Experience 09:45 Group Discussions 11:45 Feedback 12:30 Lunch	**Session 5** **Classifications** 14:00 Expert Presentation (SSB) 14:45 Discussant 14:55 Open Discussion 15:25 Summing Up 15:30 Coffee	**Session 6** **Terminology & Metadata** 16:00 Expert Presentation (BLS) 16:45 Discussant 16:55 Open Discussion 17:25 Summing Up 17:30 Close for the day
Session 7 **Data Capture** 09:00 Expert Presentation (CBS) 09:45 Discussant 09:55 Open Discussion 10:25 Summing Up 10:30 Coffee	**Session 8** **Data Dissemination** 11:00 Expert Presentation (UKDA) 11:45 Discussant 11:55 Open Discussion 12:25 Summing Up 12:30 Lunch	**Session 9** **Implications for Users** 14:00 Introduction 14:20 Group Discussions 16:15 Feedback 17:00 Conclusions 17:30 Workshop Close

Key Findings ~ Transfer of Know-How

a) Classification

- Described the background and functionality of the SSB classification server
- Numerous specific functional, organisational and human issues addressed
- Excellent example of achievement via collaborative working

b) Terminology

- Conveyed the purpose and means of managing terminology
- Described the inter-relationships between terminology and metadata
- Highlighted the level of complexity and the implications of trade-offs

c) Data Capture

- Contemplated the implications of defining metadata as:
- “Data needed to understand the data requested” Vs provided
- Highlighted the importance and role of open standards
- Discovered that the concept of “process metadata” was new to many

d) Data Dissemination

- Discussed the role of metadata in countering dissemination constraints
- Considered the feasibility and desirability of harmonisationStressed the importance of understanding users’ identities and needs

Key Findings ~ Exchange of Experiences

- Experiences on metadata definitions, metadata types and uses
- Similarities identified between international position and that of national statistics in the context of e-government initiatives
- Standardisation, harmonisation and co-ordination are the founding tenets
- Organisational issues
- Implementation issues

- User implications, i.e session sought to identify the range of activities being conducted to gain greater understanding of user identities, behaviours and needs

Initiatives such as AMRADS facilitate:

- Transfer of metadata know-how and ‘best’ practice
- Exchange of experiences & sustained networking
- Uncovering of key issues and concerns

NATIONAL EXPERIENCES (NSIS EXPERIENCES AND OTHER NATIONAL ORGANISATIONS)

This section include presentations given by representatives of National Statistical Institutes (NSIs) or other national organisations and they reflect national perspectives and experiences in the area of metadata, metadata-based systems and standards used.

Presentations on national experiences have been the following:

i) "*The ONS experience including metadata*" by Jan Thomas, Office for National Statistics, UK

ii) "*The Banca d' Italia experience"* by Vincenzo Del Vecchio, Banca d' Italia

iii) "*The SORS experience"* by Josa Klep, Statistical Office of the Republic of Slovenia"*Experiences in the Latvian Statistical Office*", by Karlis Zeila, CSB of Latvia"*The CODAM implementation exercise*", by Claude Macchi, Swiss Federal Statistical Office

vi) "Digital Government in the US", by Carol Hert, Syracuse University

These presentations are described in the following sections.

The ONS Metadata Programme

Jan Thomas, Office for National Statistics, UK
Jan.L.Thomas@ons.gsi.gov.uk

In this presentation, metadata definitions, planning, benefits for the ONS Statistical Institute as well as the system demonstration have been illustrated.

Metadata Definitions:

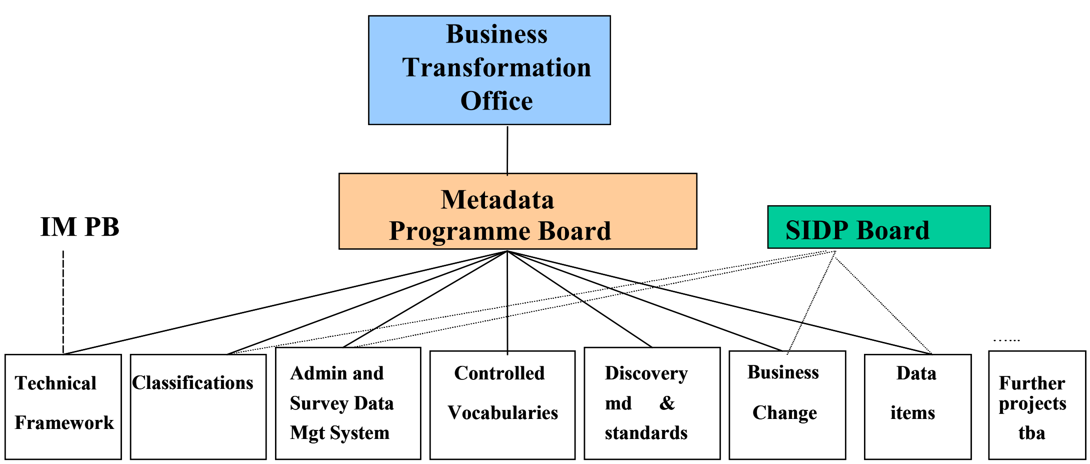
- **Statistical Metadata** – what is the object? Information about the content of statistical data, to support understanding and interpretation of the data

- **Discovery Metadata** – where is the object? Enable a user to identify and find appropriate content or information, using e.g. title, author, keywords, publication date, variable names, geography areas.

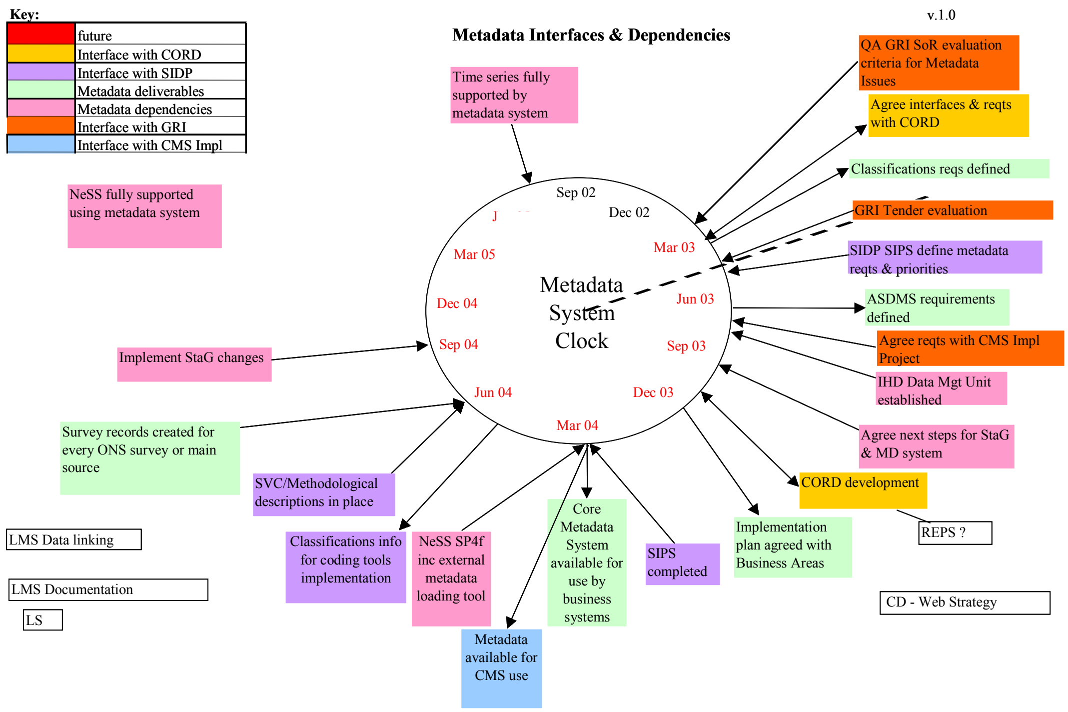
- **Control Metadata** – (also known outside ONS as Administrative Metadata) what state is the object in? Information needed to automate workflow systems and manage content, e.g. creation date, data supplier, editor, approver, release date, archive date.

- **Technical Metadata** – how to use the object. Information about the location and format of the data, used by systems for data interchange and manipulation e.g. type of file, size of dataset, record length, record layout, field types/lengths, hardware/software required to access data.

How the ONS is getting there

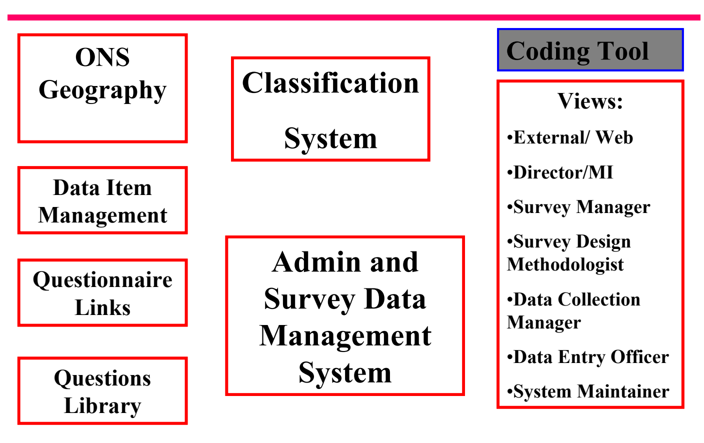


ONS Metadata Planning - What's Involved



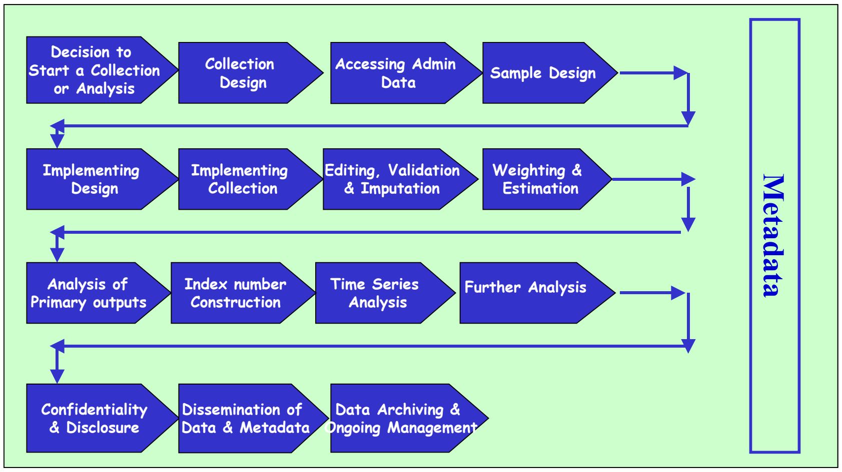
Benefits from metadata

- Improve quality of data for users;
- Improve the performance of ONS to produce more efficiently;
- Improve the data that people are supplying to us and to reduce supplier burden;
- Provide comparability for cross-survey analysis and harmonisation/integration of data and processes across ONS



ONS Metadata Systems

Statistical Value Chain



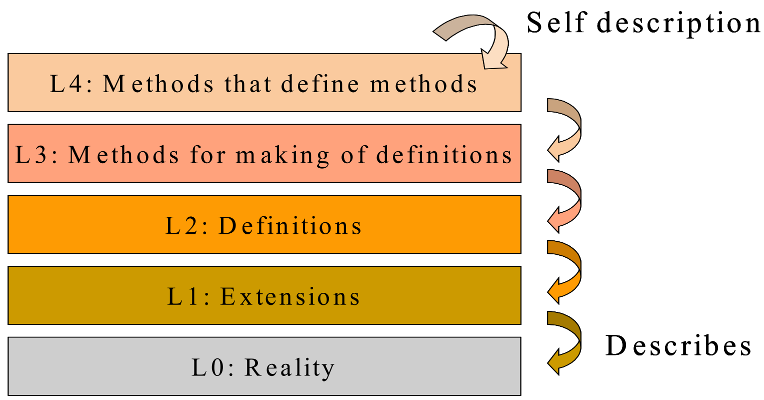
Finally, Mrs Thomas demonstrated the ONS System for the analysis and dissemination of statistics.

More information can be obtained from Mrs Thomas: Jan.L.Thomas@ons.gsi.gov.uk

The Banca d' Italia experience: hierarchical modeling statistical information systems

Vincenzo Del Vecchio, Banca d' Italia, Statistical Services Department
delvecchio.vincenzo@insedia.interbusiness.it

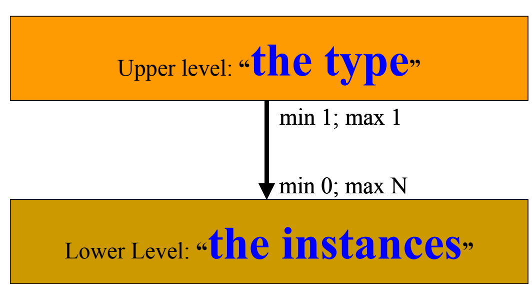
In the electronic and data processing (EDP) literature, the idea of a hierarchy of models emerged a certain time ago. The principle is that an information system can be described by different levels (layers) of modelling in a hierarchy in which the model of some level is described in terms of a model of the hierarchically upper level and it also describes one or more models in the hierarchically lower level, as displayed in the following figure:



In brief, starting from the reality that has to be described (call it the "zero" level), in the first level we have the data extensions, that is, models of parts of the reality, followed, in the second level, by the data definitions, that is, models of the data. The third level contains the methods used to produce the data definitions, that is, models of models of data (meta-models). Finally the fourth level contains the methods that produce other methods, that is models of the meta-models (meta-metamodels).[1]

The higher we go up in the hierarchy, the more abstract and general becomes the modelling: in practice, however, there is no need to have more than four levels, as the fourth level-model potentially allows us to obtain the whole hierarchy because a fourth level-model can be considered as a model that "generically defines models" and, hence, is also able to define itself, so eliminating the need of yet higher levels.

Constraints between layers exist as illustrated in the diagram on the right:



[1] Sometimes, in the EDP literature, "models" are called "schemas" and consequently "meta-models" are called "models" and "meta-metamodels" are called "meta-models".

According to the idea of a hierarchy of models, in 1986 an international standard was proposed by ISO/ANSI for the design and implementation of a generic Information Resource Dictionary System (IRDS).

An IRDS can be considered an information system that describes another information system. The proposal is based on a multi-level structure consisting in four levels in which every level has the purpose of defining the immediately lower level, as described in the previous subsection. The first level (the data) is considered external to the IRDS, the second level has the purpose of defining the data and is considered as the *content* of the IRDS, the third level contains the *structure* (that is, the model) of the IRDS and is itself defined by the fourth level, fixed by the standard. The four levels and the IRDS standard are illustrated in the following diagram:

L4 methods that define methods ==> fixed by the standard

L3 methods for making of definitions ==> IRDS structure

L2 Definitions ==> IRDS content

L1 Extensions ==> Information resource to be described

A more recent and ambitious application of the same principle is the four level-structure proposed by the Object Management Group (OMG), an organisation for the standardization in the object-oriented field of the software development in which many of the leading software production firms in the world participate. In the OMG standard, the four levels are called M3, M2, M1, and M0, respectively. M3 is the fourth level (the *meta-metamodels* level). Only one model at level M3 is necessary to define all the M2 level-models (*meta-models*). The OMG standard for the M3 model, called MOF (Meta Object Facility), is able to define itself. Examples of the third level (M2) meta-models are the UML meta-model and the relational meta-model. Correspondently, at the second level (M1) there are UML models and relational models relevant to a specific subject. First level (M0) contains data. This structure is as follows:

L4 ==> M3 : metametamodels (MOF - Meta Object Facility)

L3 ==> M3 : metamodels (UML, relational, …)

L2 ==> M1 : models (UML schema, relational schema, …)

L1 ==> M0 : data

The four level hierarchy of models was also applied to the conceptual[2] modelling of the statistical information systems that support the activity of the institutional functions of the Bank. In the statistics case a fourth level-model has the purpose of defining "structures" suitable to define third level-models. Such structures are not specific of statistics, they are instead more general and usable also in other fields (for example, the operational systems). That is to say that a fourth level-model contains structures able to define any kind of methodology, possibly shared by all of them. An important feature of a fourth level-model is its self-describing property, that is, the ability of its structures to describe themselves and, therefore, to make the existence of levels higher than four superfluous.

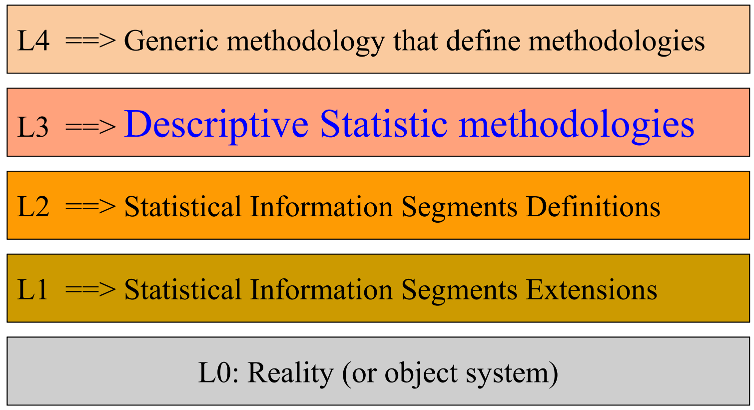
The specificity of the statistical field is located at the third level. A "statistical" third level-model, in fact, is considered the formal representation of a methodology for statistical description of the reality (that is, a descriptive statistic methodology). A third level-model contains structures able to give a concrete and possibly formal shape to statistical methodological rules.

A model of the second level can be considered the definition of a specific statistical information segment, that is, the definition of data and processes relevant to a specific subject. Therefore, second level-models are specific subject-matter models produced using a certain statistical methodology (that is, a third level-model). Note that, according to the general four level hierarchy idea, the notion of "data model" (data that is the definition of *other* data) is more specific than the more common notion of "metadata" (data that describes other data in some way). For example, a "quality datum", that is, a datum measuring, or reporting the quality of another datum, can be considered a "metadatum", yet it is not the "model" of the latter. In the four level model approach, both the original datum and the quality datum are considered level 1-data and have their definition in a level 2-model. So a level 1-model happens to contain also metadata, and the relationship between a datum and its "not definition metadata" is not of 'type-instance' type and it takes place within the *same* level, not between different ones.

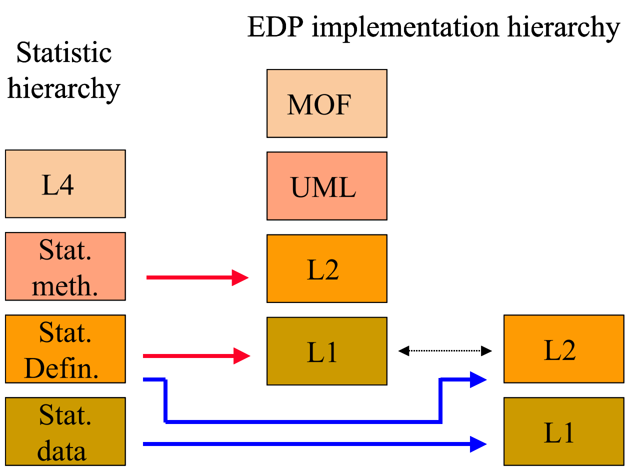
As displayed in the following figure, a model of the first level is the extension of a statistical information segment, that is, an occurrence of a second level-model. In simpler words, it is a set of values that correspond to a definition. Different sets of values, therefore, are different first level-models. More than one extension may correspond to the same definition such as in the case that the measurement process generating the data extension is performed more than once or, likewise, owing to the evolution of the data content in time: an update in the data content gives origin to a new extension, different from the previous one but with the same definition.

---

[2] The term "conceptual" is used to mean "independent from the implementation", so that many possible practical implementations can be made of a "conceptual" model, each one following its own set of implementation rules (that is, its own implementation model). In principle, there can also be models not implemented in an EDP environment



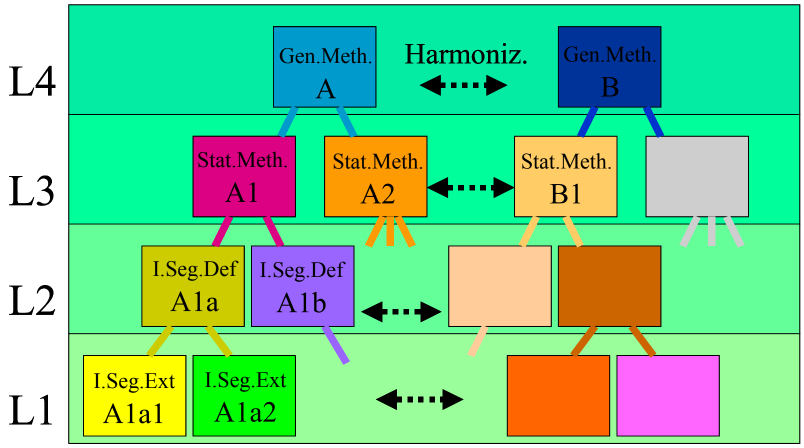
The electronic and data processing (EDP) implementation hierarchy is illustrated in the following diagram:



The Hierarchy of Models and the Practice of Harmonization

The multi-level hierarchy of models can be seen as a conceptual tool to deal with the complexity of the statistical information systems. It appears to be useful in the design and the operation of information systems as well as in the analysis of existing ones and in the effort of harmonizing and standardizing them, independently of how the implementation is done. The practical application of the idea leads to identify many models on every level and the type-instance relationships between models belonging to consecutive levels (see following figure), roughly:

- a fourth level-model for each general modeling methodology used in practice;
- a third level-model for each descriptive statistical methodology used in practice;
- a second level-model for the definition of each information segment;
- a first level-model for each extension of an information segment.



Such a schema of decomposition and description, applicable to a single information system and different information systems of different organisations alike, could provide a guideline for harmonization efforts.

The harmonization effort takes place *within* each level. Models at different levels, in fact, have different purposes and their objects[3] are different because the goal of a certain level is to describe the lower one. On the contrary, it makes sense to compare and possibly harmonise models at the same level when their objects are also partly the same.

The harmonization between different models in levels 1 and 2 can be very important. The mapping between different models enables to convert a model into another, to exchange their contents (data and definitions), to share parts of the model and to ensure some degree of coherence between them.

Models and Languages

Every model in the hierarchy gives rise to a language used for defining and naming its structures, and to the possible operations on them. In defining models, terms can be borrowed from the natural language, but they assume a more specific and formal meaning in the model context. Moreover, the same natural term can be used in different models, assuming different meanings in each one. Therefore, for the proper comprehension of the meaning of the terms, knowledge of the context in which they are used (that is, the natural language, or a more formal model) is necessary. That is to say that the meaning of a term belongs to the model in which the term is defined.

This principle, applied to the hierarchy of models, leads to four levels of languages each one corresponding to a modelling level:

- a generic modelling language for each fourth level-model;

---

[3] What the model describes.

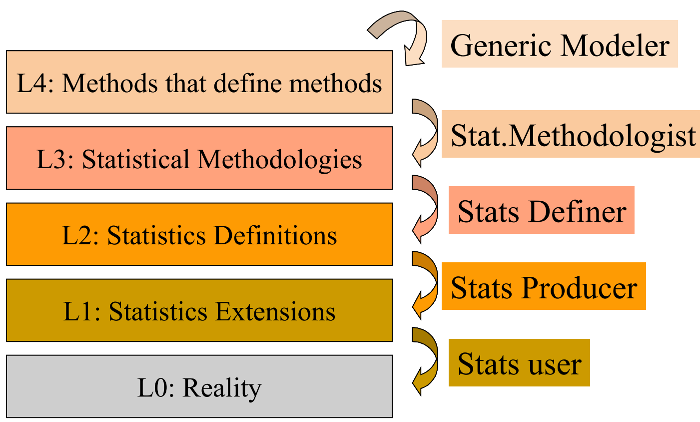
- a statistical methodology language for each third level-model (in which the terms are derived from the descriptive statistic methodology);

- a subject matter definition language for each definition model (in which the terms are basically derived from the discipline to which the model refers, like economics, medicine, physics, …);

- a subject matter extension language for each extension model (in which the terms are the symbols used in the extensional representation).

Levels and roles

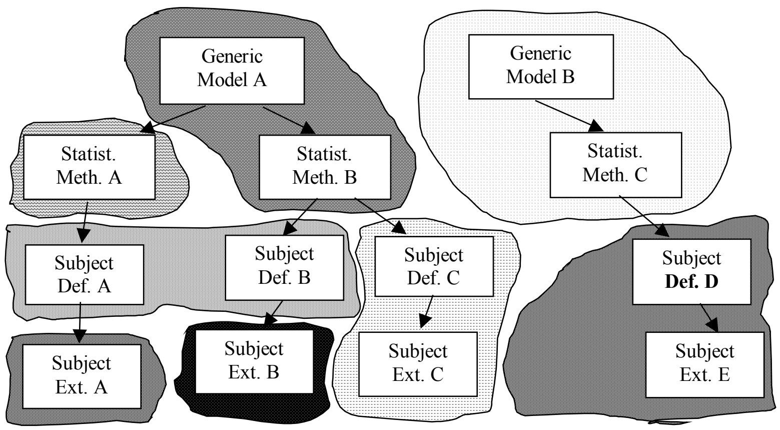
The multi-level hierarchy can also be used to distinguish different roles in the information system. Basically, the idea is that a "role" consists in using the model on a certain level in order to produce models in the lower level. Proceeding from up to down:



- the "generic modeler" produces general purpose models (fourth level-models);
- the "statistical methodologist" uses a general purpose model to produce statistical methodologies;
- the "statistics definer" uses a statistical methodology to produce subject matter definitions;
- the "statistics producer" uses a subject matter definition to produce statistics;
- the "statistics user" uses statistics to understand the reality and possibly act on it.

Models and Competencies on them

The multi-level hierarchy of models provides a method to distinguish the competencies between different units, establishing a high-level link between the information system structure and the organisation involved in running and using it. Every model, in fact, can be in charge of a different responsibility (the owner of the model). On the other hand, any number of models, also in different levels, can be in charge of the same responsibility (see following figure). The bi-dimensional schema allows implementing many configurations of competencies, spread between two extreme and ideal situations. The first one is the vertical decomposition, according to the subject of the statistics, in case the whole hierarchy of models (all of the four levels L1 through L4) relevant to a certain subject is left in charge of the same unit. The second one is the horizontal decomposition, according to the roles, when a whole level is left in charge of the same unit. Practically, the tendency to have only one model in the level 3 and 4 drives toward an intermediate situation, like the imaginary one drawn in the following figure.



Conclusions:

4L modeling to deal with:

- Infological completness
- Structuring information systems
- Harmonization
- Languages
- Roles
- Active processing
- Competencies

"The SORS experience"

Josa Klep, Statistical Office of the Republic of Slovenia
Joza.Klep@gov.si

The presentation consisted of the following topics:

- background
- classification server on the internet
- StatCop98 project
- STAT 2000 project
- metadata repositories within the target data flow
- quality concepts and metadata
- dissemination

More specifically:

Background:

It started with a "Modernisation and development of the statistical information system in Slovenia", Feasibility Study on the Architecture of Information Systems and Related Equipment Issues.

The study was carried out in the period of February - September 1997. A number of short term missions to SORS by experts from Statistics Sweden took place.

Among the main conclusions of the study were:

- SORS has an excellent potential for developing a modern, register-based statistical system, based upon administrative sources in combination with sample surveys.
- One way to obtain better focus in the development of the systems for statistics production in Slovenia is to specify a very precise and concrete target architecture for the development, and to formulate a strategy for implementing this architecture step by step, in a systematic way. As a matter of fact, this approach has been proposed by SORS itself.
- The information system architecture for a statistical office should cover a number of different information systems types and their relations to each other: registers, survey processing systems (primary systems), analytical systems (secondary systems), and metainformation systems.
- It was further recommended, that the first step in the proposed architecture should be the building of a classification database. The prototype was presented at the board of director general as of 1 March 2000 and put in production in November 2000.

Main goals of the classification server were:

- To centrally store all classifications and concordances used by SORS
- To provide common ways to update and access classifications data
- To facilitate the use of classifications stored in one central place and readily accessible to all users (internal and external)
- To provide common storage facilities which will make it easier to find and access classifications data
- To facilitate maintenance and release of new classification versions
- To facilitate maintenance and release of coding lists
- To provide report generation

Present status:

- 574 classifications and 67 concordances as per end of April 2003; 16 owner groups.
- Expected target is 2900 classifications and concordances (as reported for New Zealand in December 2001).

StatCop98 project:

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

Another component - 4.3: Development of databases and software solutions - aimed at an integrated process of aggregation and dissemination of data from the Census of agriculture, horticulture and viticulture 2000 (AC2000) and other agricultural statistics (AGRISTAT).

Within these two components, the basic common functions in the context of statistical data warehouse were defined according to Sundgren: "Statistical metadata are descriptive information or documentation about statistical data, i.e. microdata, macrodata, or other metadata. Statistical metadata facilitates sharing, querying, and understanding of statistical data over the lifetime of the data".

In a broad sense, "production" covers the whole life cycle of a statistical survey or a statistical information system, including design, implementation, operation, monitoring, maintenance and evaluation. Producers of statistical data therefore include: designers, input data providers and subject-matter statisticians. All these categories of producers of statistical data have their typical metadata needs.

Therefore, *the SORS Corporate Metadata Repository and the necessary user interfaces were built with an aim to facilitate knowledge sharing between the different producer groups in developing the computer environment for the selected statistical surveys.*

STAT 2000 - focus on dissemination procedures

SORS works with Statistics Sweden and Statistics Denmark, both in metadata modelling and in applying templates and tools (adjusted for SORS requirements) already developed in Nordic countries. SORS continues with work in the field of metadata and documentation.

In more details on the required metadata definitions regarding:

- data collection (definition of population, frame creation, necessary metadata to monitor sampling procedures, monitoring and managing contacts with reporting units...)
- quality metadata (definition of processes where the appropriate metadata is "generated", f.i. contents from the "variable" part of E-R; administrative data from "National (annual) Programme of Statistical Surveys"; availability and punctuality of the data from the records of released publications in previous year...). In view of contents, we would like to have the necessary set of metadata to meet the requirements of IMF SDDS template, SCB Quality declaration and foreseen Eurostat quality declaration template...
- we would also like to incorporate and adjust METADOK completely. We are implementing the "light" PX-Web version. It is our intention to use the complete PC-Axis family (also PX-MAP and PX-PUBL) and allow access to both - data and publications from the end-user point of view with the browser only.

Quality Concept for official Statistics, SCB, MIS 2001

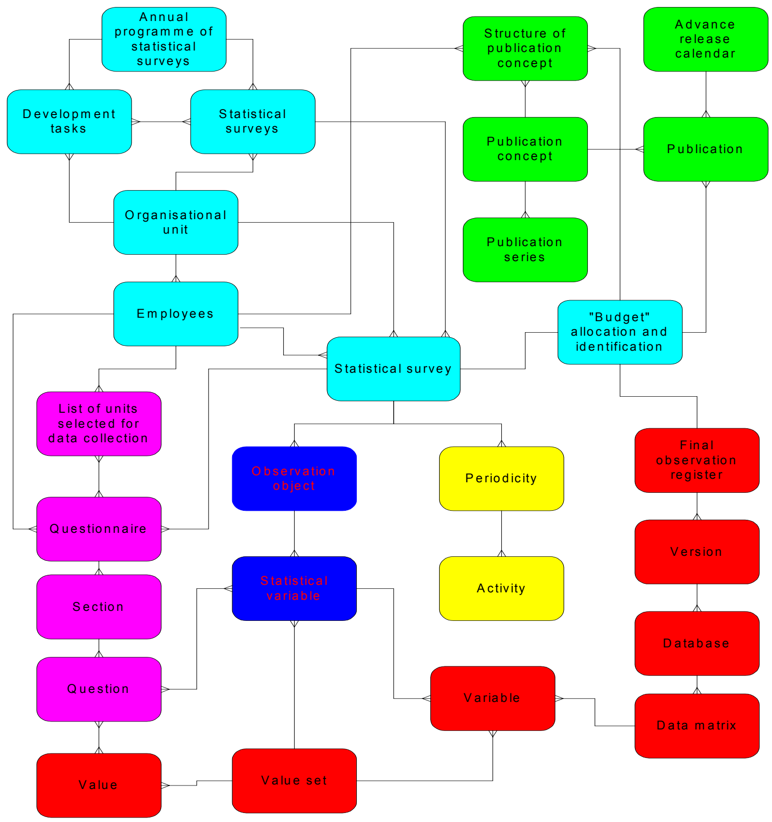
Design of quality declarations: by making a quality declaration the producer can specify the properties of a product so that it can be used in a proper way, and inform users about what quality in different aspects they can count on.

In quality improvement work: Since users opinions and preferences change over time, the producer must continuously strive to adapt the product to new needs and expectations. A user oriented quality concept facilitates communication between user and producer

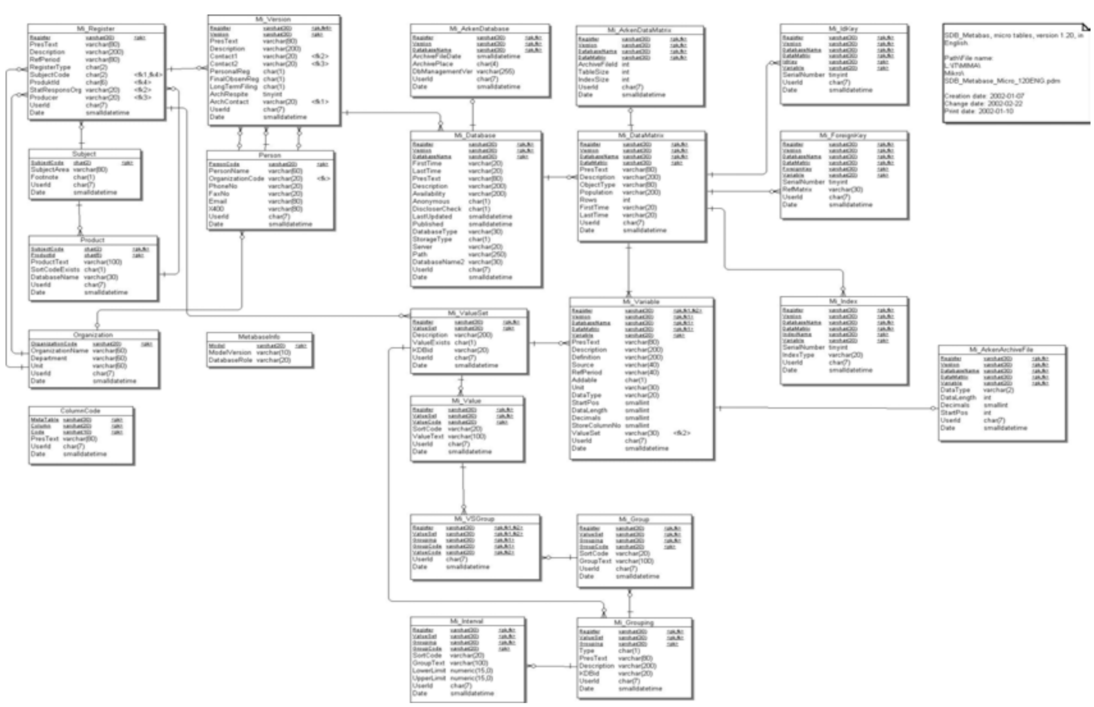
In evaluations of productivity: It is always of importance that production processes are as efficient as possible. A producer therefore wants to be able to judge the relative benefits and costs for different parts of the production process. The quality concept provides one of the instruments for efficiency evaluation and an optimal allocation of production resources.

Quality of statistics refers to all aspects of statistics, which are relevant for how well they meet user's needs for statistical information.

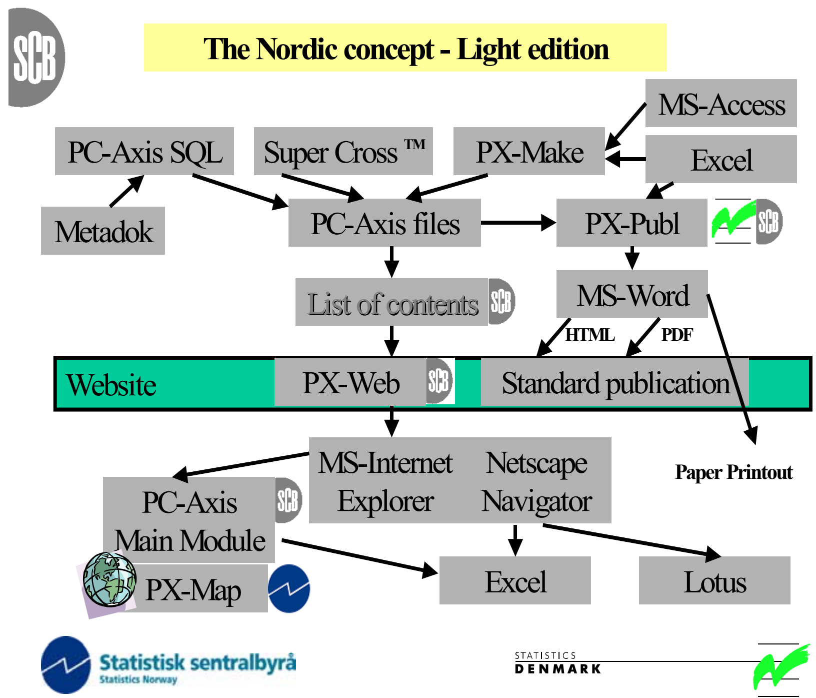
SORS METIS - main entities



SCB Metadok (see following diagram) will be used as the “red part” of the SORS METIS Main entities

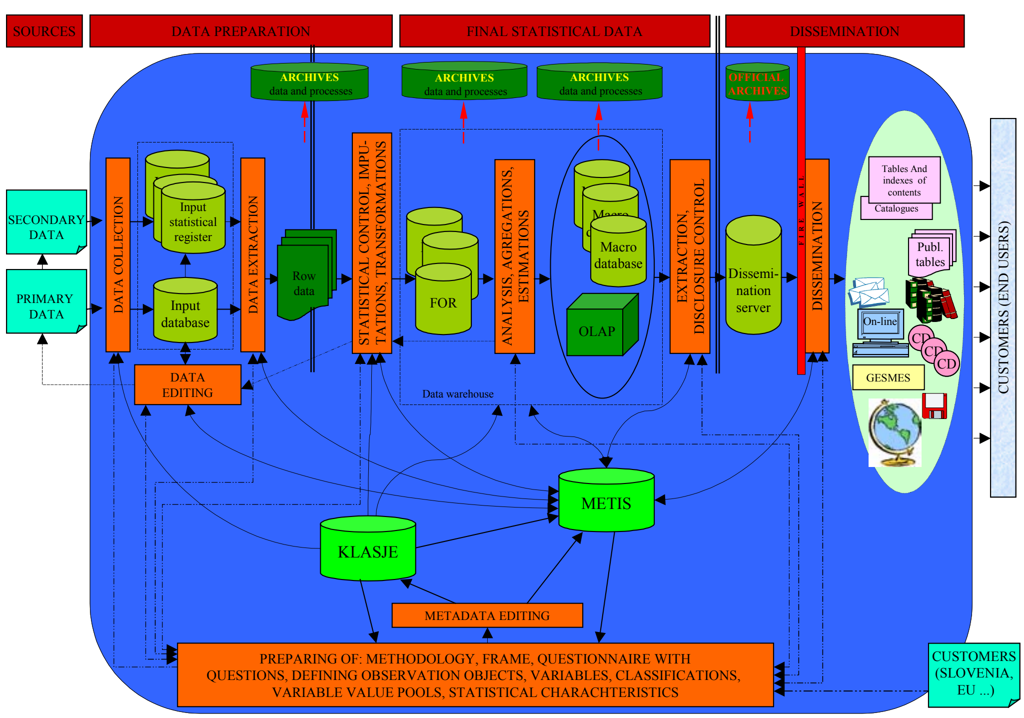


SORS moves from a publication oriented into multimedia oriented organisation with



PC-Axis family

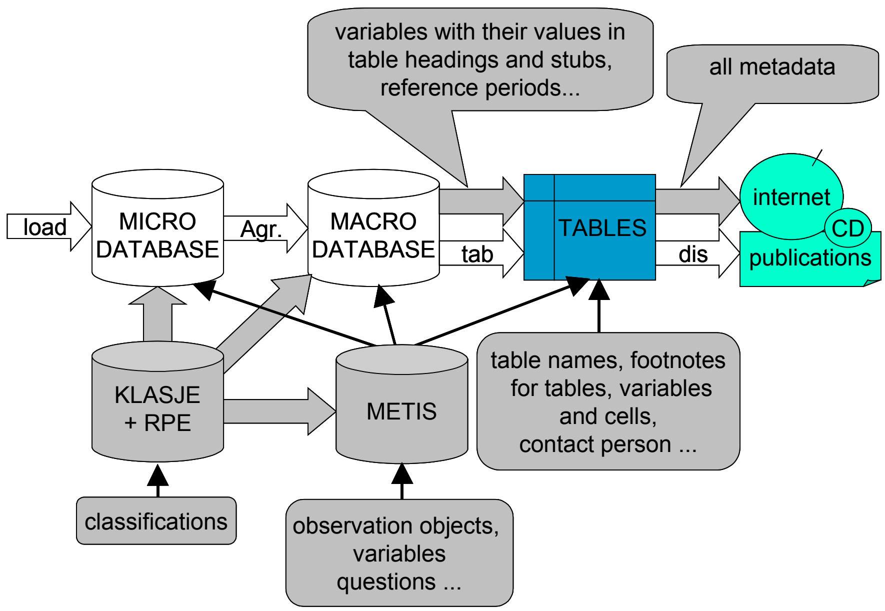
Scheme of target data flow at SORS



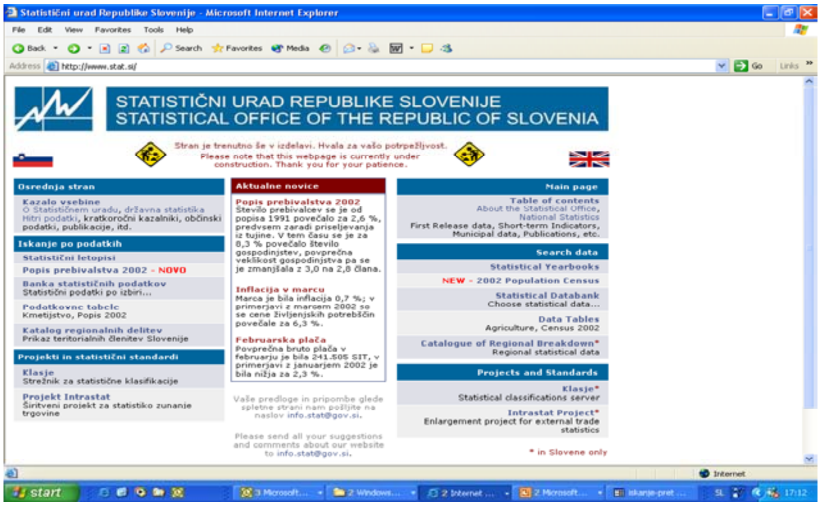
If we focus on data flows at SORS, Klasje is the source for :

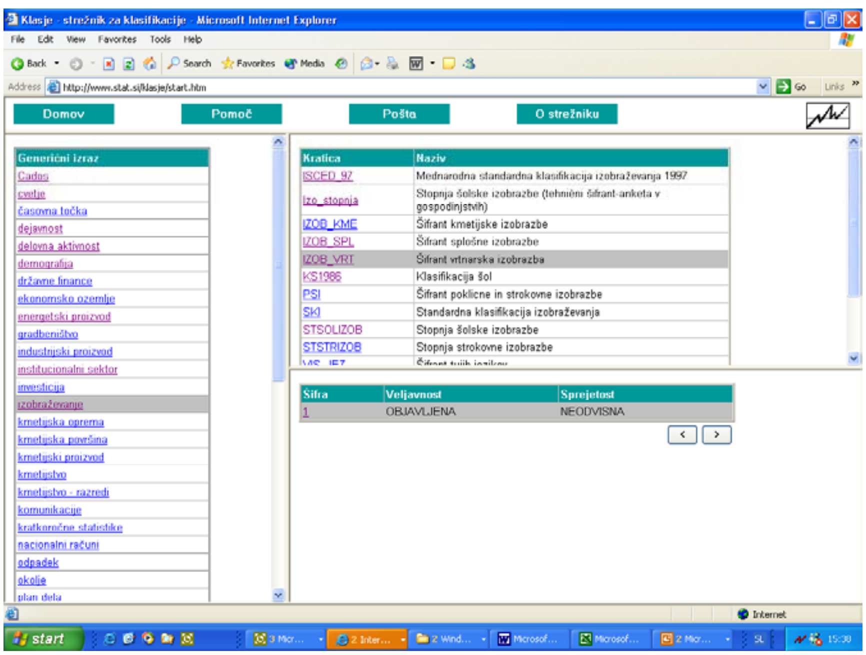
- metadata repository (Metis)
- in the process of data collection
- in the process of statistical control, imputations and transformations
- in the process of analysis, aggregations, estimations
- in the process of preparation of: methodology, frame, questionnaire with questions, defining observation objects, variables, classifications, variable value pools, statistical characteristics

Data and Metadata flow

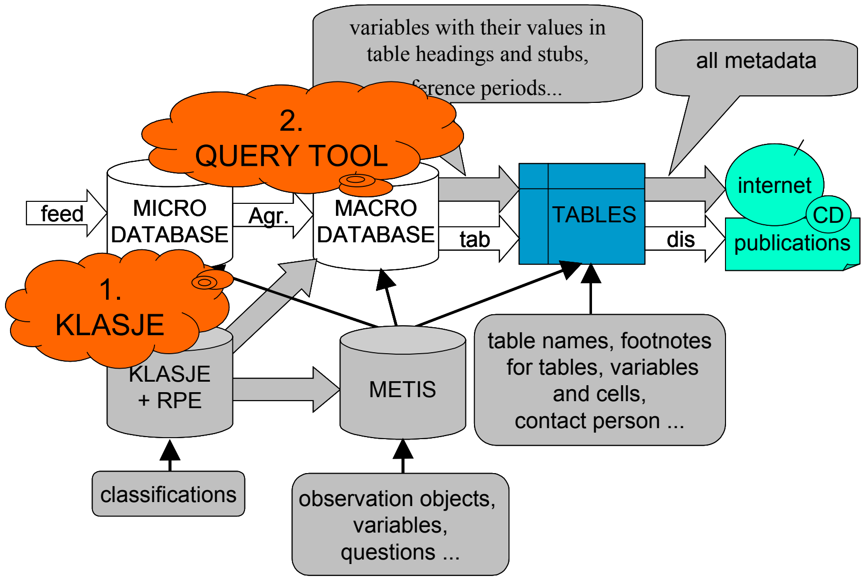


Classifications on the internet

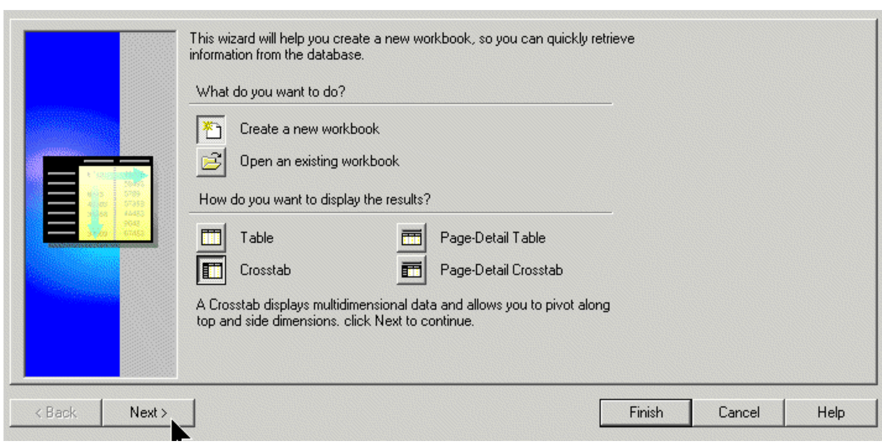


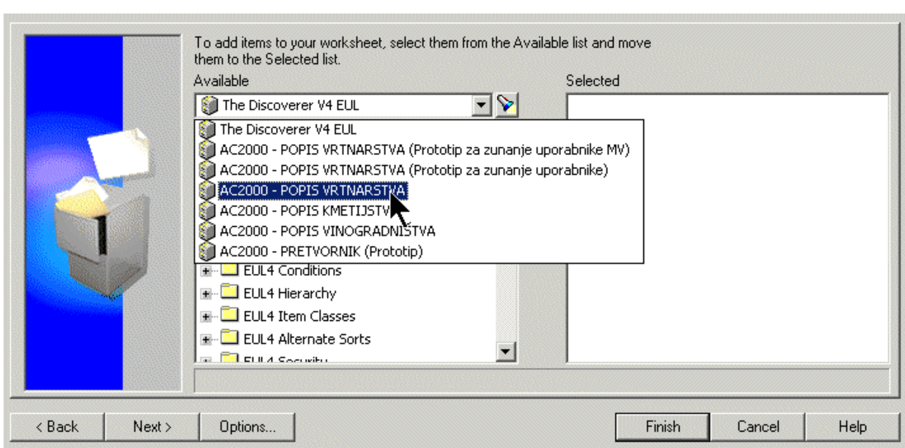


Making tables with query tool



Oracle discoverer for the users





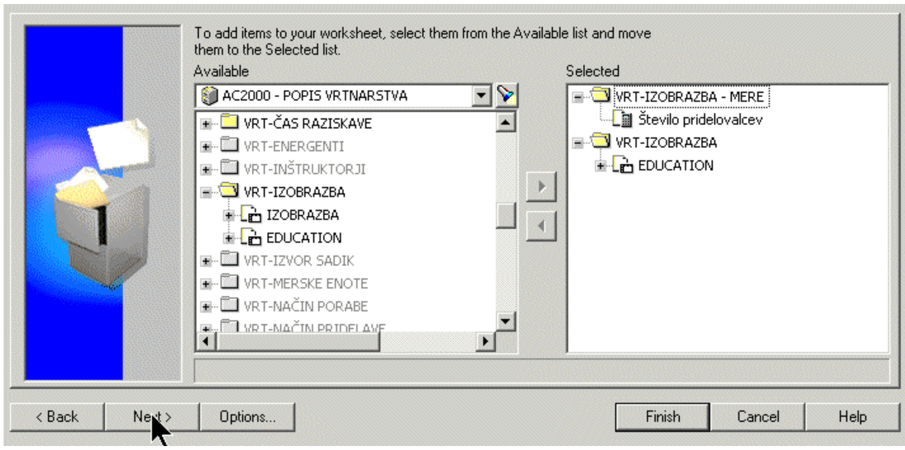
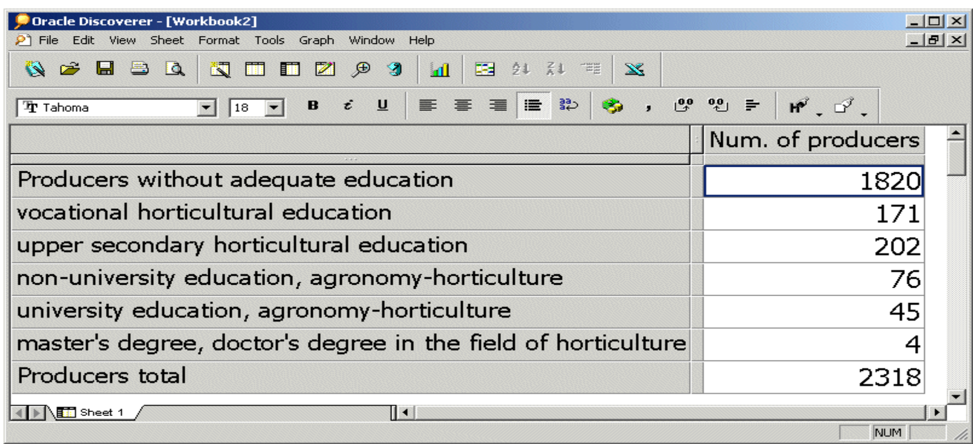
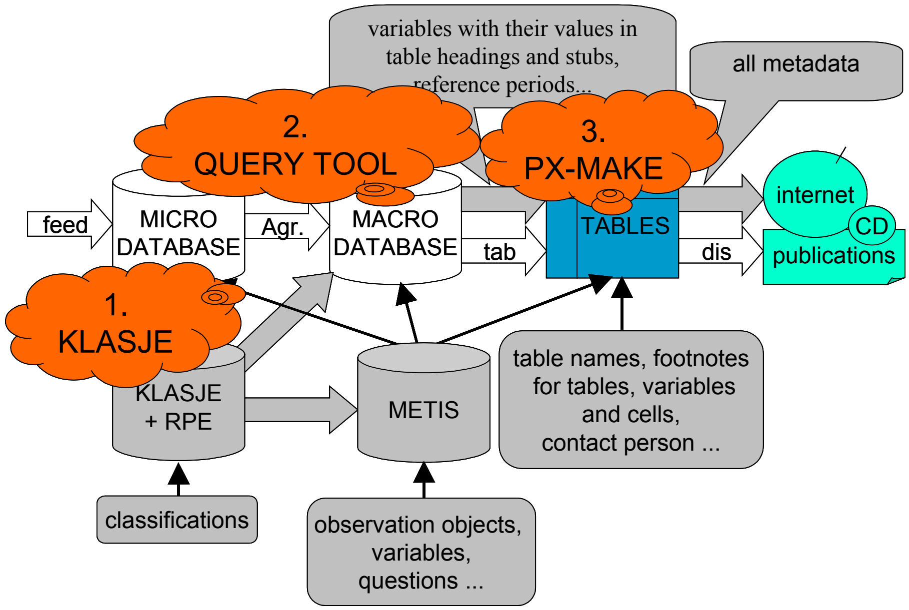


Table with data prepared with Oracle Discoverer

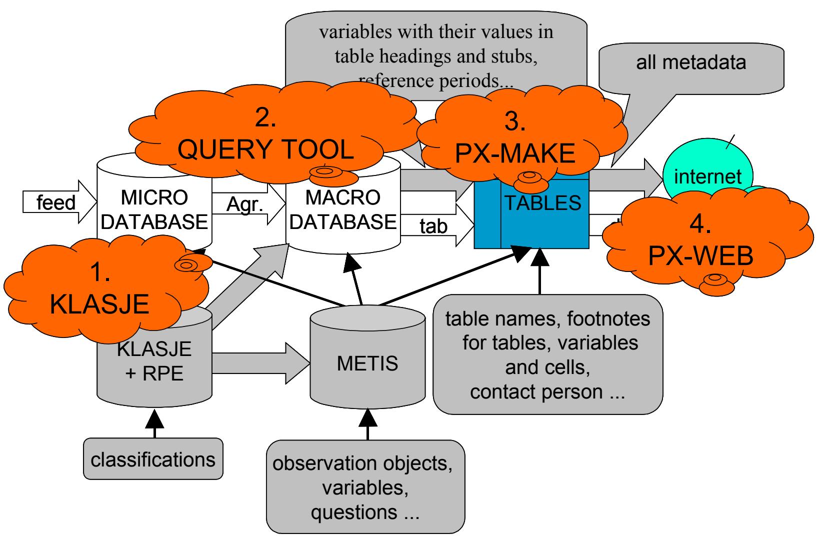


	Num. of producers
Producers without adequate education	1820
vocational horticultural education	171
upper secondary horticultural education	202
non-university education, agronomy-horticulture	76
university education, agronomy-horticulture	45
master's degree, doctor's degree in the field of horticulture	4
Producers total	2318

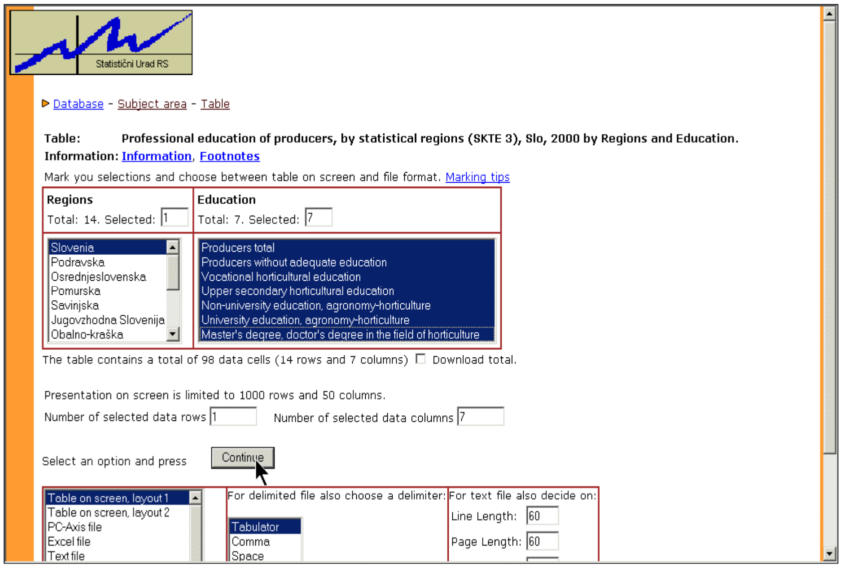
PX-MAKE



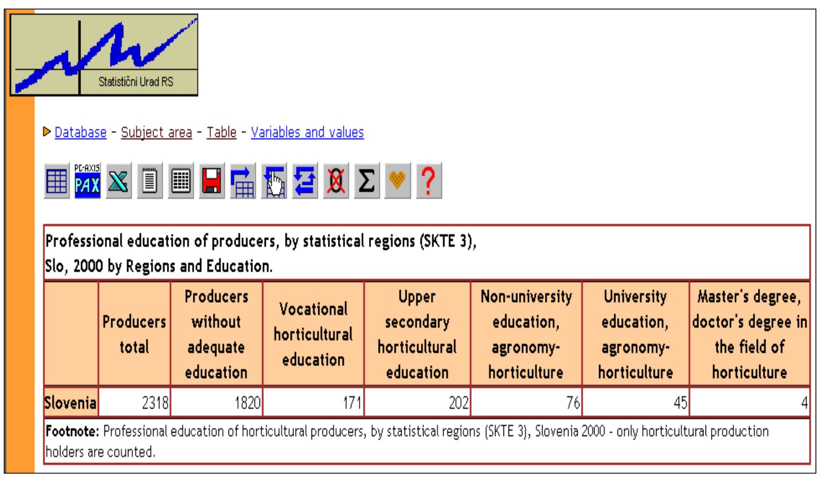
PX-WEB



PC-AXIS database-search results



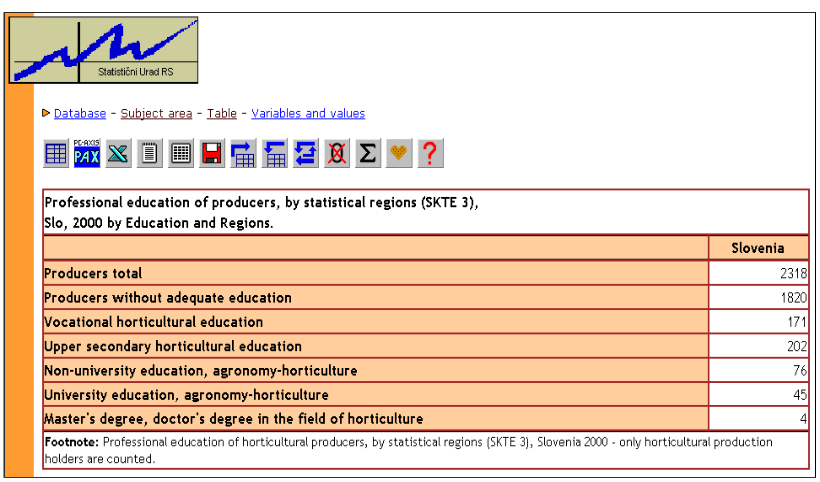
PC-AXIS file with data in the table accessible on the internet



Professional education of producers, by statistical regions (SKTE 3), Slo, 2000 by Regions and Education.

	Producers total	Producers without adequate education	Vocational horticultural education	Upper secondary horticultural education	Non-university education, agronomy-horticulture	University education, agronomy-horticulture	Master's degree, doctor's degree in the field of horticulture
Slovenia	2318	1820	171	202	76	45	4

Footnote: Professional education of horticultural producers, by statistical regions (SKTE 3), Slovenia 2000 - only horticultural production holders are counted.

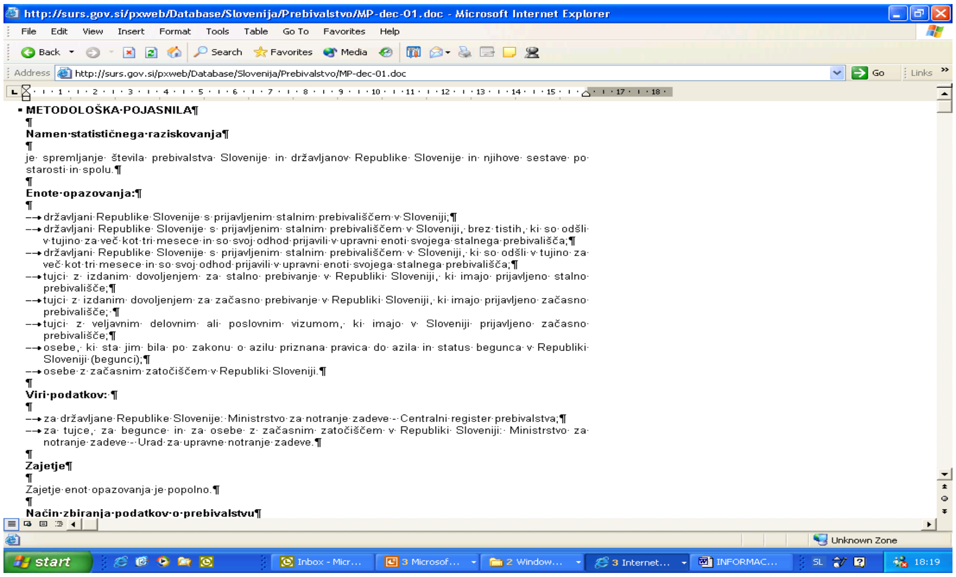


Professional education of producers, by statistical regions (SKTE 3), Slo, 2000 by Education and Regions.

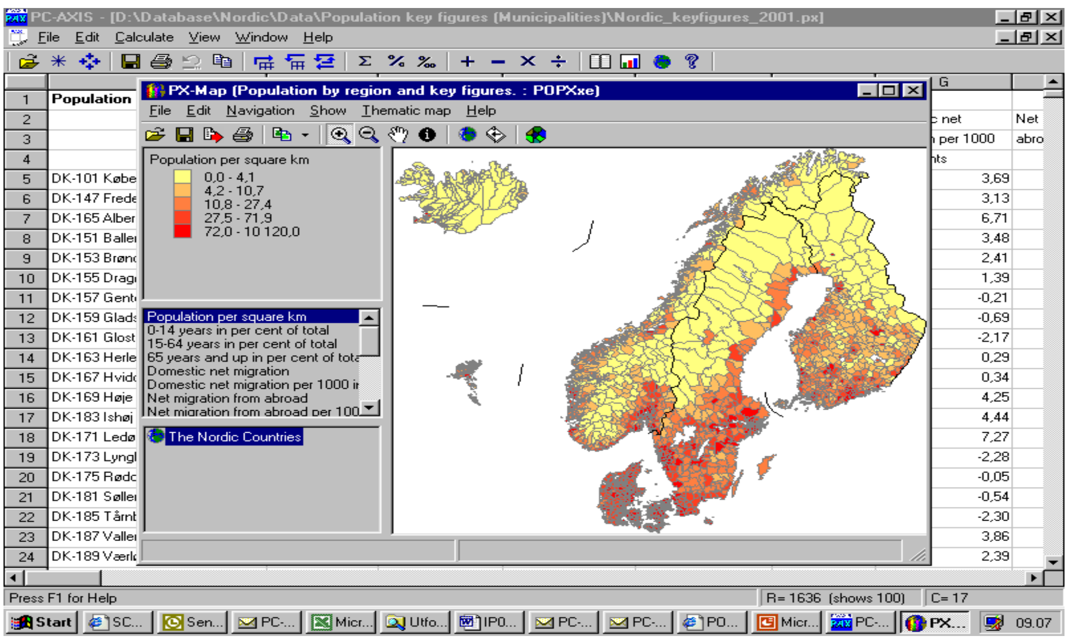
	Slovenia
Producers total	2318
Producers without adequate education	1820
Vocational horticultural education	171
Upper secondary horticultural education	202
Non-university education, agronomy-horticulture	76
University education, agronomy-horticulture	45
Master's degree, doctor's degree in the field of horticulture	4

Footnote: Professional education of horticultural producers, by statistical regions (SKTE 3), Slovenia 2000 - only horticultural production holders are counted.

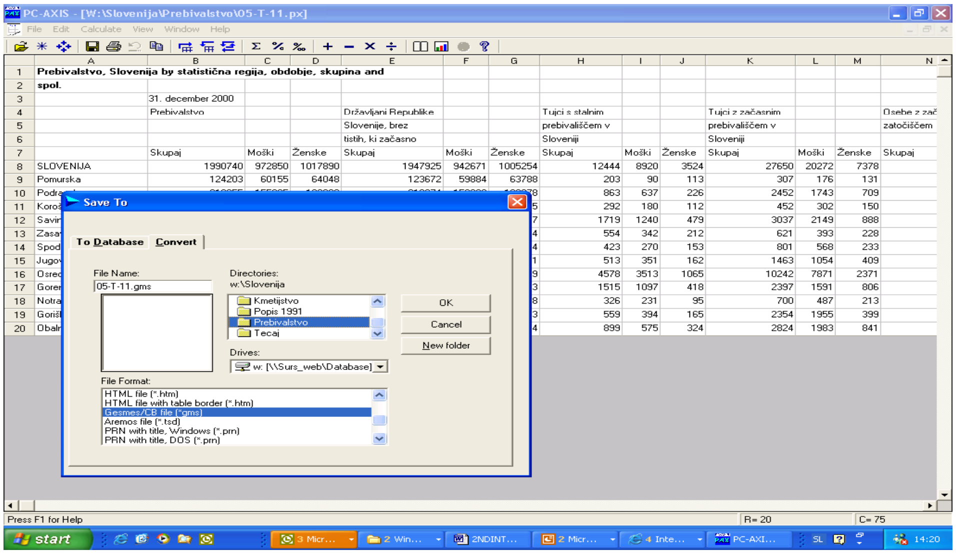
Quality declarations with PX-files



PX-Map



GESMES/CB message converted from PX-file



The end of STAT2000 project – some general problems with quality of metadata

- coherence: different (previously “stand alone”) sources; in one repository erroneous data more visible
- prevailing view – focus on the survey not on the statistics (product)
- erroneous data (periodicity, reference period types, wrong survey codes, not all questionnaires recorded, etc.) will need “editing”; if new database not used as a new source for existing needs, the procedure will have to be repeated in the following years
- in OJ listed as a statistical survey – but no output (publication) reported
- difficulties in defying object types in the f.i. economic statistics (much easier in agriculture)
- number of variables increases dramatically in statistical production process - from the questionnaire design to the dissemination (example: monthly survey on wages and employment; 13 variables on the questionnaire, 11 from the ID of statistical unit, with derived variables 43 on the output side); challenge is allowing different views for different needs
- waiting for the first draft of Nêuchatel paper on variablesinternational conference focused on the discussion of variables; with experiences from different NSI; different aspects; could be very useful and extremely important for every day practice

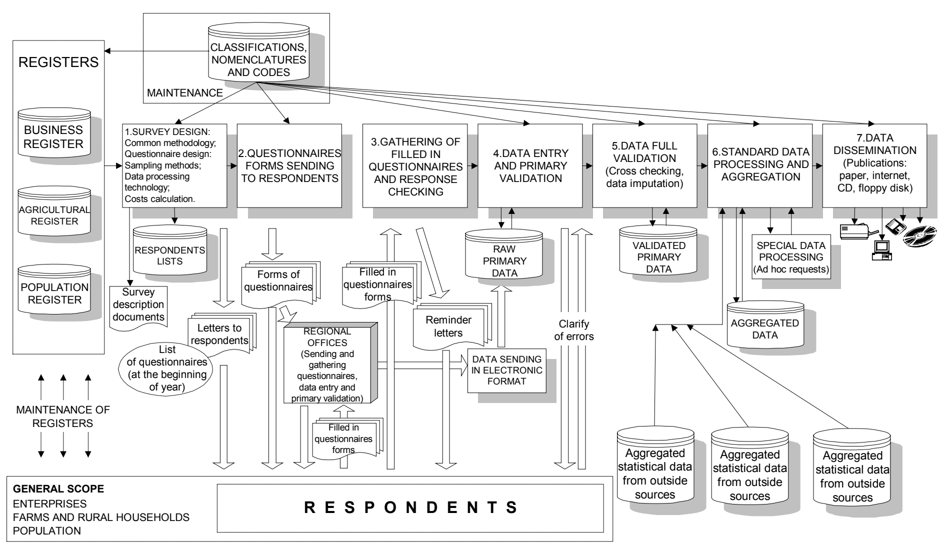
Metadata driven integrated statistical data management system - Development and implementation experience

Karlis Zeila, CSB of Latvia
kzeila@csb.lv

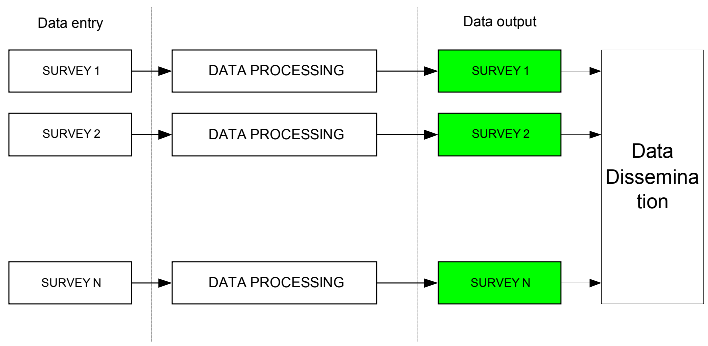
The main business and information technology improvement objectives that the CSB intended to achieve as the result of project:

- increase the quality of data, processes and output;
- further integration instead of fragmentation on organisational and IT level;
- reduce redundant activities, structures and technical solutions wherever integration can cause more effective results;
- make a more efficient use and availability of statistical data by using common data warehouse;
- provide the end-users (statistics users, statistics producers, statistics designers, statistics managers) with adequate, flexible applications at their specific work places;
- replace tedious and time consuming tasks by value-added activities through more effective use of the IT infrastructure;
- using meta data as the general principle of data processing;
- use electronic data distribution and dissemination;
- making extensive use of a flexible database management for providing internal and external users with high performance, confidentially and security.

CSB Data flow diagram

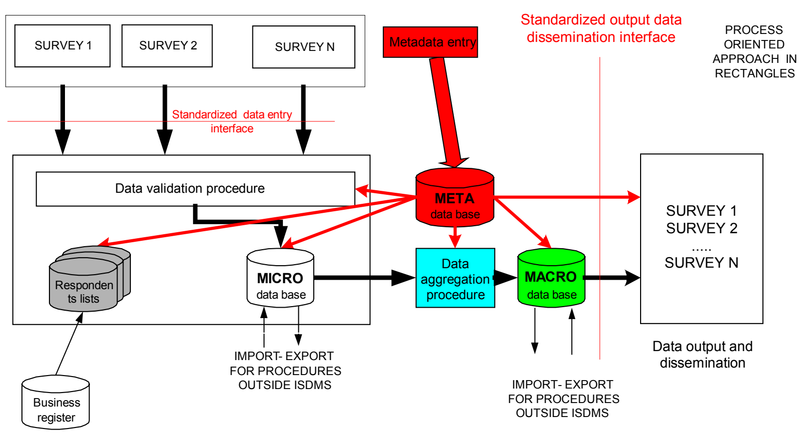


Stove-pipe statistical data processing



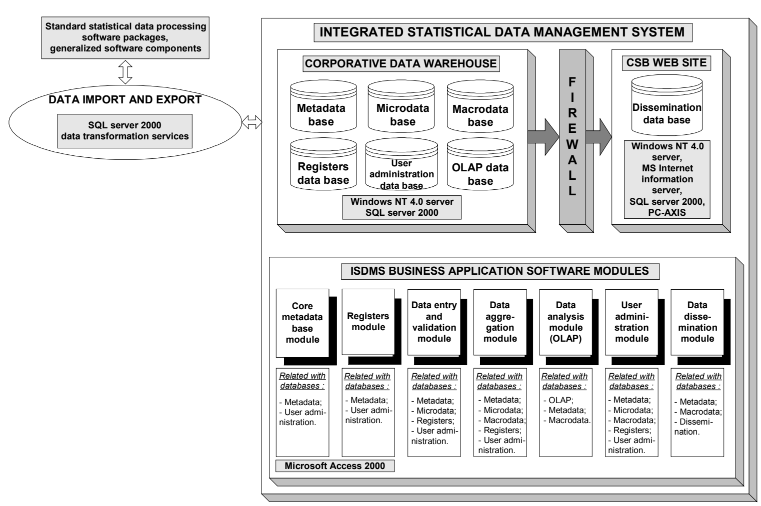
Hardware PC (Pentium I, II) and File Server   Software: MS ACCESS, FOX PRO(DOS), PARADOX (DOS)

Integrated Metadata Driven Quasy Process Oriented Technology



Client - Server technology
Hardware Workstation PC (Pentium II, III,IV) and Enterprise Server
Software: MS ACCESS 2000 - for clients Workstations,   MS SQL SERVER 2000, OLAP - for Enterprise Server

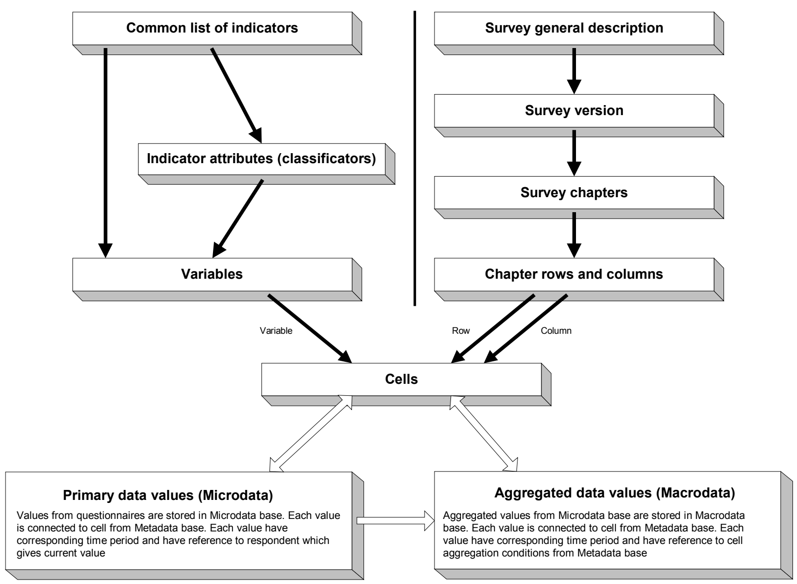
System General Architecture



Metadata module functionality

- The metadata base module is the core of the system - Metadata base data handled by this module are in use by all other modules of the ISDMS,
- Metadata base is the key element for the creating universal, common, programming-free approach for different statistical surveys data processing instead of development of software specially for certain survey, where every change of the survey will require a corresponding adaptation of the programs source code, and where it will be also necessary to develop new software for every future survey,
- Metadata base is linked at database structure model level with Microdata base and Macrodata base,
- System users can easy query necessary data form Microdata / Macrodata bases navigating via Metadata base,

<u>Metadata base link with Microdata/Macrodata bases</u>

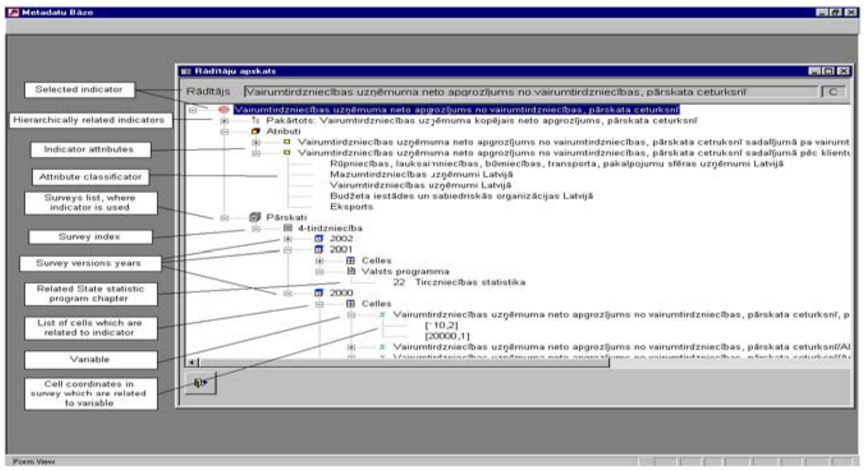


Main applications of the Metadata base module

- general description of statistical survey,
- description of survey version,
- description of indicators and attributes of statistical survey,
- description of content of statistical survey chapters,
- maintenance of validation rules of statistical survey,

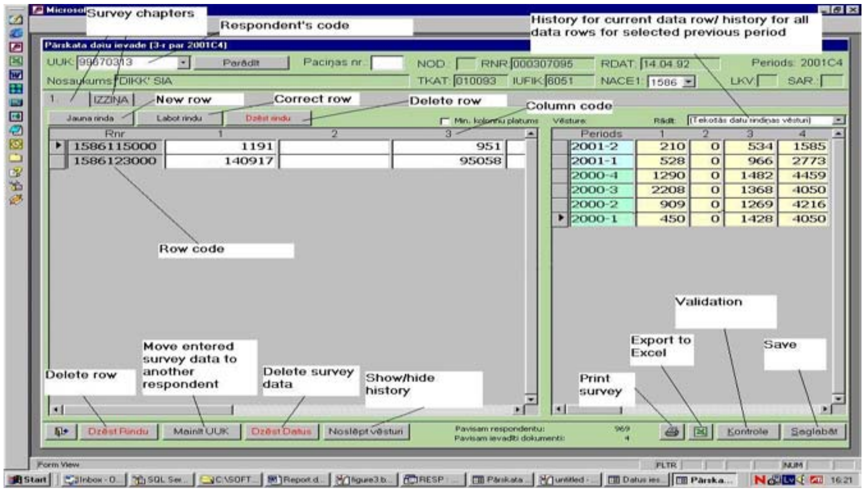
- description of aggregation conditions of statistical survey,
- grouping of classificators records,
- description of reports,
- common Metadata base data browsing

Common metadata-based browser



Survey data entry and validation module

- Module provides standardised approach to different statistical surveys data processing,
- To make any changes for survey content and/or layout, it is necessary only to change survey description in the Metadata base,
- For each selected survey for selected period following main functionality is available:
    - respondents list maintenance;
    - data entry and validation;
    - data aggregation;
    - reports creation;
    - data export/import.
- ISDMS ensure linking of the statistical survey to a particular list of respondents obtained from Business register.
- Each survey version for each period have its own list of respondents.



Lessons learned

- Design of the new information system should be based on the results of deep analysis of the statistical processes and data flows
- Clear objectives of achievements have to be set up, discussed and approved by all parties involved
    - Statisticians
    - IT personal
    - Administration
- Within the process of the design and implementation of metadata driven integrated statistical information system both parties statisticians and IT specialists should be involved from the very beginning
- Both parties have to have clear understanding of all statistical processes,which will be covered by the system, as well as metadata meaning and role within the system from production and user sides
- Initiative to move from classical stove-pipe production approach to process oriented have to come from statisticians side not from IT personal or administration
- Improvement of knowledge about metadata is one of the most important tasks through out of the all process of the design and implementation phases of the project
- Clear division of the tasks and responsibilities between statisticians and IT personal is the key point to achieve successful implementation
- To achieve the best performance of the entire system it is important to organize the execution of the statistical processes in the right sequence.

The CODAM Implementation Experience

Claude Macchi

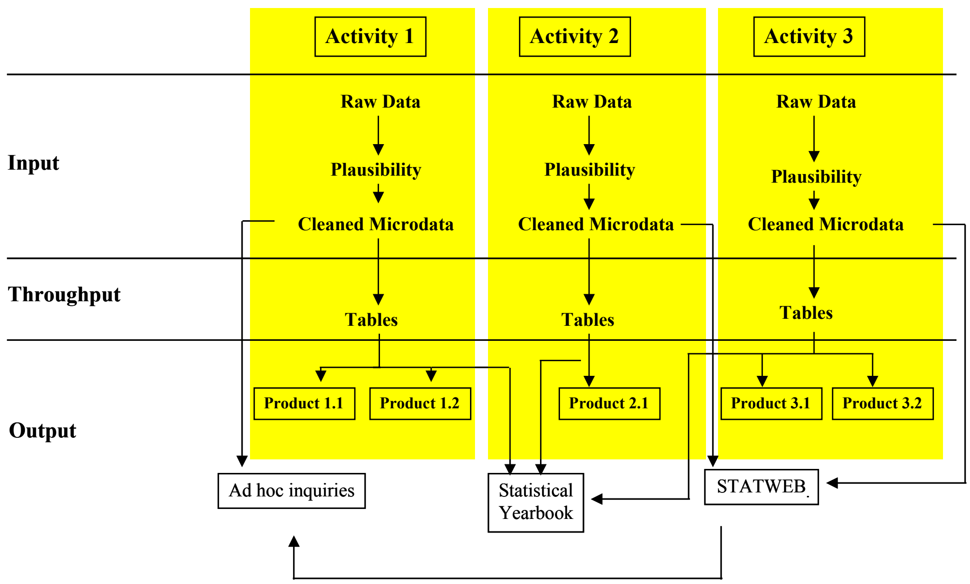
Swiss Federal Statistical Office

claude.macchi@bfs.admin.ch

Main points of the presentation:

i) *The Implementation of the Metadata System*

ii) *The Integration in CODAM*



The 'Metadata Chaos':

– Lack of Communication

– Lack of Documentation

– No Collaboration – no Co-ordination

The New Metadata System

– Centralized Metadata Administration

– Use of Standard Metadata

– Implementation of Standard Tools

Problems of the Implementation

– Technical Problems

– “Political” Problems

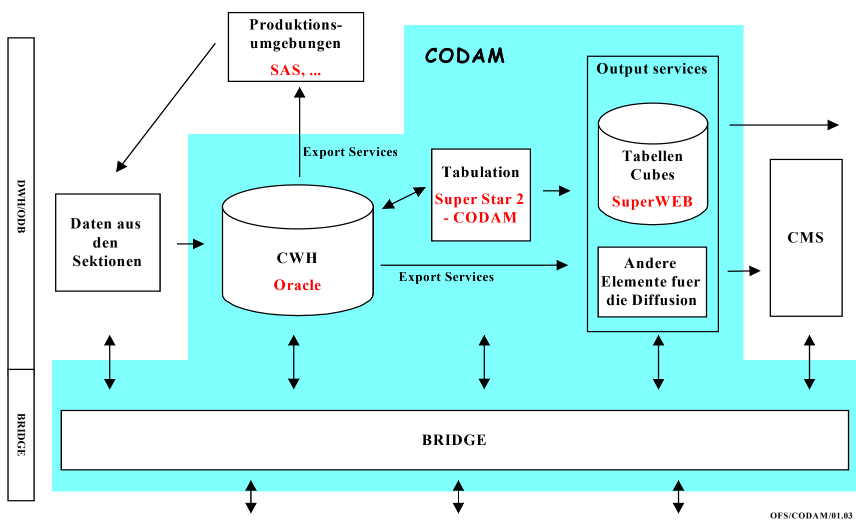
  - The Change of Habits
  - Less Autonomy
    - ⇨ Acceptance

The Implementation

- Integration of the Future Users
- Support of the General Management
- Step by Step Implementation
- Concept of Corporate

CODAM Corporate Data Management

- Central DWH, ODB and Metadatasystem
- Harmonization
- Standard Tools



Digital Government Initiatives in the United States: Statistical Information and Metadata

Carol A. Hert
Syracuse University
cahert@syr.edu

Overview of the presentation:

- Funding digital government research and projects
- The Technology transfer challenge
- Coordinating activities of statistical agencies and researchers
- The role for researchers in metadata activities
- Joining internal metadata activities to user-end tools

U.S. Funding Structures

Projects relevant to statistics, statistical information, metadata funded by

- Agencies themselves via operating or research funds
- National Science Foundation
  - Division of Social and Economic Sciences
  - Directorate of Computer and Information Science and Engineering
  - Digital Government initiative

Digital Government Initiative Goals:

- To support partnerships among researchers and across levels of government
- Support research on design and use of IT in support of democracy, citizen/government interactions, & governmental collaboration

Statistical Projects in DGI

- Research projects with statistical focus have been well represented and include:
  - Delivery of statistical information to citizens
  - Statistical literacy, metadata, statistical representations such as graphics
  - Various aspects of internal agency processes such as data collection
  - Confidentiality
- Technologies investigated:
  - Natural language processing,

  - information retrieval,
  - data mining,
  - Ontologies
  - Wearable computers
  - Databases
  - Etc.

Metadata-related Research Projects

- Language Modelling Approach to Metadata for Cross-Database Linkage and Search
- The CARDGIS Energy Data Collection
- Information Discovery in Digital Government: self extending topic maps and ontologies
- Integration of Data and Interfaces to Enhance Human Understanding of Government Statistics
- Collecting and Using Geospatial Data in the Field

Full list of projects at website: www.digitalgovernment.org

Challenges in Technology Transfer

Traditional Transfer Process

- researchers work independently creating basic research
- basic research results passed on to organizations/vendors to be realized as a commercially-viable product

Traditional model of transfer is inappropriate because:

- Expertise is equally shared among agency personnel and researchers
- Researcher solutions need to fit within agency constraints
- Researchers can not provide fully realized products, vendors don't see market for products so role may fall to agencies

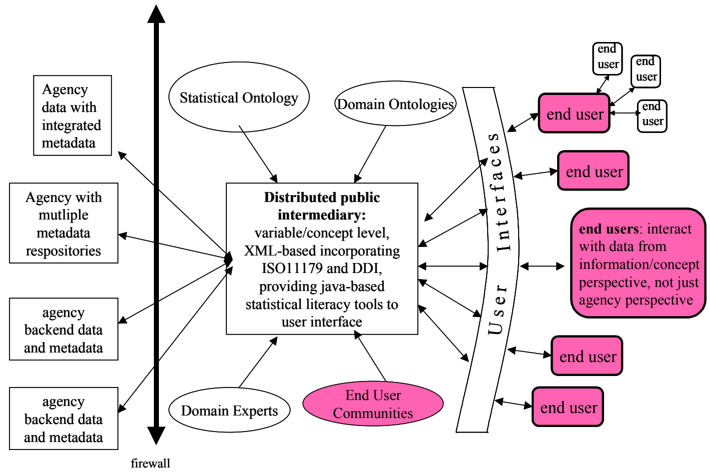
Technology Transfer: The Role of Researchers in Metadata Activities

Identification of user-centered metadata needs

Providing basic research in:

- ontologies/terminology systems using statistical metadata
- Information retrieval and database systems that might support integration efforts

Investigating metadata usage in end user tools



Metadata and End-User Tools: An Extended Example

- A user's task suggests specific metadata incorporated into end-user tools with specific functionality
- The example: A high school student needs to find economic indicators to understand how the economies of her county and state compare to that of the United States as a whole.

User Needs

- To understand what statistics might be appropriate for task
- Understand the statistics retrieved
- Whether seasonally-adjusted numbers can be compared to non-adjusted, etc.
- User problems identified in a user-study recently completed
- The SKN resolves these problems by providing appropriate metadata or metadata-driven tools.

Specifically, the SKN might need to provide:

- The definition of "economic indicator"
- Definitions of specific indicators
- Example economic indicators
- Mappings of geographic entities about which she might find data (such as county, city, MSA)
- Available geographic granularity for a given indicator

- The currency of a given indicator (e.g. latest available date or periodicity of indicator release)
- Rules for valid comparisons

Associated Functionality

- Retrieval of data for user-specified indicators
- Display of definitions (in context)
- Display of relationships among geographic units
- Display of geographic granularity for a given indicator
- Display of indicators for which desired geographic granularity is available

Metadata and Tools

Information to Support the User	Possible "tool"
Definition of economic indicator, Definition of specific indicators	CMR; Ontology, Glossary, Terminological system
Example economic indicators	CMR; Ontology, Glossary, Terminological system
Mappings of geographic entities about which data might be found (such as county, city, MSA, others)	CMR in conjunction with TIGER-based tool

Functionalities and Tools

Function	Tool to Support
Retrieval of data for user-specified indicators	CMR
Displaying of definitions	Statistical Interactive Glossary (SIG)
Displaying relationships among geographic units	A mapping tool
Display geographic granularity for a given indicator	relationship browser
Choosing geographic granularity and displaying indicators available	Relationship browser

Next Steps:

- Further articulation of the modelling effort
- Consideration of scalability
- Incorporation into specifications for public intermediary

EU Indicators - Compliance with EU Requirements as Metadata Problem

Zdenko Milonja
Central Statistical Bureau of Croatia
MilonjaZ@dzs.hr

In Central Statistical Bureau of Croatia there was a need for estimation of the time and effort needed for harmonization with Eurostat and other international requests for statistical data. A project has been developed where the metadata is the tool for estimating the size of the project.

This presentation mainly stresses on the efforts to comply with the international standards and the steps taken for this purpose by the Central Statistical Bureau of Croatia.

The very first steps in that direction were done by the international consultants. They defined the global frame for future work and set the global compliance with Eurostat statistical system. They have produced the documentation about status of Croatian statistical system but that requests the further elaboration and estimate of future activities.

The difficult management task was on the horizon. How to manage the important project without the better knowledge about resources needed and the estimation of available time. We found that the main problem is to measure the actual distance between the European and Croatian statistical system. The question was how to do that having in mind that we need the objective measure what we can observe over the time.

We can define the compliance problem as metadata problem. The difference between two statistical systems is the difference in metadata; the difference in data is obvious and trivial. In a project we would like to declare how the difference could be observed in metadata and what the possible measure of similarity (dissimilarity) could be. Thus, the required steps were considered as follows:

- Collect the data about data (EU / CBS)
- Define the difference (in metadata)
- Measure the difference

We consider inequality could be observes in four dimensions. The first, and probably the most important one, is the <u>inequality in statistical variable definition</u>. If we try to find where the possible inequality could be the basic element of the survey is variable and two surveys could have the two variables, which are not the same variable in both surveys. As second dimension of inequality we consider <u>the possible inequality in value set, but only for qualitative variables</u>, because the different values in quantitative variables does not make the difference. The third dimension is <u>a unit of measure.</u> Normally, it is possible to recalculate the values to the different measures, but this fact has to be observed as inequality. Finally, the last dimension of inequality is <u>the precision of measuring.</u>

The purpose of the survey or statistical system is to measure something. In that sense we can think about the concrete variable or concrete value set as an instrument. What we can see comparing the instrument for two surveys (statistical systems)? There are three possibilities: i) the instrument can be the same, (for instance, the same definition of the variable), ii) the instrument can be derivable and iii) the instrument can be different. The derivable instrument is the instrument in one survey, which can be derived (computed) from the instrument in another survey.

For example: if one survey observes enterprises with more 15 employees, and another survey observes the enterprises with more than 50 employees we can say that the instrument for the second survey (>50) can be derived from the first survey (>15). Of course, we have to have the variable with number of employees. If we cannot do such a computation we shall say the instruments are different (not derivable).

The algorithm for obtaining the distance for the whole statistical system starts *with every variable* in every statistical survey. The first step is estimating the difference in the definition of the statistical variable (equal, derivable, not derivable). If the variable is quantitative then we shall calculate the difference in the value set. Otherwise, the variable is qualitative and we shall calculate the difference in unit of measure and the difference in measurement precision. The calculation has to carry out for the whole statistical system, for each statistical variable.

Probably the easiest way to do the calculation is to establish a table (matrix) with a row for each statistical variable. On the right hand side we have three columns, one for equal, derivable and not derivable relationship. The experienced person, subject matter specialist could mark the appropriate column with a mark reflecting the actual situation with statistical variable. In that way we can mark all of the variables in the statistical system.

Calculation

Stat. variable	=	⇨	~~⇨~~
Unemployed person	X		
…			
Classification for domestic animal			X

For the management purposes it is useful to have the estimate of the resources needed for executing the task. If we establish a matrix which says how many man-days is the estimate for solving the problem of equal, derivable and non derivable variables. Of course the effort for equal variables is zero, because there is no need for any change. The same

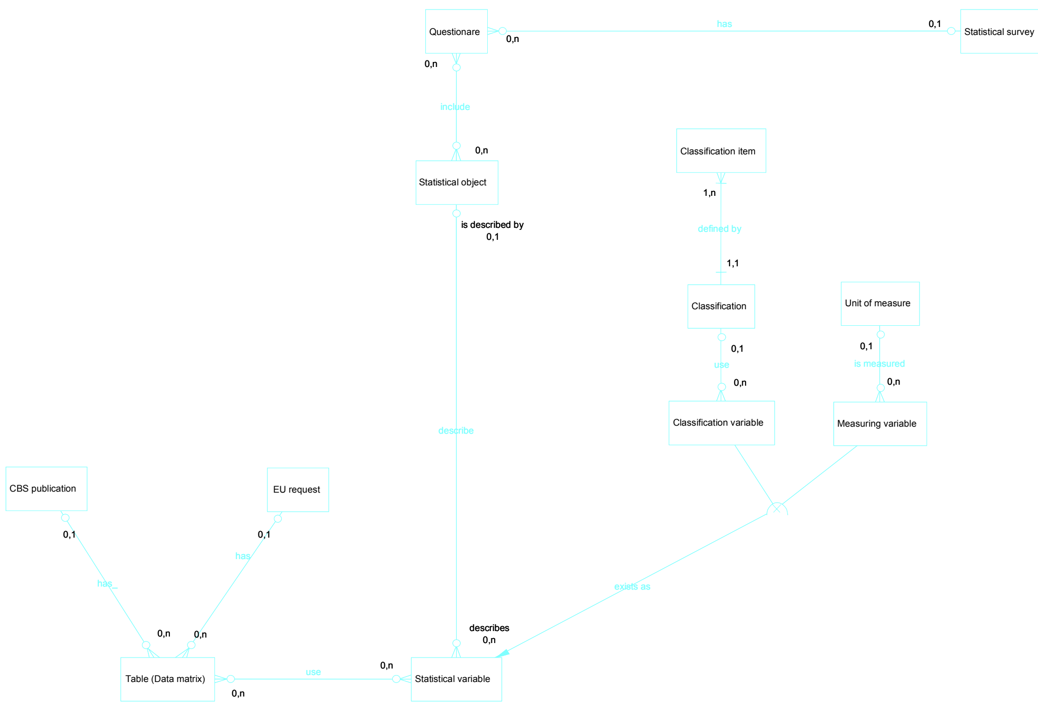
Dimensions	=	⇨	~~⇨~~
Variable definition	0	1,2	3,5
Value set	0	0,4	1,8
…			

numbers shows the effort in [man/month]

estimate must be established for the value sets and other instruments. Multiplying this matrix with the matrix from the previous step we can calculate the total effort for the compliance project.

Concerning project phases the database model for the metadata is developed and implemented. This was a first step in project development.

The main metadata elements are shown in the database schema. The schema is implemented in Microsoft SQL Server, except the part for classification variables which is implemented in the BRIDGE system (RUN software).



The next phases, some of them has already been developed are:

- Software development – software is developed in Microsoft Visual Basic and covers the input data for the statistical variables and other elements from the schema.

- Collecting the CBS variables – the pilot phase for five surveys has been conducted and the input variables have been collected. Now we start the same work for the output statistical variables.

- Collecting the EU variables is planned for September 2003.

- The last phase of the project is planned for the end of the year which concludes the scope of this project.

INTERNATIONAL EXPERIENCES AND PRESENTATIONS OF GENERAL INTEREST

Models and Metadata

Andrew Westlake, Survey & Statistical Computing, UK
ajw@SaSC.co.uk

1. Introduction

Models are abstractions from real-world situations, designed to support some particular context. With Statistical Metadata we are mostly concerned with software to support the processing and analysis of statistical information. Models provide the opportunity to specify how information can be shared between stages of processes (so that later stages can make use of information entered in earlier ones) and how information and specifications can be moved between independent applications. Because we are supporting the development and use of software, our models need to be detailed and precise in their specification of the structures and semantics of the information. However, the model also determines a conceptual framework for process designers and software users, so they must be able to view elements of or generalisations from a model, with less detail than is needed by software developers. Furthermore, when developing a model we need to work with domain and subject specialists to discover their needs and to help them to agree on model components and structures. These people will probably need assistance to express this knowledge in ways and with sufficient precision for use in the model, and will need help in understanding the model representation of their knowledge, so that they can confirm that the model represents this knowledge correctly.

1.1 Statistical Metadata

For completeness, we reproduce here our definition of statistical metadata.

*Statistical Metadata is any information that is needed by people or systems to make proper and correct **use** of the real statistical data, in terms of capturing, reading, processing, interpreting, analysing and presenting the information (or any other use). In other words, statistical metadata is anything that might influence or control the way in which the core information is used by people or software.*

It extends from very specific technical information, used, for example, to ensure that a data file is read correctly, or that a category in a summary table has the right label, or that the design of a sample is correctly taken into account when computing a statistical summary, right through to descriptive (intentional) information, for example about why a question was worded in a particular way or why a particular selection criterion was used for a sample.

Thus, metadata includes (but is not limited to) population definitions, sample designs, file descriptions and database schemas, codebooks and classification structures, processing details, checks, transformation, weighting, fieldwork reports and notes, conceptual motivations, table designs and layouts.

1.2 Acknowledgements

The ideas presented here have been expounded and discussed during the MetaNet project, particularly in work groups 1 and 2 and at the final conference. Of particular importance is work with Chris Nelson, of Dimension EDI, but the author takes full responsibility for all the ideas and opinions expressed here.

2. The role of Modelling

Models are abstractions, designed to meet a particular need in a particular context. Thus the form and roles of models can be very different. Some examples may help to show some of the range.

Conceptual Models are an attempt to form a frame of reference for some domain or collection of constructs or concepts. These are often similar to classification structures, and such structures (for example the International Classification of Diseases – ICD – or NACE, the General Industrial Classification of Economic Activities within the European Communities) can be seen as conceptual models.

The *Relational Database Model* is a formal specification of the structures and behaviour for databases formed from sets of rectangular tables. This provides a conceptual framework for thinking about databases (one that is widely used) but is also sufficiently detailed and precise to be the basis for the implementation of many database software systems.

The *Object Oriented* approach is an alternative (more general) way of thinking about databases and program structures (an alternative paradigm), built using a different set of primitive constructs, assumptions and conventions.

The statistical *Generalised Linear Model* is a mathematical specification of the way in which a set of predictor variables influence a dependent variable, together with the form of the variability about that relationship. This model is very flexible and is widely used for estimating statistical relationships (using suitable software to calibrate the model to a particular data set), and for discussing the potential form of such relationships. Of course, there are many situations where the GML is not an appropriate form of model.

Structural Models concentrate on the objects and attributes that are used to represent information structures. This is necessary for the exchange of information between systems, but needs to be accompanied by clear specifications of the intended purpose and use of the various elements. Inconsistent interpretation by independent users or implementers working with such a structure is a continuing concern, unless some enforcement mechanism can be specified and implemented.

With statistical metadata we are looking for models that allow us to interchange information between processes and systems and that provide a stable conceptual framework for users to work with complex information structures across processes and systems. We want to support users of statistical systems, support the automation of statistical processes, and exchange information between systems and processes.

We can have more than one model, focussing on different parts of the statistical process, but they should dovetail together when a wider picture is needed. And we should aim to get suitable models accepted as standards, agreed and used across the statistical domain.

3. Elements of Models

To these ends, we need models that provide formal specifications of components and relationships, avoiding misinterpretation. They must address:

- Structure: how are the elements organised, how are elements grouped and related, what attributes are needed for each type of element.
- Semantics: what do the elements represent, what rules and constraints apply to their attribute values, to their states and to the way in which they are used.
- Methods: specifications of algorithms and processes that apply to the elements and the data they refer to.
- Concepts: complete and detailed definitions of the terms and concepts that are the subjects and objects covered by the model and of the relationships between them. In some situations this may correspond to the idea of a thesaurus.

To construct models quickly and accurately we also need a modelling framework or workbench, which provides generic building blocks for model components and tools to support the design process.

4. Levels of Model

Models exist at various levels of abstraction, and confusion can arise from not recognising the level to which a particular construct contributes, or at which a discussion about the model is taking place.

For example, the metadata about a particular survey forms an instance of the more general model that can be used to describe other surveys (of the same general type). This in turn will draw on both a conceptual model of the application domains for which the model is appropriate, and on a more abstract model of statistical processes and surveys in general. These abstract models of statistics are sometimes called meta-models, and are themselves constructed as instances of an even more abstract model for the process of defining models.

Once grasped, the fact that there are different levels of model does not need to cause confusion, but the failure to recognise the levels can be very confusing.

5. Existing Models

Within the MetaNet project we have identified a large number of models for metadata, most of them addressing a general problem but within a limited domain of application. Some are extensive and detailed, such as the DDI Codebook for documenting datasets from surveys, others much more limited in scope, such as the *triple-s* proposal for survey data exchange. The deliverable from work group 1 contains details of most of the models and related systems identified by the MetaNet project.

Two extensive and more general models have been developed during MetaNet.

5.1 The UMAS proposal

Karl Froeschl, Wilfried Grossmann and Vincenzo Del Vecchio have produced a significant report on the concept of Statistical Metadata (MetaNet Deliverable 5) which includes the specification of a model (UMAS) for the main components of statistical systems. This extends from the abstract concepts of populations and samples to detailed descriptions of datasets (similar to DDI).

The complexity of statistical metadata is a consequence of the many entangled facets of the subject. In order to help reduce this complexity, the following *five* (at least) *canonical* metadata dimensions can be singled out, giving rise to a five-way metadata framework:

(1) *Structure:* (the “entity” dimension: what things are); A break-down of metadata entities into classes like population units, populations, variables, value sets (including measurement units), values, datasets, and the (formal) interrelations between them.

(2) *View:* (the “role” dimension: the different ways things are considered);A four-fold distinction of metadata into semantic (determined mainly by substantive matter considerations), statistical (referring to statistical properties of the data), computational (referring to all issues of data representation) and administrative (referring to the organisational aspects of the institutions producing, processing, and using the data) roles.

(3) *Stage:* the “process” dimension: how and where things are used); Metadata in each of the statistical processing life-cycle’s phases such as data definition, data production, data transformation (a rather broad area itself including many subtopics such as data modification and data aggregation), data exchange, and data dissemination.

(4) *Form:* (the “material” dimension: how things are represented);A broad subdivison of metadata into intensional form metadata (i.e. more or less textual information about the data targeted at human interpretation) and extensional form metadata (i.e. a formal representation amenable to machine processing).

(5) *Function:* the “agent” dimension: the purpose things are used for); A pragmatic dissection of metadata according to the needs emanating from specific retrieval/usage scenarios, including also aspects of data quality/quality management.

5.2 The Terminology Reference Model

This system has been developed by Reinhard Karge, based on experience working with various other MetaNet partners and on other projects, to identify the important concepts and structures for discussing and representing statistical metadata. The resultant models are called Terminology Models.

Terminology models can be defined by defining names for statistical concepts such as **classification** or **classification item** and their attributes and relationships to other concepts (characteristics or details). Such two level terminology definition correspond directly to a conceptual metadata model were the concepts are considered as metadata object types and the details as properties (attributes and relationships) of these object types.

The meta-model for a terminology model is simple and easy to understand for non-technical persons. The terminology model consists of two object types: Concept and Characteristic. In contrast to other terminology definitions terminology models differ between context independent terms (concepts) and concept related terms (characteristic). Thus, context related terms might be defined with different meanings in different contexts or for different concepts. The terminology model should also include rule definitions for defining rules for concepts and characteristics.

This approach has been used to construct a generic (reference) terminology model. At an abstract level this defines some of the basic elements of a statistical system (focussing on statistical production rather than statistical theory), and it also contains a large number of specific concepts needed for specific statistical applications.

6. Representation of Models

Whenever software is built there is always a model that represents those aspects of reality that are implemented by the application. However, usually this model is not made explicit, and exists only in the implemented code. As software projects and systems became more complex, the need for proper tools to support software development (or software engineering) became more and more apparent, and various methods were proposed. In the late '90s an effort was made to bring the most successful and important proposals together, and this resulted in the development of UML, the Unified Modelling Language.

6.1 The UML standard

The UML standard was developed within the Object Management Group (OMG) as a way to design and represent object models, especially for software development. It is a collection of diagram types and components for representing various types of object and behaviour. It is a formal specification, with semantics and conventions for representation of every element of a model. The model and the diagrams exist separately but not independently – nothing can appear in a diagram unless it is in the model (so adding things to a diagram adds them to the model), and their role in the model dictates the way they can appear in the diagram..

UML recognises that complexity is at the heart of most modelling, and it provides specific functionality to support this. For example, the same items (whether classes or objects or some other element) can participate in multiple diagrams, with different emphasis or different level of detail or abstraction. This corresponds to the idea of views in relational databases, where the same information can be viewed in different arrangements to meet different needs, or to reveal different aspects of its structure or behaviour.

It also recognises that designs must exist at different levels of detail and need to represent different aspects of the behaviour of a system. This extends from User Requirements (in Use Case diagrams) through Class and Object definitions, down to coding and implementation (Statechart, Activity, Sequence, Component and Deployment diagrams).

The origin and emphasis in most UML descriptions is on software implementation, but there is potential for much wider application for the design of any system that can be conceived in terms of objects. It is rich, complex and extensible, and not tied to any particular implementation language.

A number of tools for designing in UML exist, and it is a requirement of the standard that they are able to exchange design information (which is done using an XML structure called XMI – XML MetaData Interchange). A limitation of this standard is that XMI contains only the specification of the model; it does not contain any information about the diagrams. This is due to be addressed in UML version 2.

Several design methodologies have been developed (generally for software development), consisting of rules and guidelines about how to design good systems. UML thus provides a potential mechanism for a system to be designed in a way that supports interchange between development teams and extension over time.

6.2 Levels of Modelling

Models can be built at various levels of abstraction, from the description of a specific instance (say the metadata for a particular dataset), through the specification of the allowed structure (and behaviour) for a particular type of metadata, through the generic description of the types of structure that could exist in a model for metadata, right up to the specification of what it means to build a model.

The OMG approach explicitly recognises these levels. Within this structure, an actual instance of metadata for a model is at level 0, while the model itself is level 1. The specification used to build the model is at level 2, often called a meta-model, and UML is an example. UML modelling tools allow us to produce models at level 1, by understanding the structure and semantics of a model at level 2 (the UML specification). Level 3 defines the components (classes, attributes, associations, etc) from which actual modelling frameworks (such as UML) can be built.

6.3 The Role of XML

XML (eXtensible Markup Language) is a method for representing complex structures as linear text in an XML Document. Because an XML document is just text it is easy to construct the technical layers of protocols for passing such documents between applications. This is very important for information interchange, since text files and streams are very easy to exchange. So XML overcomes an important obstacle to effective interchange, by providing a simple solution to the exchange transport problem. It is a major contribution to solving the plumbing problem associated with the interchange of complex information. Of course, we still have the problem of deciding what the proper structure is for the complex information that we wish to communicate and interchange.

XML is a syntax, within which we can build communication languages by choosing the vocabulary that is needed for a particular subject area. It is a *markup language* (based on a much older system called SGML) and the text contains marker tags, in the form <*word*> ... </*word*> – these identify the content (between the start and end markers) as being of type *word*. We choose the names for the tags (the vocabulary), so we can identify whatever we need to. The tags can be nested, so we can produce complex structures (and we can have sophisticated rules specifying what is allowed).

The expected structure for a document can be specified in one of two ways, a Document Type Definition (DTD) or the more recent XML Schema Definition (XSD); they are similar, but not exactly equivalent. An XML document that obeys the XML rules about matching tags is said to be *well-formed*, whereas one that conforms to such a definition is said to be *valid*. Both these specifications are restricted to the expected structure of the XML document – there is no way (except as comments) of specifying the semantics of the elements of the data structure.

6.3.1 Basic Structures

For very simple situations where there are few elements to be exchanged, and where the number of parties to the exchange is small, the structure can be designed as a DTD or XSD in any text editor. An example of such a simple structure is the *triple-s* standard for exchange of basic statistical data and metadata. The structures are sufficiently simple that the whole DTD or XSD can be written on one page and easily understood by anyone familiar with that technology. The problem of agreeing on the semantics of the structure can be addressed by verbal agreement between the parties, or by annotating the definition using comments.

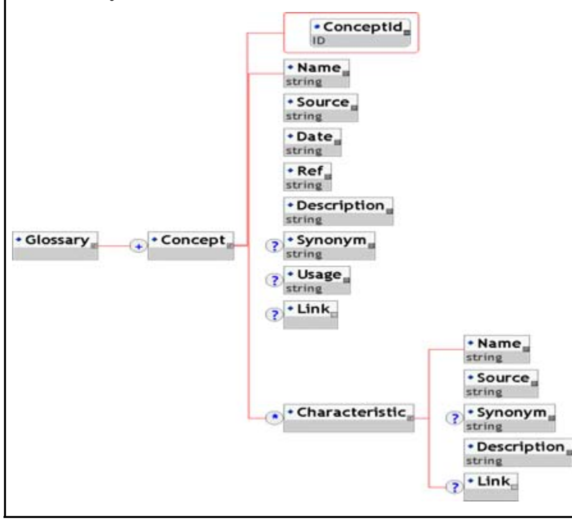
6.3.2 *Complex structures*

For more complex structures the use of some more specialised form of editor is recommended. Any XML editor can be used to design a schema to specify a structure, because an XSD is an XML document. This has the advantage that an XML editor will automatically check whether a document is well-formed.

If a DTD is to be built then a specific DTD editor is needed, because this is not an XML document. In general there are advantages in using specialised DTD or XSD editors, since they can know that the task is to design a structure, not just build a general XML document.

A number of specialised applications are available to support the design of document structures. Some are little more than text editors with syntax checking, while others provide much more extensive facilities (at a price). For example, in XML Authority (part of Turbo XML) a structure can be designed graphically, and then saved as either a DTD or an XSD. An example structure diagram is shown in Figure 1.

Figure 1: XSD Structure for Concepts in a Glossary



6.3.3 Semantics

Neither a DTD nor an XSD allows for the specification of any formal semantics for an XML structure (beyond the strong data typing of XSD). Comments can be inserted into the design, but these are not enforceable or executable, and it is not even sure that a new user of the structure will read the comments. The Codebook proposal from the Data Documentation Initiative (DDI - www.icpsr.umich.edu/DDI/CODEBOOK/index.html) includes some comments within the DTD, but is also accompanied by many pages of description of the intentions for the use and content of the many elements in the structure. More formal and extensive methods for reaching agreement on the meanings of terms can be employed, such as the Terminology Models mentioned later. These are useful, but remain limited to textual descriptions, albeit with more structure.

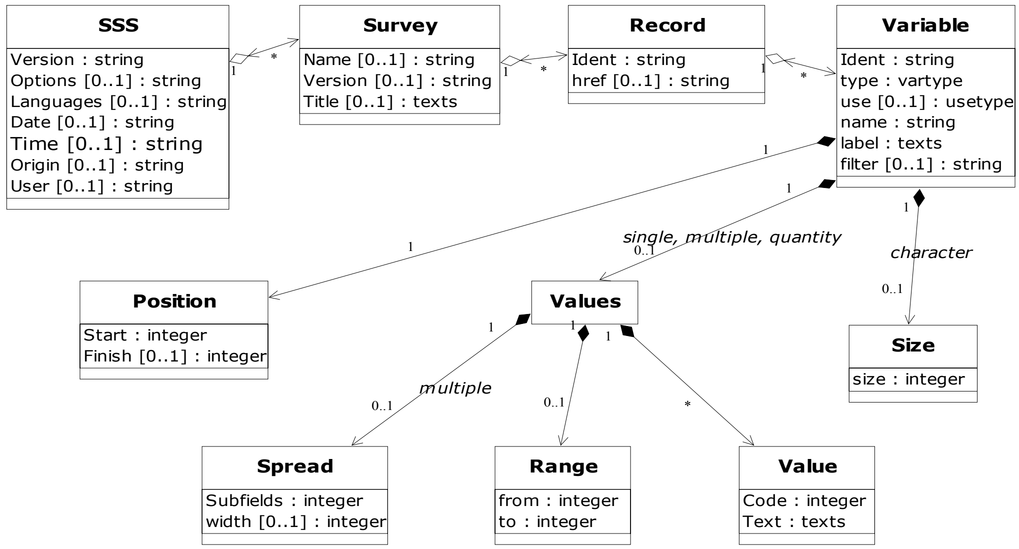
In contrast, Object Modelling techniques such as UML include the specification of behaviour as well as structure.

6.4 Examples

6.4.1 Class Diagram for triple-s

By designing in UML it is possible to express structure and the semantics associated with the behaviour of the elements of the structure. Figure 2 shows the data structure of the *triple-s* metadata proposal as a UML Class Diagram – this diagram does not include all the semantics that can be expressed in UML, and there are a few aspects of the *triple-s* semantics that cannot be expressed in UML. In UML, the various forms of shape and line all distinguish aspects of the semantics of the model. For example, the diagram shows (among other things) that the SSS, Survey, Record and Variable elements form a hierarchy, whereas Variable is a composition of elements (Position, Values, Size) whose presence depends on the Type of the Variable.

Figure 2: UML Class Diagram for the Triple-S structure



A structure designed in UML can easily be converted to an XML structure (though there are some possible difficulties), and the behaviour can be implemented in code that uses the APIs to access the XML documents. Not all semantics can be expressed in UML, so there is still a need to obtain agreement among users on the meanings of terms based on language, and here the Terminology Models are useful.

6.4.2 Special Values: an example of semantics

Missing values are commonplace in statistical data. In relational databases we have the concept of *Null*, to represent the absence of any information, but in statistics we often have some information, represented by codes, about why the expected information is not present. This idea we refer to as Special Values, meaning that such values in a field are not examples of the measure for the field, but are indicators of some other type of information. In the metadata we would want to identify such values as Special, and give some indication of their type. This is a difficult area, where agreement is not easy, but some examples of types of special value are:

- Refused
- Not Known
- No Answer
- Question not asked by interviewer
- Not applicable

One of these possibilities is fundamentally different from the rest. The first four indicate situations where the question was asked (or should have been asked) and code the reason why the respondent did not give any answer. The last indicates that the structure of the questionnaire is such that the respondent is not expected to answer the question (for whatever reason). This distinction has an implication when the data for this field is tabulated: for not applicable codes the respondent should usually be excluded from the tabulation (does not contribute to the denominator), whereas for applicable but missing codes the respondent should be included in some residual category. However, this may not be the case if the question was skipped because the answer could be inferred from some other answer.

All this indicates that there is considerable semantic content associated with this statistical concept, and the semantics need to be understood by systems that process or manipulate the data with reference to the metadata. These semantics can (generally) be represented explicitly in a model defined in UML, and then implemented in corresponding code. In pure XML the only option is to include them as comments, and to rely on the reader to take note and understand.

6.5 UML in practice

Almost all training and discussions about modelling now focus on UML. It is probably not reasonable to claim that UML is the dominant method for actual modelling work, since a great deal of modelling is still done using older tools and methods, but no one is promoting these older tools as preferable to UML.

With use, some limitations have been identified in UML. These can be addressed through the extensibility mechanisms, but it would be better if they were in the standard, to avoid duplicate or inconsistent extensions, and so that their semantics were implemented directly in the modelling tools. These issues will be addressed in the forthcoming design of UML 2.

UML is a rich and complex concept, and is not easy to learn comprehensively. Training material usually starts with the Class Diagrams, as the easiest to learn. This represents the static structure of the model, and corresponds to the entities (components) that are needed in a metadata model. Dependencies and relationships are shown in this type of diagram, and the structure can readily be converted to an XML structure definition to facilitate the exchange of information about an instance of a model between applications that implement the whole model.

Methods (behaviour) can easily be defined for classes, with their functionality described in words. Specifying the functionality formally (through state, sequence and activity diagrams) requires rather more familiarity, but is worth doing since it can then be converted into code that enforces the behaviour.

7. Model Development

The conclusion we draw from the forgoing is that UML (or some similar equivalent) is the correct tool to use for building models of statistical processes and the associated metadata. It allows us to express structure and behaviour with sufficient detail and precision for us to be confident that different implementations based on this model (if they correctly use the contained specifications) will be able to interoperate.

However this still leaves us with two problems. We have to discover the characteristics that are appropriate to include in the model for the area of statistics that we are modelling, and we cannot expect the domain specialists who have this knowledge in their heads to be able to express it directly in the terms needed to build a model, let alone to express it directly in UML. Once we have built systems from the model we will have to train users to use the systems correctly. That will involve them understanding the concepts and procedures built in to the model, and we again cannot expect them to do that by examination of the UML diagrams.

The extraction of domain knowledge involves the identification of concepts, structures, relationships, processes, constraints, rules, etc. This is based on discussion and agreement among domain specialists, and has to be done in terms that they understand. However the knowledge will need to be organised and expressed in a way that is coherent and amenable to transfer into a UML model, so the process needs to be moderated by modelling specialists, probably one who specialises in the domain area and one who is expert in the representation of models in UML.

Note that this elucidation and identification process implies the development of conceptual structures to represent the knowledge, and this may bring new insights to the domain specialists and alter the way they structure their thinking about the domain. This implies an iterative process that includes training the domain specialists to understand a perhaps more abstract expression of their knowledge than they normally use.

This process of discovering domain structures is akin to processes of User-Centred Design that are current in IT, the main difference being that we are trying to tease out underlying generalities, rather than reproduce some existing manual system.

The method of Terminological Modelling, espoused by Karge and mentioned previously, has a major role to play at this stage, as it provides a simple but formal structure to record concepts, definitions, structural relationships associations and semantics (in descriptions). These will need to be converted to more appropriate structures later for the UML model, but are sufficiently formal to allow precision without introducing unnecessary obscurity. Experience in developing models for various statistical processing components with the Nordic and other NSIs (for example in the Neufchatel Group working on Classification Schemes) shows the value of this method.

From such a domain-focussed specification the UML specification will need to be produced. In general, domain concepts will map into UML Class structures (usually not all at the same level of abstraction), and the semantics of these will need to be extracted from the concept descriptions, and refined (with iteration through the domain group) until the necessary degree of detail and precision is attained. Process specifications will similarly need to be developed and agreed.

Notice that as part of the process of getting agreement that the UML model is correct, some method is needed to map the content of the UML specification back to the terms that can be understood by the domain specialists. In part this can be done by training them to understand more of the UML, and producing UML diagrams focussed on their needs, but ideally we need tools that allow us to view the UML in less precise but more approachable terms.

8. Communicating Models

After developing a model in UML, and possibly implementing software and processes based on it, we need to be able to communicate appropriate aspects of the design to potential users of the model or the systems. The UML may meet some of this need for some users, but in general we will need simpler (less detailed and precise) presentations that meet the particular needs of users.

For example, users of a system based on the model will need introductory material about the principles of operation of the system, will need guidance on the operating procedures, and will need reference material about the rules and assumptions that are embedded in the model (and hence in the system). In addition there will be aspects of the conceptual view of the application domain that are included in the model (in the conceptual part) but which do not have any equivalent in the implementation.

In practice, getting users and operators of a metadata-based system to understand the concepts and nuances behind a model may well a more difficult task than building the model or the system. This is particularly true where a new system involves changes to working practices – unless these are seen as directly contributing to the substantive objective of the system. Similar considerations will apply in persuading other system designers and developers that a particular model is appropriate for their purposes. These issues would not normally be a primary concern for technical model designers, but must be included in the overall plan for developing a model and bringing it into practical use.

If a model is truly complete, then most of the detailed information needed by users should be present in some form. However, since the presentation of this material will need to be specialised to the domain of application of the model, it is not reasonable to expect that a UML modelling tool will come with appropriate views and reports built-in. There will be viewing and reporting tools available (diagrams and textual reports), but these will need to be used to produce the customised material required for the domain.

Technical and reference material can clearly be produced from the model using tools in this way, but explanatory material, which is not necessary for systems implementation, will only be present if has been intentionally included in the model descriptions. Some of this may flow from the conceptual modelling, but much will have to be specially written. This is hardly a surprise, since all system development should include the development of appropriate training and explanatory materials. Modelling with UML is no exception. It should be possible to include this material within the model, so that it remains closely linked to the more technical elements. The Holy Grail is to construct the informal materials in such a way that where appropriate they refer to the formal components and can automatically remain consistent if the model if modified.

9. Conclusions

UML is the ideal form to hold the master version of the design of models or systems related to statistical metadata. This is because it is able to represent precisely and in detail almost all aspects of the structures, processes, methods, constraints, relationships, interactions, etc. involved.

UML is too complex to be a natural form of communication except for technical specialists, so other forms of representation or documentation are needed to support the discovery of knowledge to feed into the model, and to support designers or users of related applications and systems. These are standard issues for User-Centred Design.

XML is the ideal format for the exchange of actual instances of statistical metadata between applications, because it is simple to transmit over standard protocols, and because it is well supported by manipulation APIs at the application level. However, it is not adequate to merely design an XML structure as specification of a model, since this cannot include a rigorous specification of the semantics of the elements of the structure, nor anything about processes or behaviour. Instead a suitable XML structure should be derived from the structural components of a full UML design.

Practical issues over the introduction and acceptance of new systems can be as important to an organisation as the detailed correctness of a model, so must enter in to the overall development process.

Appendix 1: The Object Paradigm

Much current (and recent) work on modelling is based on the Object-Oriented paradigm, and this is the approach assumed for UML.

An *object* is a structured collection of information, an *instance* of a particular component (such as a classification). An object must conform to its definition, and the general definition of a particular type of object is called a *class* (not a particularly good choice of name). The specification of a class determines the structure and semantics of the objects that are instances of that class – the objects can contain different information, since they describe different instances, but their structure and behaviour is the same[1].

The specification of a class includes the *attributes* which form its structure – these may be simple (such as numbers or strings) or complex (effectively links to and collections of other objects).

Every object (instance) has a unique *identity*, and this can be referenced by other objects. Object identities are global, so object references do not need different forms for different types of object.

Classes support the idea of *inheritance*, *specialisation* and *generalisation*. One class can be defined as based on another, so that it inherits all the properties (structure and semantics) of its parent class. New structure and semantics can be defined for the child, but only those things that are different have to be specified. The child class is a specialisation of the parent, which in turn is a generalisation of the child. In particular, this means that a child class is also valid anywhere that the parent class can appear in a structure or an operation (because it inherits all its' parent's structure and behaviour). A child class can substitute its own behaviour for that of its parent if appropriate – this is called *polymorphism*. For example, a child could respond to a 'print' command differently from its parent, because it has extended content and/or more specialised understanding of how this should be presented.

A class can be *dependent* on another, in that it needs to know about the structure and semantics of the depended class, so that it can make use of it. This is a one-way relationship. Where one object makes reference to another (of the same or a different class) this is called an *association*, and this is usually bi-directional. For example, a data cube may be constructed with reference to a particular classification for one of its dimensions. In the implementation of the model the dimension could contain a reference to the classification, and the classification may maintain a list of all the dimensions that reference it.

References can be traversed without knowing what type of object is at the other end. Generally it is useful to design an object structure (class) so that references are organised according to their type, but it is always possible to follow a reference first, and then find out what type of object has been reached afterwards.

---

1 The behaviour of an object may depend on the values it contains, but only in a way defined for the class as a whole.

Practical Problems of Implementing Metadata Standards in Official Statistics

Josef Olenski, Warsaw University, Poland
Jozef.Olenski@mail.nbp.pl

Presented by D. Prazenka

0. ABSTRACT

Metadata are information standards. In the era of the globalization of information systems there is the need for the metadata - driven standardization of infrastructural information systems. The effective strategy of standardization of official information systems of countries and of international organizations is introduction of official standards based on metadata. Statistical metadata are predestined to play leading role in that process. However statistical agencies are not ready to take that role and responsibility. The prerequisites of the statistical metadata - driven standardization and practical problems and approaches to the development and implementation of metadata driven standardization of informational infrastructure are discussed in the paper. General considerations are exemplified by practical experiences of the countries of the European region and of transition countries.

1. INTRODUCTION.

Homogeneity, integrity, coherence, transparency, precision, timeliness are the prerequisites of good quality of information and of information systems. Those qualitative features of information depend mainly on the quality of relevant metadata. Good metadata are the basis of good data. Practical experiences of many information systems have proven, that the quality of data may be provided by implementing good metadata - driven standards. It was also proven however that the elaboration and implementation of those standards is not easy. Many ambitious projects of metadata systems were not successful. Thus, metadata are essential for the standardization of information.

Why metadata - driven standardization of information and of information systems is so difficult ? What are the reasons of so strong resistance of information systems against standardization in practice while all developers and users of data systems declare their willingness and readiness to adopt all relevant standards and to subordinate their information and information systems to adopted standards ? What are real needs, requirements and limits of standardization of information ? What is the role of metadata and of metadata systems in the process elaboration of standards and in their implementation ? Who may and who should take the responsibility for the unification and integration of information and of information systems, based on the use of metadata as standardizing tools ? What strategies of metadata-based standardization of information systems in the economy are efficient ?

Those are the questions to be answered by the people and organizations involved in the development of information systems in the national economy. The list of questions may be continued. Many efforts have been undertaken to provide the unification and homogenization of data on the microlevel, i.e. on the level of microdata, and on the level aggregates of the national economy and on international and global levels. The results of those efforts are significant. However successful metadata projects are - as a rule - single metadata holdings, which are integral parts of larger information systems, or large metadata holdings, which are the part of official information infrastructure of the national economy.

Most of other projects of development of autonomous integrated metadata systems have not been successful. The question is why ? What should be the strategy of development and maintenance of metadata systems to enable the use of metadata as the instruments for organization, coordination, integration and standardization of economic and social information systems ? What should be the role of metadata standards developed by official statistics in that? We will try to answer some of these questions in this paper.

2. METADATA, DATA AND INFORMATION SYSTEMS.

2.1. METADATA AND DATA.

The following relations between data and metadata may be specified:

1) Metadata create (or generate) the meaning of data. E.g. the name of a statistical indicator creates (generates) the meaning of a given number - the value of the indicator ; without metadata the number would be meaningless.

2) Metadata interpret the data. E.g. the definition of a statistical indicator helps to interpret the economic or social meaning of an indicator.

3) Metadata identify the data. The identification should be differentiated from the address of the data. The identification helps to differentiate data and metadata from each other and is necessary to determine the characters which belong to the given chunk of data or metadata. E.g. the identifier of an indicator in the catalogues of indicators.

4) Metadata address the data. The addressing function of metadata is the collection of information where data are stored and how the place of the storage of data is labeled. E.g. the catalogue number of a statistical yearbook and the numbers of pages of the statistical table requested by the user.

5) Metadata organize the data. Metadata are used for the defining and description of the organization of information systems.

2.2. METADATA AND INFORMATION SYSTEMS.

From the point of view of metadata the following classes of information systems may be specified:

1) Information system in which metadata are integrated with relevant data (e.g. traditional data processing systems).

2) Information systems, in which data and metadata are organized in the form of separate files (e.g. database systems, data retrieval systems etc.).

3) Information systems in which at least some metadata are organized in the forms of separate subsystems (e.g. factographic information retrieval systems, knowledge based systems, expert systems etc.).

4) Entire metadata holdings (e.g. catalogue of indicators, catalogue of tables, glossary of terms and definitions etc.).

Modern information technologies, i.a. database systems, knowledge based systems, expert systems, etc. enable and require the organization of metadata in the form of metadata holdings integrated with respective data according to the specificity of the information system.

3. CLASSIFICATION OF METADATA

3.1. ELEMENTARY AND COMPOSITE METADATA.

Elementary metadata are the metadata which can not be divided into the meaningful part, e.g. name of an ethnic language, international alphabetic code of a currency, SI code of an unit of measurement, proper name of a national statistical office, an identifier of an enterprise in a business register. etc.

Composite metadata are the metadata which are constructed from elementary metadata according to the structural patterns (models) relevant for the respective metadata holdings, e.g. formatted description of an enterprise in the business register, description of a file in the catalogue of archived files, description if a statistical indicator in the data dictionary etc.

3.2. IMPLICIT AND EXPLICIT METADATA.

Implicit metadata are the metadata which are entirely integrated with respective data. They exist within a given information system associated with respective data. Implicit metadata are not organized in the form of separated metadata holdings. E.g. unit of measurement as the part of the name of an indicator, number of the year as the part of the date.

Explicit metadata are the metadata which exist separately from the data and from other metadata. They are organized in the form of metadata holdings. Most of the metadata may exist as explicit metadata.

Implicity or explicity of metadata is the attribute of an information system and not the attribute of metadata as such. In one information system some metadata may be implicit, while the same metadata may be explicit in the other information system. E.g. in a simple electronic data processing system the description of the structure of the input data stored on questionnaires is an integral part of data processing and is described implicitly in the computer program , while in the database system this metainformation is explicitly represented in the catalogues and data dictionaries/directories of the database system.

3.3. INTERNAL AND EXTERNAL METADATA.

Internal metadata are the metadata which are an integral part of a given information systems. They are generated by and for the particular information system. Those metadata may be used by other information systems and by other users, but on their own decision and responsibility. The use of internal metadata can not be mandatory or recommended for external use. E.g. the classification of age groups for the analysis of demographic processes by scientists is used for statistical purposes, but it classification of businesses by size on the basis by the number of employees for analytical purposes etc.

External metadata are the metadata which are generated by the information systems for the external use by other information systems. They often are elaborated for the standardization purposes. E.g. ISO SI system of the units of measurement, UNESCO classification of languages, GESMES standard for statistical table, COFOC classification of the SNA developed by the UN Intersecretariat Working Group on National Accounts and endorsed by the United Nations Statistical Commission, ISCO 88 Classification of Occupations of the International Labor Organization etc.

Internal metadata of one information system may become external if they are used by other information systems. In such case the external users of metadata should follow all updates of metadata introduced by the generic information system, i.e. the information system which generates the metadata as internal metadata. E.g. SWW, the nomenclature of commodities originally developed for statistical purposes by the CSO of Poland was adopted for tax purposes by the Ministry of Finances; in the same way the CPA based Polish statistical classification of products (including services) was adopted by the revenue service for tax purposes. The acceptance of updates of external metadata developed as internal metadata for different purposes, by different organizations, may cause some problems.

3.4. FUNCTIONAL CLASSIFICATION OF METADATA HOLDINGS.

For the needs of this paper the classification of metadata holdings based on the criteria of basic functions of metadata in an information system is proposed. From that point of view the following classes of metadata holdings are specified:

1. Registers of real entities (e.g. population register, business register, territorial register, tax payers register, register of automobiles, list of countries, list of languages, etc.).

2. Catalogues of data files (e.g. catalogue of archived files, catalogue of current files, catalogue of tables, time series etc.).

3. Classifications, nomenclatures, thesauri and typologies (e.g. CPA, CPC, NACE Rev. 1, ISIC Rev. 3, PRODCOM, typology of age groups, classification of forms of property, classification of occupations, statistical thesaurus, index of key words etc.).

4. Data dictionaries (e.g. catalogue of statistical indicators, list of indicators processes in a particular edp system, list of indicators of a bookkeeping system of an enterprise, data dictionary stored in a database system, etc.).

5. Data directories (e.g. list of addresses of files, catalogues, registers, descriptions of the access to databases, administrative records and metadata holdings, etc.).

6. Glossaries (e.g. glossary of definitions of concepts, terminological dictionaries etc.).

7. Organizational metadata (e.g. list of organizational units of an information system, functional specification of an information systems, organization of statistical data production etc.).

The above presented specification is not exhaustive. One should note that each class of the metadata holdings play different role in the standardization processes i.e. in the processes of defining standards and in the processes of their implementation in practice.

4. METADATA AS INFORMATION STANDARDS.

4.1. METADATA ARE STANDARDS.

All metadata are information standards. The standardizing role of metadata is an integral feature and function of any metadata. This role does not depend on the will of the data owner, on information system manager, on the database administrator. By introducing any metadata in an information system, information standards are introduced. Metadata are petrifying for some period of time the semantics of data, their structure and organization of information system, they create significant part of the language of users. Any decision concerning metadata is in fact standardization of information and of information systems. This fact should be well understood and always remembered by all information systems analysts, data processing programmers, data managers in a public sector and in businesses.

The standardization functions of metadata and their role in standardization processes of information may be different for each class of metadata. The standardization functions of implicit metadata are limited to concrete data and to the relevant respective part of a given information system e.g. standardizing role of names of indicators defined for one particular survey based on one statistical questionnaire are limited to this part of the official statistical system. On the other hand, the classification of products developed by official statistics and used mandatory for customs, taxes, for licensing of businesses etc. plays the role of the information standards for many information systems: official statistics, tax services, local governments, customs, and a large number of businesses (tax payers, statistical units, exporters and importers etc.).

The standardizing role of metadata is also related with the information system in which the metadata are generated and used. This role depends on the functional specification of the respective information systems. Autonomous metadata holdings play rather often the role of standards for other information systems. Principal function of some metadata holdings is the standardization and unification of information systems, e.g. it is one of basic functions of catalogues of statistical indicators, tables and time series, catalogues of classifications and of classifications and nomenclatures (instance level), catalogues of statistical units, territorial register, metadata based surveys frames, .

4.2. CLASSIFICATION OF METADATA-DRIVEN INFORMATION STANDARDS.

The following classes of metadata driven standards are specified:

A. FORMAL STATUS

1. Standards de iure, i.e. standards adopted by respective authorized institutions as the official standardization documents, e.g. ISO standards for metadata elements, statistical code list of territorial units introduced by the decision of the official statistical agency, standard for description of an economic entity in the official business register etc.

2. Standards de facto, i.e. the metadata element which, because of their common use, are standards and are accepted in practice by the users as standards, although they were not introduced officially, e.g. the format for a statistical questionnaire used in the BLAISE software becomes de facto standard for all users of the BLAISE.

B. FRAME OF REFERENCE

1. Local standards, used by one information system as internal standards. Local standards may be used by more then one information system or more units, but the decisions of using local standards are taken independently by each user, e.g. standard format for the questionnaire for face to face interview adopted by an official statistical agency is mandatory standard within that agency only; other agencies may use that as their own standard on the basis of their own decisions.

2. Regional / branch / discipline related standards, used regionally by all relevant users of the region / branch / discipline of research, e.g. code list of territorial units of the region developed by the local government, classification of disciplines of sciences .

3. National standards referring to the national economy, e.g. official national standards ANSI, DIN, PN for metadata elements: date and time, names of countries, codes of languages, codes for units of measurement etc.

4. Supra-national and international standards, adopted by respective international organizations (UN Statistical Commission, UNESCO, WTO, WHO, Eurostat).

C. PRECISION OF STANDARDS.

1. Specific standards: concrete classification, nomenclature, register, code list, glossary of terms.

2. Generic standards: standards which determine metadata models, e.g. standard format for statistical classification, standard for representing names in the business register or in the catalogue of indicators, GESMES, etc.

5. PRACTICAL PROBLEMS OF THE DEVELOPMENT OF METADATA - DRIVEN STANDARDS .

5.1. GENERAL REMARKS.

In official statistical agencies the following practical approaches to the development of metadata driven standards are adopted:

1. Single survey: metadata are developed for one specific survey. This approach is adopted for very large surveys (censuses of population, industrial censuses, agricultural censuses), for the surveys which are conducted one time only (if they are repeated, the methodology is different) or for the surveys with very low frequency.

2. Family of surveys. Some surveys are linked with each other by subject matter issues, by statistical units and by methodology. These surveys are or should be organized in the form of the families of surveys. In those cases the metadata are developed for the group of methodologically integrated surveys e.g. household surveys, industrial production surveys, agricultural surveys.

3. Subject matter area. Some metadata are developed not for a particular survey or a group of surveys, but for an area of social or economic analyses, e.g. national accounts, analysis of poverty and welfare, etc.

4. Autonomous metadata holdings developed for standardization purposes as a part of an information infrastructure, independently on any particular surveys or any particular analysis. Those standards are introduced for the whole statistical agency and should be used by any relevant survey.

5. External metadata generated by non-statistical systems used for statistical purposes. Those metadata may become statistical standards, if the statistical agency accepts it in their methodology.

In each case different approaches of development of metadata and different strategies of their introduction as standards are adopted.

5.2. SINGLE SURVEY METADATA.

Single survey metadata is the case beloved by traditional statisticians. The author of the survey is fully authorized to define the methodology and to elaborate any metadata relevant to the survey: classifications, nomenclatures, typologies, definitions and terms, statistical units, sampling procedures. He is not bounded by external standards. The standards developed are valid for one single survey only. Standardization is limited to one single survey or to the series of surveys of the same kind. So comfortable situation exists only in case of some social surveys and in new experimental applications in social statistics.

The elaboration of the metadata - driven standards for single autonomous surveys is very important and should be undertaken with the responsibility. Those new surveys and autonomous analyses are the places where the freedom for experiments, for trials and errors in statistics exists. Those are also the sources of progresses in statistical methodology and in the modernization of standards, in the introduction of new standards.

It seems to me that in each statistical subject matter area there should be found and offered to the statisticians the field for experiments which objective would be the elaboration of new methodological approaches and of new metadata representing those approaches in the form of metadata - driven standards. However the need of the right for experiments in official statistics is understood by very few official statistical agencies, happily those which are on the top of the ranking list of official statistics.

The other side of the coin is the responsibility for experimenting with metadata. It happens that those statisticians whom were given the possibility to experiment use this opportunity for using „trial and error approach" with much more errors then trials. I would like to stress that any proposal of metadata in an experimental survey should be made with the feeling of responsibility for potential dissemination of errors. In the practice of official statistics, also on national and international levels, one may find many examples of non optimal solutions and even errors which became standards, because the experimental results were adopted by other surveys or by other information systems.

5.3. FAMILY OF SURVEYS METADATA.

Two or more surveys are the family of surveys if and only if they use the same set of metadata covering:

- one subject - matter area
- definitions of statistical units
- basic classifications and nomenclatures
- basic terms and definitions.

All generic standards may be used by any survey belonging to the family of surveys. All metadata should be coherent for all the family of surveys. Organization of statistical surveys in the families of surveys is recommendable. It facilitates to coordinate the methodology of interlinked surveys, it provides better integrity and comparability of data, it economizes the costs and time needed for the elaboration and maintenance of common metadata base for many surveys belonging to the family of surveys.

The concept of the family of surveys is an efficient approach to the methodological coordination of statistics. Main problem is how to define the families of surveys in practice, how to divide the program of statistical surveys into the sets of surveys, what should be the criteria for defining the families of surveys. The basis for defining the families of surveys is the classification of subject - matter areas of statistics. All official statistical agencies have elaborated the classifications of subject matter areas of statistical information and surveys. However, in many statistical agencies there are used several classifications for different purposes:

a) subject - matter classification of surveys used for the structuring of the program of statistical surveys ,

b) subject - matter classification used for the classification data and tables in statistical yearbooks and other publications,

c) subject - matter classification used for the classification of data and publications in statistical libraries,

d) subject matter classifications used for the classification of data in statistical database systems, archives etc.

It seems to be necessary to adopt one and only one classification system of subject matter areas used for the classification of surveys, input data, output data and all data stored in a statistical system in archived files, database systems, libraries etc. This classification should be adopted as the standard for the official statistical system as a whole. It is not easy to convince the statisticians that the unification of classification of information by subject matter topics is necessary. The statisticians responsible for one or a few surveys to not see the need for methodological coordination based on the concept of the families of surveys. That is the reason why the metadata holdings which should be used as the standards for methodological coordination should be developed as external metadata holdings and maintained by special metadata service.

Each family of surveys has its own set of metadata holdings used for all surveys belonging to the family. There should not be overlappings or inconsistencies in the metadata for a specific family of surveys. Coherent set of metadata holdings for each well defined family of surveys may be used as the standard de facto, and may be also introduced as the standard de iure .

5.4. SUBJECT - MATTER AREA METADATA.

Subject matter area may cover one or more surveys, one or more family of surveys, and the data and methodology related with the frame of social or economic analyses. The defining of the subject matter area is not a statistical survey, but the analytical problem which should be solved by statistics. The difference between the concept of the family of surveys and the subject matter area may be explained by the following exemplification: in price statistics the family of surveys covers all surveys needed for the compilation of the CPI, while the subject matter area will cover all information and methods relevant for the analyses of inflation processes in the national economy, or the analyze conducted within the ICP (International Comparison Program).

The identification of subject matter areas is very helpful in defining real needs of comparability and integrity of statistical data. The basic tools for that are metadata. So, the defining of subject matter areas for statistics helps to determine the necessary limits of metadata - driven standardization.

5.5. AUTONOMOUS METADATA HOLDINGS.

Autonomous metadata holdings are those which are developed and maintained independently on particular statistical surveys. They should be maintained and updated according to their own routines even when they are not used for any survey for some period of time. Autonomous metadata holdings are often used not only for statistical purposes, but also by non statistical users. E.g. business register, classification of products, HS/CN nomenclature of commodities etc.

Autonomous metadata holdings of official statistics are standards de iure. Their standardization function is of particular importance, because it refers both to statistics and to other information systems, including those which are the sources of information for statistical surveys, e.g. administrative tax records, social security records, custom declarations, administrative records of enterprises etc. Their development and maintenance takes into account the needs of statistics and of other non statistical users. From economic and social point of view, the non-statistical applications of autonomous statistical metadata holdings may be of dominant importance. E.g. the function of the HS/CN nomenclature as the basis for custom tariffs, the CPA based nomenclatures of commodities as the basis for taxes - from the point of view of many businesses and of the government - are more dominant then their use for statistical purposes. Statistical applications if those metadata holdings are often an additional function to their non-statistical applications, although official statistical agencies are responsible for the development and maintenance of those classifications and nomenclatures.

All autonomous metadata holdings should be developed, maintained and officially introduced as if they were information standards de iure, even if they are, for the time being, local standards de facto . Such conclusion is the consequence of the fact that the modern information technologies i.e. knowledge based systems are based on the autonomous metadata holdings.

Autonomous metadata holdings as standards should also obey the general rules for standardization. Those rules are represented by generic standards for statistical metadata. The proposals of generic standards for statistical metadata have been elaborated by the SCP Joint Expert Group on Statistical Metadata (so called METIS Joint Group) and by the CES Expert Group on METIS. More concrete proposals of generic standards for some metadata holdings and metadata element were developed by the UN/EDIFACT MD6 group under the leadership of the EUROSTAT, e.g. GESMES . The standards elaborated by other UN/EDIFACT groups are also of importance (e.g. SAD, standards for data transfer in banking and finances etc.).

5.6. EXTERNAL METADATA .

External metadata are the metadata used for statistical purposes which originally are generated, for non statistical purposes, as a rule by non statistical systems. E.g. CN nomenclature of products used for customs, classification of professions for collective bargaining, nomenclature of diseases used in medical sciences, register of tax payers, register of social insurance beneficiants, format of custom declaration SAD, format of travel checks etc.

External metadata may be

(a) single metadata e.g. standard format for custom declaration,

(b) separate metadata holding e.g. register of tax payers.

They may be developed and maintained by non statistical organizations, e.g. registers of tax payers are developed in many countries by tax services, classifications of professions and employees are developed by ministries of labor or by other offices not necessarily statistical. But often non statistical metadata are developed and maintained by official statistical agencies, e.g. in many countries the nomenclatures of commodities and services used for tax purposes or customs are developed and maintained by official statistical services. The fact that the given metadata holding is maintained by the official statistical agency does not necessarily mean that it is an internal statistical metadata holding. The maintenance of external, non statistical metadata by official statistical agencies has many advantages for statistics. It helps to obey by those metadata holdings the rules and generic standards for metadata. It also enables to control the coherence between the external metadata and the relevant statistical metadata.

The official statistical agency should be given the possibility and authority of adjusting the external metadata to the needs and rules of statistics. The development and maintenance of non statistical metadata by official statistical agencies is an efficient approach of the harmonization of non statistical information systems with official statistics.

External metadata are rather frequently used by statistical offices for statistical purposes. But statistical metadata are also often used for non statistical purposes, e.g. statistical classifications are used for tax purposes, statistical classifications of occupations and professions are incorporated into collective bargaining, labor market or social insurance regulations.

The establishing of feedback between statistical metadata and the metadata of governmental and para - governmental institutions may facilitate the development of information standards both for statistics and for other non statistical information systems. To achieve that, the „owners" of information systems which belong to the information infrastructure of the country and official statistical agencies should harmonize their metadata holdings and should agree the „division of responsibilities" in developing, maintaining and implementing metadata as information standards.

6. PRACTICAL PROBLEMS OF IMPLEMENTATION OF METADATA - DRIVEN STANDARDS .

6.1. CONFLICT BETWEEN STANDARDS DE IURE AND STANDARDS DE FACTO.

The owners of information systems have strong propensity to develop their own metadata driven standards without their formal adoption as standards de iure. The reason of that is simple: the procedure of modification of standards de iure is rather complicated and long lasting, while standards de facto seem to be easy to change. It is an illusion that standards de facto may be modified easier. The freedom of choice and the freedom of modification of metadata driven standards depends rather on the commonness of their implementation and use. The classification used for many surveys and administrative records even without forma establishing as standard de iure, becomes the common standard and can not be changes easier then any other standard, if one would like to be responsible. The procedures of change of standards de iure and standards de facto are different. The changes of standards de facto do not require formal legal acts and decisions, while the changes of standards de iure, also those of local character, need formal procedures, which are sometimes long lasting, complicated and expensive. But one should remember that the real costs of the change of metadata driven standards are not those paid for the elaboration of new standards, but for the replacing of an „old" standard by the „new" one in all information systems in which the standard is used, including the costs of modification of software, costs of training, costs of the increase of errors caused by the changes of metadata, causes of misunderstanding and of misinterpretation of data by end users.

The conflict between standards de iure and standards de facto is the conflict between the responsibility and irresponsibility. Commonly used metadata driven information standards de facto should be treated in the same way like any commonly used standards de iure .

6.2. CONFLICT OF COMPETENCIES.

Metadata driven information standards are usually used by many information systems. E.g. statistical standard classifications are used for data collection, for analysis, for the production of tables disseminated to many users, which include those data and metadata into their information systems and databases. Analyzing the frame of implementation of statistical metadata we should not forget the information systems of statistical units and the information systems of end users. It may occur that the metadata holding used in one statistical survey is in fact used also by many thousands of information systems of enterprises and by hundreds of information systems of users.

One metadata holdings are used as standards de iure by more then one information system, for both statistical and administrative purposes. E.g. the CPA based classification of products is in several countries used as standard de facto (or de iure) by official statistical agencies, and as standard de iure by revenue services or local governments and economic self governments. The requirement of maintenance and updating of that classification for statistical purposes may differ from those of tax authorities or of economic self governments. Those differences may cause the conflict of interpretations and updates.

The question is, who is competent to interpret the classification for different purposes, who is competent to initiate modification, how should the procedure of elaboration and approving of updates look like, who finally is authorized to introduced changes, and - on the other hand - what is or what should be the legal basis of official introduction of changes and their implementation for statistical and for administrative purposes.

To avoid such conflicts, the procedures of elaboration, evaluation, acceptation and introduction of basic metadata driven statistical standards, which are of common use for statistical and administrative purposes, should include as partners all responsible offices and authorities: governmental institutions, non governmental official organizations representing the interests of data suppliers and data users. Official statistical agencies should not avoid to discuss and to agree the metadata driven standards with all important users even if formally they are not obliged to do that.

6.3. THE „HEN AND EGG" DILLEMA.

The „hen and egg" conflict in the field of the metadata driven information standards appears when there is the „chain" of related standards which should be harmonized. The eternal question is, what metadata holding is the first one in the loop of the chain. E.g. NACE (of Eurostat) === PKD (national NACE based classification of activities of the Polish CSO) === KWiU (national CPA/PKD based classification of products) === SWW 96 (national KWiU compatible classification of commodities) etc. Other example: ISCO 88 (international classification of occupations of ILO) === KZ 95 (national ISCO 88 compatible classification of occupations) === KZiS 95 (classification of occupations used by labor offices) === KZW (nomenclature of professions used for education purposes).

To harmonize information systems it is necessary to control and maintain the homogeneity of all metadata holdings of each chain. The needs for changes and updates may appeared in any place of the chain. As a rule, the „owner" of each „ring of the chain" may be different organization e.g. official statistical agency, the ministry of labor, ILO, self governments etc.

It seems to be necessary to define for all basic metadata driven standards the „chains of metadata holdings" and to established formal procedures of harmonization and updating of whole chains instead autonomous metadata holdings. Statistical offices, developing their metadata systems, should structure their metadata holdings in the form of chains with references to all relevant information systems. Reference models for metadata driven standardization should also be based on the „chain model" of METIS.

6.4. THE „META - METAINFORMATION" GAP.

One of the frequent reasons of the inconsistencies of related information systems and of the lack of common use of adopted metadata driven standards is the insufficient knowledge of statisticians and other information systems developers on existing metadata driven standards, both de iure and de facto. Even ISO adopted or UN recommended standards are not commonly known by many statisticians. The gap in the information on metainformation , i.e. the meta - metainformaiton gap is growing parallel with the growth and globalization of information systems.

It seems that now the information system on metainformation and on metadata - driven standards is an indispensable subsystem of any official statistical agency. The meta-

metainformation should be accessible for all potential users. It should be disseminated by force as an integral part of relevant data and information.

6.5. THE COORDINATION GAP.

Any metadata driven standard has its owner who is responsible for its development, maintenance, updating, implementation, interpretation. Sometimes one metadata holding as the standard has more then one „owner" . Often, different functions of the maintenance of the standards are fulfilled by different agencies and offices. E.g. official statistical agency is the owner of the CPA based classification of commodities, which may be used for tax purposes. The CSO is competent to introduce updates and changes to the classification schema and its items, but some competencies of the interpretation of that classification may be reserved by law or only by practice, for tax offices. The same is valid for other metadata holdings used both for statistical and administrative purposes or the holdings to one metadata chain shared by many competent offices.

The competence gap between the „co-owners" of metadata holdings may cause growing inconstancies of information standards for many information systems. The competence gap is extremely dangerous for the integrity and coherence of information systems. All legal and formal regulations of official statistics and of all related information systems and the users of metadata driven standards, which create the informational infrastructure of the country or of the region should be harmonized by law.

7. OPTIMISTIC CONCLUSIONS.

7.1. INTEGRATION THROUGH STANDARDIZATION

Recently the information infrastructures are in the phase of transition from local, autonomous information systems to integrated, interlinked information networks. Standardization is the prerequisite of integration and interchange of data. Metadata driven standards are basic tools for integration and unification of the content of information, end user languages and ways of communication and navigation. Although the understanding of the role and importance of metadata in that process, and the necessity of joint development of metadata systems for joint use by many users is not common, there is growing the consensus, that the standardization of metadata and the use of metadata as common standards is the necessary approach to the development of informational infrastructure of countries, regions and of the world.

7.2. INFORMATION NETWORKS.

National, regional and global information networks stimulate the development and use of metadata as standards. One should distinguish two kinds of networks: universal networks, offering technical facilities of data interchange by many users(e.g. Internet), and specialized networks for the interchange of some specific data by specific users (e.g. networks for banks). The standards required by the operators of universal networks are of technical nature. The specialized information networks require not only technical standardization, but also standards for data formats and for metadata (e.g. SWIFT, MARC, DERWENT).

7.3. STANDARDIZATION OF INFORMATION TECHNOLOGIES.

Modern information technologies, i.a. database systems, knowledge based systems, expert systems, end user friendly interfaces, offered by software producers dominating on the information market, introduce standards de facto for the representation of information. The standardizing role of those technologies is not direct, but they create the background for unification and coherency of information. The use of homogenous metadata and data formats becomes easier and cheaper.

7.4. THE METADATA - BASED CASE TOOLS.

Most of the CASE tools are based on autonomous metadata holdings. Implementing CASE for the design of statistical surveys one may achieve the standardization of metadata and the use of metadata as standards for all surveys and information systems designed that way.

7.5. UN/EDIFACT.

Standardization of messages in the UN/EDIFACT projects contains the standardization of metadata. That is true for generic (e.g. GESMES) and specific (e.g. SAD) standards. One standards message introduces many standards for metadata holdings, practically for the most of data elements of the standardized format. E.g. by introducing the standard for custom declaration SAD the standards for over 20 metadata holdings were introduced. The role of UN/EDIFACT project for the standardization of metadata and for the introduction of metadata driven standards on global scale can not be overrated.

7.6. THE COMMERCIALIZATION OF INFORMATION SERVICES.

Commercialization of information services and systems also stimulates the standardization of metadata and the use of metadata driven standards. Commercial information systems must be effective, offering good and cheap services. Commercial information services understand well the benefits of metadata standardization then official information systems financed from the public sources. We are of the opinion that the process of commercialization, which also enters the public sector, including the commercialization of official statistical services, will also stimulate the metadata based harmonization and standardization of information systems.

7.7. THE INFORMATION CULTURE.

Wide use of modern information technologies by end users changes positively the level of information culture of the societies. More and more people are involved in the process of creation of information systems, data files, data bases. Public access to the information via local, national and global information systems, and the possibility of introducing the data to the public networks, helps to non - professionals to understand better the requirements of modern information technologies and systems.

Amendments 7, 8 and 9 of the Fundamental Principles of Official Statistics (The Resolution of the UN Economic Commission for Europe adopted 15th April 1992 and the Statement of the UNSTAT) have direct implications on the development and use of metadata as standards on national, international and global scale. The precision of those amendments is done on the level of laws and rules regulating the organization and functioning of official statistics and other infrastructural information systems.

BIBLIOGRAPHY

1. Fundamental principles of official statistics in the region of the Economic Commission for Europe adopted at the 47th Session of the ECE, 15th April 1992, Geneva.

2. METIS - Glossary of terms, SCP/UNDP/CES, edited by Prazenka D., Bratislava 1986.

3. METIS - the concept of the statistical integrated metadata system, SCP Expert Group on Statistical Metadata, SCP/UNDP/CES R.1. Geneva 1984.

4. Oleñski J., Information standards in official statistics, Proceedings of Euro-Info Poland, 1993, Warsaw 1993.

5. Oleñski J., Standard format for a statistical indicator, METIS Expert Group of the CES, Working Paper, Geneva 1992.

6. Sundgren B., Infological approach to data bases, Urval Skriftserie 7/1973, National Bureau of Statistics, Stockholm.

Warsaw, March 1996.

METIS – METAINFORMATION SYSTEMS IN STATISTICAL OFFICES

(A Synopsis of the ECE/UNDP/Statistical Computing Project and its two METIS related basic documents)

Dusan Soltes

Faculty of Management, Comenius University, Bratislava, Slovakia

dusan.soltes@fm.uniba.sk

INTRODUCTION

The years 1970s among various other development trends in statistics were marked also by the introduction of the new ideas and concepts on the metadata, metainformation and metainformation systems. This new development trends in statistics has been launched in particular by the three most visible representatives of the so-called Nordic school viz. S. Nordbotten of Norway, B. Langefors and in particular by Bo Sundgren of Sweden and his famous book on "An Infological Approach to Data Bases" (Urval no. 7) Stockholm 1973.

It has been no surprise that these new concepts have been introduced within and/or very close to the statistical systems as right they have always been typical by the enormous amounts of data to be identified, collected, processed, stored and disseminated to numerous users. So it has been quite logical that there was a natural need to find some new means and ways how to handle efficiently all these huge amounts of various statistical data and information more over collected from various and sometimes very different sources through various statistical surveys, etc. For achieving this goal it has been needed to have some new specific kind of data and information about the data and information itself i.e. data and information which would "depict" and/or "describe" other (object) data and information i.e. to have metadata and metainfrmation and both of them to be organized into a specific information system i.e. a metainformation system or the METIS.

One of important contributions to this ongoing process of introduction of these new concepts i.e. metadata, metainformation and metainformation systems have become among others also international ISIS – Integrated Statistical Information Systems – Seminars regularly organized since early 1970s by the Computing Research Center – UNDP Programme at Bratislava (Slovakia) under the sponsorship of the United Nations ECE – Economic Commission for Europe – and its CES – Conference of European Statisticians. During their course, there was spontaneously formed an informal, ad hoc group of experts from statistical offices of different countries as e.g. already mentioned B. Sundgren of Sweden, J. Olenski and T. Walczak of Poland, D. Soltes of the former Czecho-Slovakia, J. Philip and W. M. Podehl of Canada, K. Neumann and L. Rauch of the former GDR, R.van der Abeele of Belgium, R.B. Graves and J. Walland of the UK, D. Altman of the former Yugoslavia, M. Euriat of France, H.O. Hougast of the Netherlands, etc., who annually presented papers on the subject of metadata in statistics and discussed their overall impact on the future statistics, etc. Thus gradually metadata, metainformation and metainformation systems in statistics became one of the regular and very popular thematic blocks in the deliberations of these seminars regularly attended by the representatives of European Statistical Offices as well as of those from the USA and Canada and various other parts of the world.

On the basis of this development it is then no surprise that the growing cooperation among the statistical offices in solving the problems related to the sphere of metainformation systems finally led to a more formal forms of international cooperation. Finally, in 1981 a regular regional project on the METIS started under the framework of an United Nations regional project. It was under the Statistical Computing Project (SCP) where one of the four basic joint groups was devoted to the problems of metainformation systems or METIS. The SCP was launched by the United Nations Economic Commission for Europe in cooperation and by funding from the UNDP – United Nations Development Programme with the project's head office and coordination at Geneva. Just for illustration in addition to the Inter-country METIS Joint Group there were another three other joint groups of experts from statistical offices devoted to some other most important issues of the statistics at that time viz. Table Generation, Data Editing and Relational Approach to the Statistical Data Base Management Systems. The SCP and also the METIS Joint group under the leadership of D. Soltes was launched in 1981 and completd its activities in 1984 so it has been in operation for four years.

In the end of its four years joint work and project activities the METIS Joint Group has , as already mentioned above, produced two final products – official United Nations project documents:

-ECE/UNDP/SCP/H.4 - User Guide to Metainformation Systems in Statistical Offices (144 pp)

-ECE/UNDP/SCP/H.6 - Selected Chapters for Designing METIS in Statistical Offices (131 pp).

USER GUIDE TO METAINFORMATION SYSTEMS IN STATISTICAL OFFICES

The first of those two basic documents as produced by the ECE/UNDP/SCP/METIS Joint Group has been "The User Guide to Metainformation Systems in Statistical Offices" published by the United Nations Geneva in October 1984. The User Guide consisted in addition to the Foreword and six appendixes altogether of five chapters where the most important issues of METIS for statistical users have been processed as follows:

Chapter 1: Basic Concepts and Purpose of Metainformation Systems

At the beginning, the basic concepts related to metainformation have been defined on the basis of B. Sundgren's works and his infological approach i.e. what are and what are differences between (statistical) data and information on the one side and metadata and metainformation on the other side. Following that, the metainformation systems in statistical offices i.e. "METIS for SIS" has been characterized. In subsequent parts of this chapter, the impact of computerization on statistical data processing and metadata handlings has been dealt with as well as the issue of the current state of computerization of metainformation in statistical offices. In the remaining parts of this chapter the categorization of metainformation systems users have been dealt with from the point of end users, managers and designers (of SISs). In the end, the chapter dealt with the objects of metainformation systems as well as with the proposed approach to metainformation system from the point of view of formalized and non-formalized statistical data. As far as the objects of the METIS for SIS are concerned, the following main objects have been identified:

- a SIS as a whole
- subsystems and processes of the SIS
- pieces (elements) of data and information
- logical and physical constituencies of data/information as handled by the SIS
- any other relevant entities related to above objects of the SIS.

In practice of the METIS for SIS it means that METIS should contain "metainformation" on such objects as :

- statistical surveys
- statistical units
- populations of statistical units
- variables/indicators
- data files and bases
- tables and publications
- computer programs and manual routines.

Chapter 2: The Information Function of METIS

In this chapter, the basic function of METIS for SIS have been identified and defined. Among them the following basic functions have been defined in particular relating the object statistical data/information:

-information function itself as the most important one

-identification sub-function

-interpretation (of statistical) sub-function.

In this connection the following main metainformational objects of primary interest have been identified:

-socio-economic phenomena and processes

-statistical indicators

-statistical surveys

-statistical units

-populations of statistical units

-classifications, nomenclatures and code lists

-time series

-statistical publications and tables

-rules and algorithms

-data files/data bases

-users.

The chapter deals also with basic sub-functions of METIS for SIS and in particular with the following main sub-functions:

-locational sub-function – general/overall navigation in the repository of data of a national statistical office

-navigational sub-function itself may work in two basic modes as the SLS – a simple locator system or DIS – detailed informing system

-retrieval sub-function – as a part of the particular DBMS it works on the direct retrieval of data from the data bases.

Chapter 3: Basic Forms of the Metadata Holdings and Instruments in the METIS

The chapter has been devoted to the following main parts. The first one deals with the basic typology of metadata holdings:

-catalogue

-dictionary

-register

-directory.

In addition to the above main four types of metadata holdings, the chapter specifies also some other more specific types of (statistical) metadata holdings such as:

-glossary

-thesaurus

-classification

-nomenclature

-code-list

Among the typical metadata holdings for SIS the following metadata holdings have been described:

-master catalogue of the METIS holdings

-catalogue of statistical indicators

-catalogue of statistical surveys/forms

-catalogue of statistical populations

-catalogue of socio-economic classifications, nomenclatures and code-lists

-catalogue of time series

-catalogue of statistical publications

-catalogue of statistical units

-catalogue of rules, algorithms for aggregation and calculations of statistical indicators

-catalogue of statistical output tables

-catalogue of data files

-glossary of terms of the socio-economic phenomena and processes

-thesaurus of statistical information

-register of users

-register of statistical unites – instance level.

For each and every of the above metadata holdings, the set of attributes have been identified and defined. Due to the overall extent of all attributes for the above holdings, herewith, only attributes for the catalogue of statistical indicators are listed:

-code of a (socio-economic) phenomena-process

-(identification) code of indicator

-code of unit of measurement

-code of time characteristics

-indicator name

-indicator type

-origin

-code of documents/statistical forms

-subject system code

-cross-section

-algorithm code for calculations

-codes of output publications

-codes of data files

-type of price (in case of indicators of value)

-definition of indicator

As stated above, in the same way the particular attributes have been defined and described in more details for each and every of the above metadata holdings accompanied by illustrative examples and more detailed specifications.

In the end of the chapter, the Conceptual schema of METIS – metadata base and mutual relations among individual metadata holdings with dividing into the (content) blocks have been presented.

Chapter 4: Usage of METIS Holdings

In this chapter the usage of the METIS holdings as specified in the previous chapter have been elaborated according to the basic three categories of users:

-users

-management

-designers

Each and every type of usage of METIS holdings has been presented in the textual as well as graphical forms of presentations regarding some most typical queries, links between individual metadata holdings to be involved, forms of output presentation, etc.

Chapter 5: The Role of the Statistician during the METIS Life-Cycle

The chapter deals with some basic questions and roles of statisticians to be assumed actively by them during the whole process of preparation, development, implementation and efficient utilization of METIS in the conditions of statistical offices. The main purpose of the chapter has been to clearly state that the development and efficient utilization of METIS in statistical offices is not possible without an active participation of statisticians themselves in the whole process. In this connection, two basic possible approaches have been presented in the chapter viz. the top-down and bottom – up approaches as the two most widely utilized approaches to the development of METIS in statistical offices.

In the end of the User Guide, altogether six appendices have been prepared and presented dealing with various practical aspects of the proper interpretation and utilization of all the previous five basic chapters in concrete statistical situations such as:

-examples of METIS usage

-finding indicators on manpower in statistical publications/tables

-the values of fixed assets per industry worker

-statistical data available for analyzing the seasonality of investment processes

-redundancy of data demanded by a proposed survey

-consequences of a decision to introduce a new version of the branch classifcation

-kinds of surveys for engine production

SELECTED CHAPTERS FOR DESIGNING METIS IN STATISTICAL OFFICES

This second basic output and/or document of the METIS Joint Group has been prepared as a specific, in difference to the User Guide, more technically designed and oriented document dealing with some most important aspects of the designing, development and implementation of a metainformation system in a statistical office.

There are altogether four selected chapters for designing METIS which in addition to an introduction and two appendices deal with the following four most crucial aspects of designing METIS:

Chapter 1: A Systematic Approach to the METIS Development

The chapter starts with some basics of the necessity of a systematic approach to the METIS development. There are five basic stages of the METIS development further elaborated in more details within:

-Preparatory stage – Setting up the task, specification of investigation and its planning

-Identification and analysis of user requirements for metainformation

-Design of inputs into the metadata base

-Design of metadata base

-Design of outputs from the metadata base

-A general procedure followed when designing METIS

Chapter 2: METIS Implementation

This second chapter deals with some selected and most important or crucial problem areas related to the issues of implementing METIS in a statistical office. In this respect there are discussed the following problem areas of METIS implementation:

-The basic strategy of METIS implementation as a top-down and/or bottom-up or step-by-step approaches to this specific kind of implementation

-Prerequisites of METIS implementation such as are creation of implementation conditions, planning and scheduling

-Organizational aspects of METIS implementation

-Selection of software and hardware for METIS implementation

Chapter 3: Special methods and techniques of metadata systems design

The chapter deals with some selected problem areas and/or approaches to the designing METIS in statistical offices regarding:

-designing statistical information languages including such issues as basic language forms of statistical metadata languages, classifications, nomenclatures, systematic lists, indexes, multi-faced classifications and thesauri

-analysis and synthesis of texts

-metadata links

-methods of statistical metadata coding.

Chapter 4: The Outline of the Jigsaw Methods

This final chapter is devoted to some non-standard methods and their assessment from the point of view of their potential future utilization in the designing METIS in statistical offices. In this respect, the chapter deals with the following issues:

-the general premises of the puzzle methods

-the need for some new approaches

-why the puzzle method

-an outline of the puzzle approach.

This document again as in the case of The User Guide is accompanied by two appendixes:

-Scenario of a dialogue for METIS

-Illustrative examples of the practical use of special tools and methods of statistical metadata design in the development of METIS holdings.

FINAL REMARKS

Although both documents have been prepared and finally published already in 1984 and 1985 respectively i.e. almost twenty years ago, they still represent an important source of information for all those who in one or other ways need to deal and/or to learn something on metainformation systems, metadata, metainformation, etc. In this respect these documents and in more general terms the whole joint work of the METIS Joint Group as conducted under the ECE/UNDP/SCP Statistical Computing Project has become an important source of information and advice for the designing, development, implementation and utilization of metainformation systems in numerous statistical offices not only in the region of operation of the United Nations Economic Commission for Europe i.e. in Europe and the Northern America but also in many developing countries what was after all also one of the main goals and reasons why the UNDP was funding this joint work although in that time the concept of METIS seemed to be quite a theoretical, abstract and futuristic one. However, all the further development in the statistics and in particular the rapid grows in implementing the latest computer networking and Internet into the daily practice of the contemporary statistics have fully confirmed the full relevance and practical utilization of metadata concepts as processed under the particular METIS Joint Group almost two decades ago.

The main positive aspect in the approach to METIS as carried out by the METIS Joint Group and as processed and published in the particular two its basic documents has been that it has in very systematic way and well ahead of time transferred the ideas and theory of metainformation as initially conceptualized by its inventors such as B. Sundgren to the practical guides for statistics and statisticians. The METIS in this sense has been the very first systematic attempt to transform those theories and concepts into the practice of the statistics. In this way, the METIS has well ahead of time responded to the future challenges of statistics vis-à-vis its needs for metadata, metainformation, as we know them in the current era of Internet and the World Wide Web.

Implementing Standards

Simon Musgrave
Nesstar, Colchester, Essex, UK
simon@nesstar.com

Structure of the presentation

- Introduction to Nesstar
- External user perspective
- DDI standard
- Linking the aggregate to the microdata
- Examples of external use
- Examples of production tools
- Conclusions

NESSTAR

- Builds upon the success of Information Society Technology projects (NESSTAR and FASTER) funded by the European Commission (Note: FASTER web site is not live, please see nesstar web site for contacts)
- Nesstar is exploiting the project outputs and is continuing to develop an innovative suite of 'DDI based' data publishing, discovery, analysis, and dissemination software
- Company formed to serve the wider market for web based metadata and statistical software
- Specialising in the management and dissemination of surveys and tables

Implementing Standards

•General approach: To use existing open standards whenever possible and to contribute to standards development and harmonization whenever appropriate

•Developed a Metadata Object Model (MOM) drawing on and including elements from:

–The Data Documentation Initiative (DDI)

–The Cristal Model (CBS)

–ISO 11179

•Contributed to the DDI development process

•Participated in metadata initiatives/processes like Metis, Metanet, Cosmos etc.

•ACU language specification delivered to OASIS standardization committee (XACML)

•The Nesstar Object Oriented Middleware (NEOOM) placed in the public domain

Driven by User perspectives

•User Categories (Joanne Lamb)

- –Management
- –Data collection
- –IT & Data Administration
- –Statistical
- –Publication
- –External users

•Simple web users

•Sophisticated web users

•External agencies

•Model will, inevitably, reflect organisational priorities

External User

•Types of data

- –Indicators
- –Tables
- –Microdata

•Complementary content

- –Information architectures
- –Factsheets
- –Reports

Characteristics

<ddi>
Metadata powered by the Data Documentation Initiative

- **End-user perspective**
  - provide the end-user with the information needed to locate relevant data sources and to use data-sources in a sound way
- **Initial emphasis on survey-data**
  - developed to describe independent surveys on study, file and variable-level (rudimentary support for other types of data)
- **Emphasis on codebooks (survey-data dictionaries)**
  - metadata seen as a complete “book” or document
- **Library-orientation**
  - strong on catalogue information,
  - mapping to Dublin Core

Structure

<ddi>
Metadata powered by the
Data Documentation Initiative

- **Document description**
  - describes the metadata document and the sources that have been used to create it
- **Study description**
  - information about the entire study or data collection (content, collection methods, processing, sources, access conditions etc)
- **File description**
  - describes each single file of a data collection (formats, dimensions, processing information, etc.)
- **Data description**
  - describe each single variable in a datafile (format, variable and value labels, definitions, question texts, imputations etc.)
- **Other Study-related Materials**
  - include references to reports and publications, other machine readable documentation (referenced by URI's) etc.

Achievements

<ddi>
Metadata powered by the
Data Documentation Initiative

- **Acceptance**
  - fast take-up in the community of data libraries and international development world-wide
- **Community building**
  - revitalised the co-operation and sharing of know-how and technologies
- **Strengthening of the ties to the data producers**
  - Direct link to CSPro and other tools under development
- **Software development**

The DDI in action – what do we know?

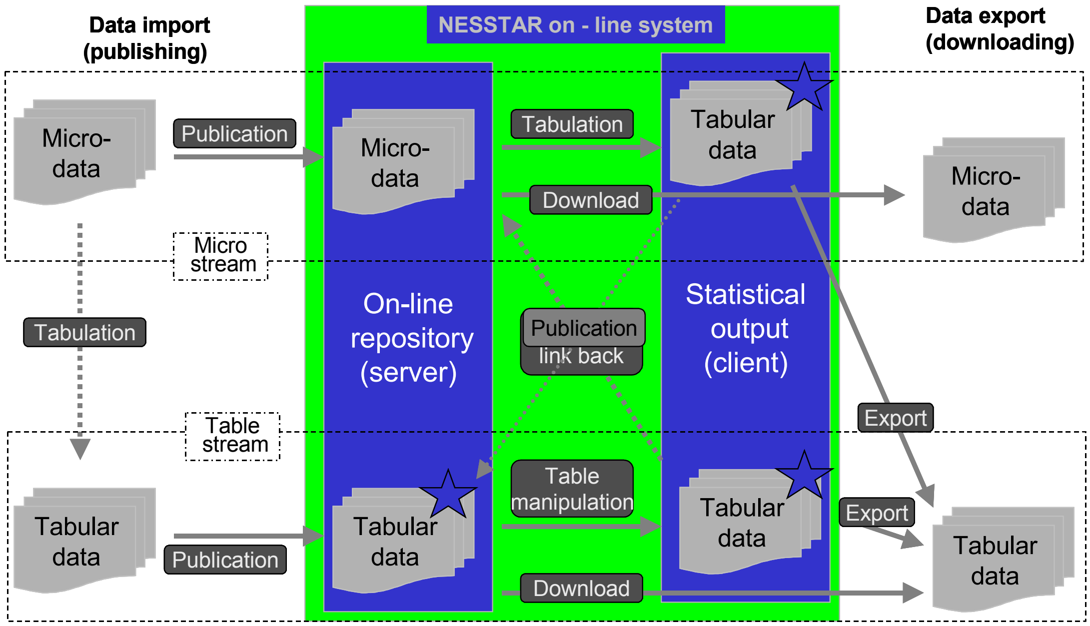
<ddi>
Metadata powered by the Data Documentation Initiative

- The costs of migrating a data archive to the DDI is high (much higher than the cost of any DDI-compliant archiving software)
- The DDI is a “cathedral standard” (no data provider buys the full package – they are all using the DDI building blocks to build their own more modest “parish churches”).
- The DDI is a very loose structure loaded with alternatives and ambiguities (a single study described by two different organisations will probably look quite different)
- The DDI only tells half the story (data providers have to add their own local guidelines on top of the DDI (controlled vocabularies, mandatory elements etc) to secure internal standardisation.)
- The DDI is inflexible (there is no extension mechanism that allows a data provider to add local elements without breaking the standard)
- A pure “bottom-up” approach: The DDI is used to describe concrete files or products coming out of the statistical process. It has no level of abstraction above or beyond a physical statistical product
- Machine-understandable versus human-understandable: Using XML does not automatically create metadata that is complete and logical enough to drive software processes

Main differences between ISO11179 and the DDI

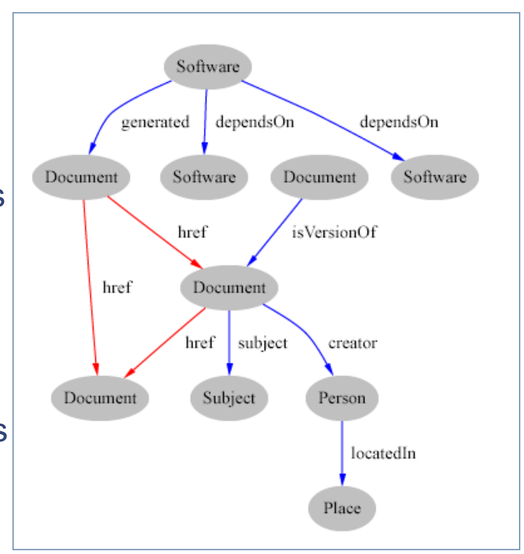
DDI	ISO11179
End-user perspective – to bring relevant metadata from the data producer/archive to the end user.	..data producer perspective. An instrument to document and administer concepts and data elements within and organisation. Central tool for the data production process
Represented in XML – no explicit data model	..a metamodel represented in UML, no bindings to a syntactic language defined yet
“Metadata after data” – only used to describe instances of a survey or a dataset.	“metadata before data” – mainly used to describe concepts and data elements independent of survey or dataset instances
Dataset (study) oriented	Data element or variable oriented
Document oriented – codebook metaphor: a document describing a set of data	Database oriented: relationship between elements in a constantly updated environment
Studies as the highest level of abstraction – variables live in the context of a study and cannot be abstracted out of that context	Concepts as the highest level of abstraction – variables/data items seen as representations of abstract concepts, which might be present in one or more instance of a study.

Linking the microdata to the table



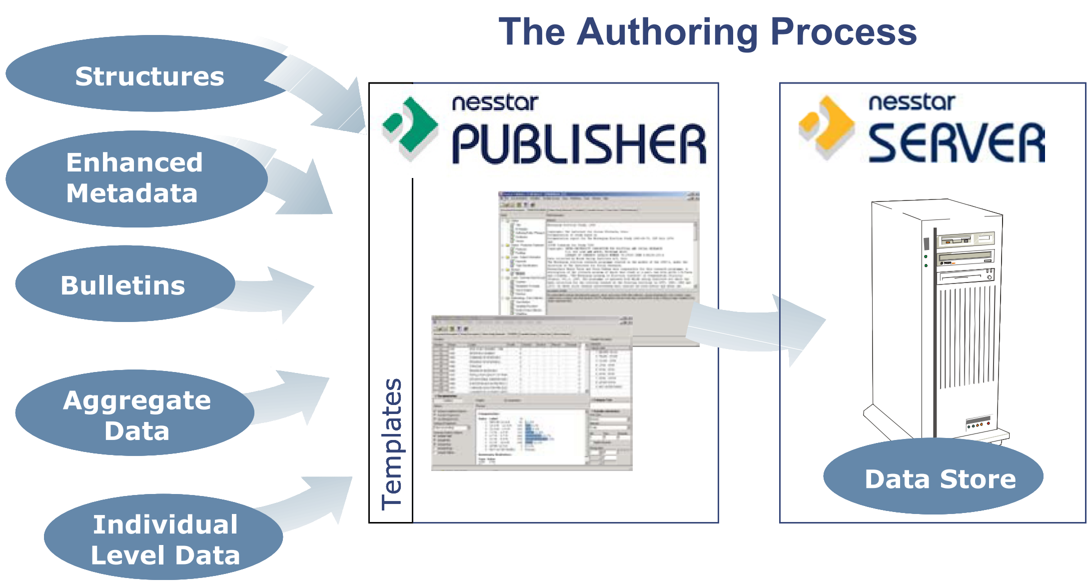
Semantic web

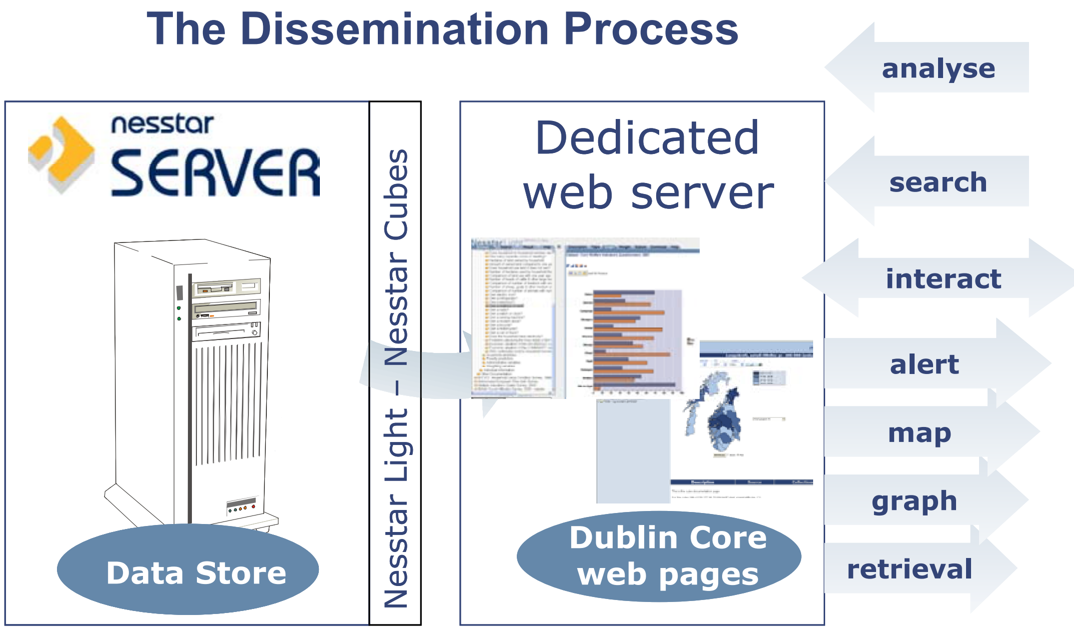
- Metadata – the glue of the Semantic Web
- A framework for “knowledge representation” – RDF
- The introduction of namespaces (allowing different system of terms and concepts to cohabitate in a single information system)
- Partial understanding/agreement
- The vision: the creation of a dynamic framework facilitating cooperation/interoperability across domains and communities - gradually expanding the “web of understanding”.

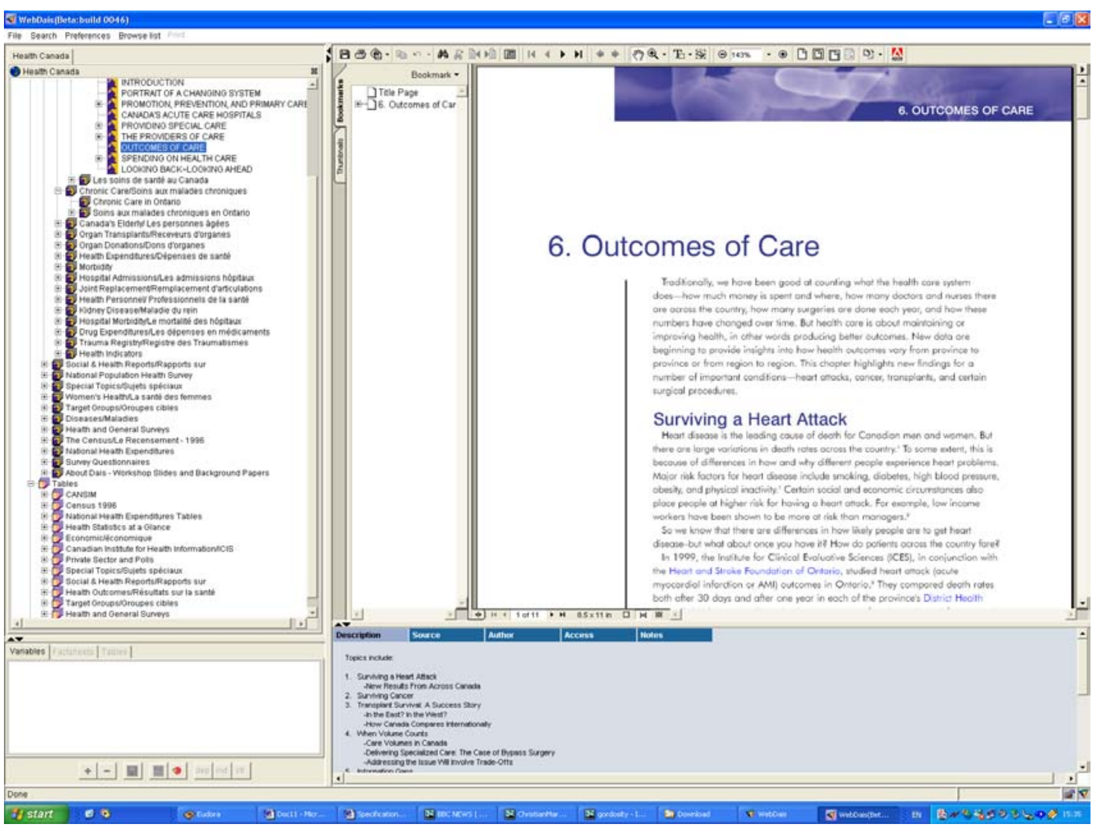


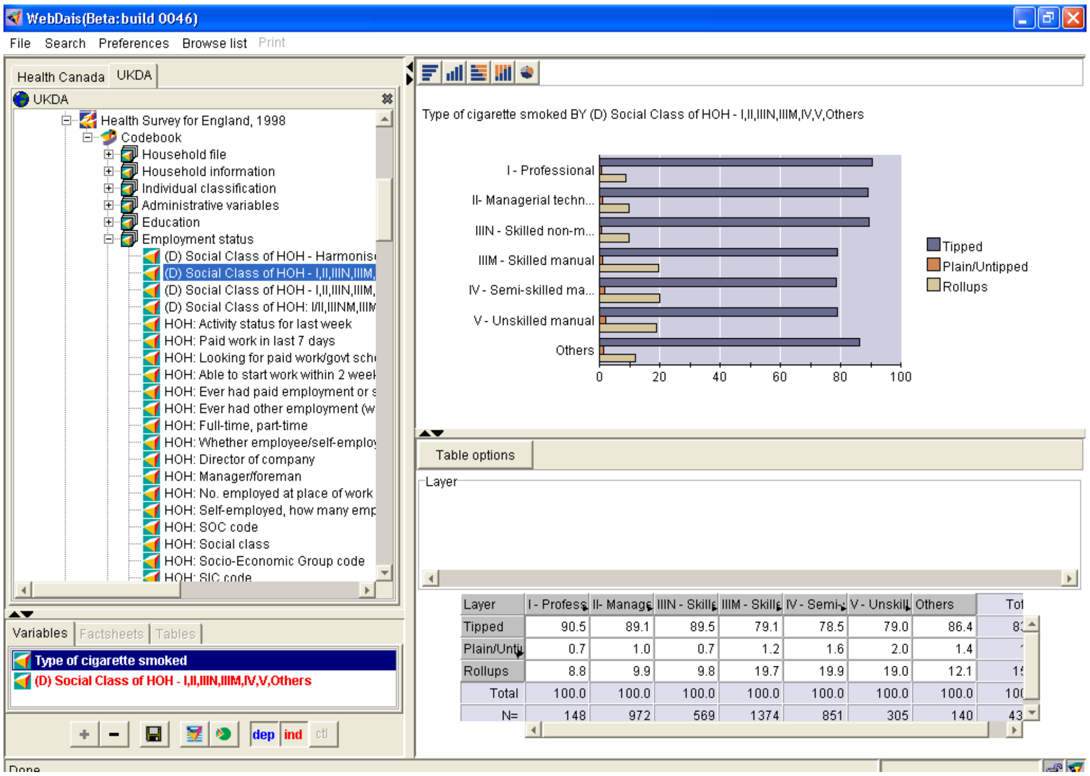
NEOOM – Nesstar Object Oriented Middleware

- All statistical objects "live" at a URL
- Objects are self describing – when a client access the URL of the object, the object returns a description of its current state (and its available methods) in RDF (using RDF as an Interface Description Language)
- Remote object-oriented calls are performed by a simple protocol running on top of HTTP. The calls can be stored as a URL, specifying the location of the relevant object as well as the method parameters.
- This allows for client side storage of statistical operations that can easily be rerun at a later stage thereby creating a simple batch language for operations on remote statistical objects.




The Dissemination Process
nesstar
SERVER
Data Store
Nesstar Light – Nesstar Cubes
Dedicated
web server
Dublin Core
web pages
analyse
search
interact
alert
map
graph
retrieval


6. OUTCOMES OF CARE
6. Outcomes of Care
Surviving a Heart Attack



The Nesstar perspective

•Building systems for the secondary user

•Context is all important

•Quality measurement is driven by provenance

•Focus is on integration with other types of content

–Link via Dublin Core (and e-GMS)

–Content management systems

•Classifications increasingly important

•Analysis – generating tables on the fly and viewing them in various ways (tables, graphs, maps)

•Ease of use (for different types of users) is key

Conclusions:

•Nesstar's mission is to build web based information systems that are easy for a new community of users to access

•Consequently focus is on

1. Long term management and descriptions of static data
2. Easy creation of new tables integrated in reports
3. Easy interaction with live tables

•Standards serve all three objectives in different ways**Helping People Understand Statistics Terms:**

The Statistical Interactive Glossary (SIG) and the GovStat Ontology

Stephanie W. Haas, Ron T. Brown, Jesse Wilbur, Cristina Pattuelli
School of Information and Library Science, University of North Carolina
{haas, browr, pattm}@ils.unc.edu, jdwilbur@email.unc.edu
Presented by Carol Hert

Consumer Price Index - All Urban Consumers

Series Id: CUUR0000SA0
Not Seasonally Adjusted
Area: U.S. city average
Item: All items
Base Period: 1982-84=100

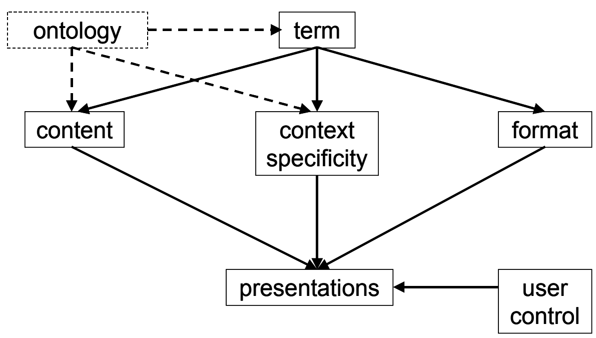
Year	Annual
1993	144.5
1994	148.2
1995	152.4
1996	156.9
1997	160.5
1998	163.0
1999	166.6
2000	172.2
2001	177.1
2002	179.9

The presentation stressed at the development of SIG and its content, especially focusing on the Consumer Price Index (CPI) case.

Let us consider the time series presented in the figure on the right of the CPI.

It is crucial to help users understand important statistical terms and concepts in the context in which they are used, by integrating glossary tools as seamlessly as possible into the statistical resources themselves.

Development of SIG

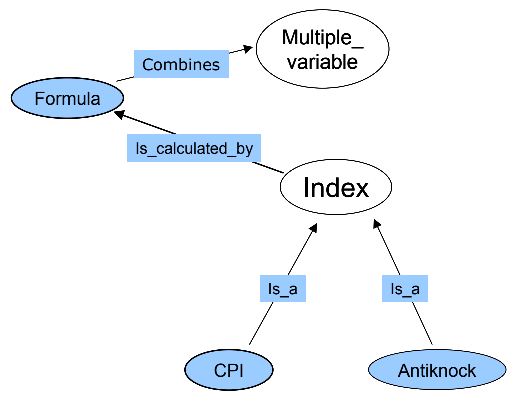


Terms:

- Terms that users frequently encounter on agency sites, not comprehensive dictionary
- Basic level of statistical literacy, not highly technical resource
- Strategies for term identification
  - examination of frequently-visited pages
  - anecdotal evidence from agency and non-agency consultants
  - metadata user study
  - webcrawl of agency sites

GovStat Ontology

- Hierarchical structure identifies related terms. Inheritance of taxonomic relationships supports context-specific presentations.
- Semantic relations among concepts suggest opportunities for combining related concepts into more comprehensive explanations.
- Explanations of a concept follow templates for definitions or examples established for parent concepts. Templates streamline the creation of additional presentations for other subclasses or for additional contexts.



Content

- Provide basic level of explanation
- Means of delivering information about concept
  - definition
  - example
  - brief tutorial
  - demonstration
  - interactive simulation
  - combination
- May incorporate related terms and concepts
- Give pointers to more complete and/or more technical explanations

Context specificity

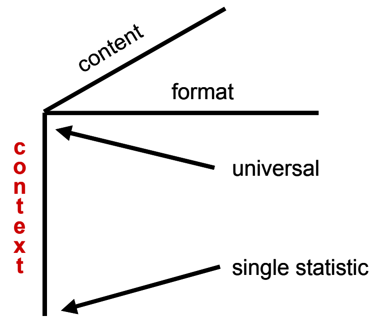
- Seek to incorporate explanations in users' work context
  - Table- or statistic-specific, based on a single row, column, or statistic, e.g., CPI, national death rate, gasoline prices in NY state, etc.
  - Agency or concept-specific, incorporating entities from agriculture, labor, science R&D, energy, etc.
  - General, context-free, "universal"
- Provide explanations of term or concept that are as relevant to user's current context as possible.
- When user invokes help on a term, the most specific explanations available are offered.

- If there is no explanation for that specific statistic or table, more general (e.g., agency-specific) ones are offered. Default is “universal” level.
- Path from specific to general is based on the ontology.

Format

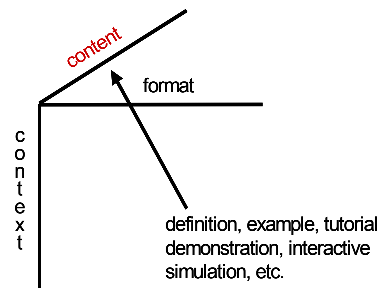
User can choose desired format of explanation based on interest, learning style, reading level, hardware/software limitations, etc.

- text
- text plus narration
- graphic
- graphic plus narration
- animation
- animation plus narration
- interactive
- and so on



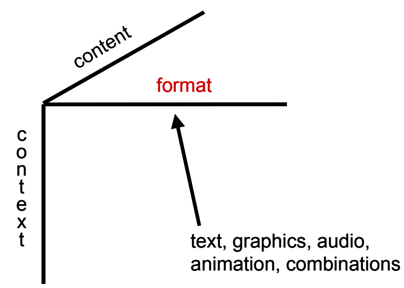
For each term, choices from the three dimensions are combined to create a presentation.

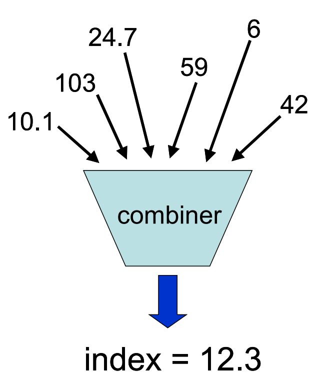
Context: a presentation may be universally applicable, or specific to a table or statistic.



Content: how to “house”, or deliver, the information.

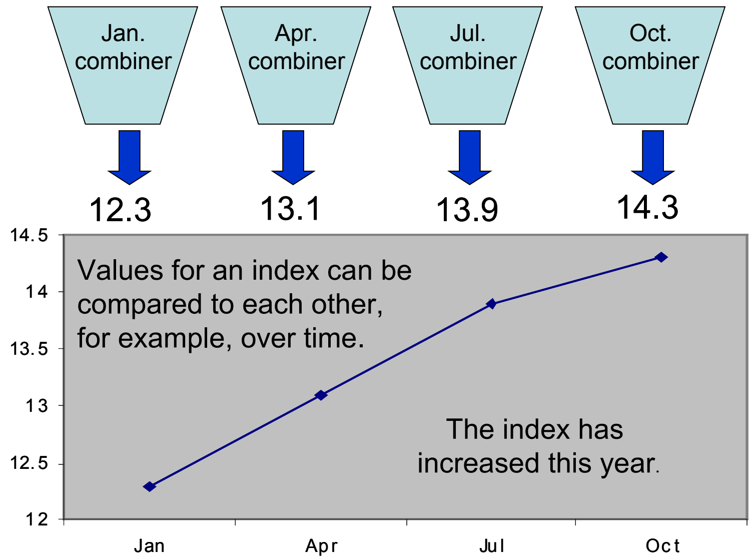
Format: effective formats may depend on hardware limitations, learning style, and content





An **index** combines numbers measuring different things into a single number. This number summarizes the measures. It can be compared with other values for the index over time.

An index combines numbers measuring different things into a single number. The single number represents all the different measures in a compact, easy-to-use form.



The *Consumer Price Index (CPI)* represents changes in prices of all goods. It combines prices into a single number that can be compared over time.

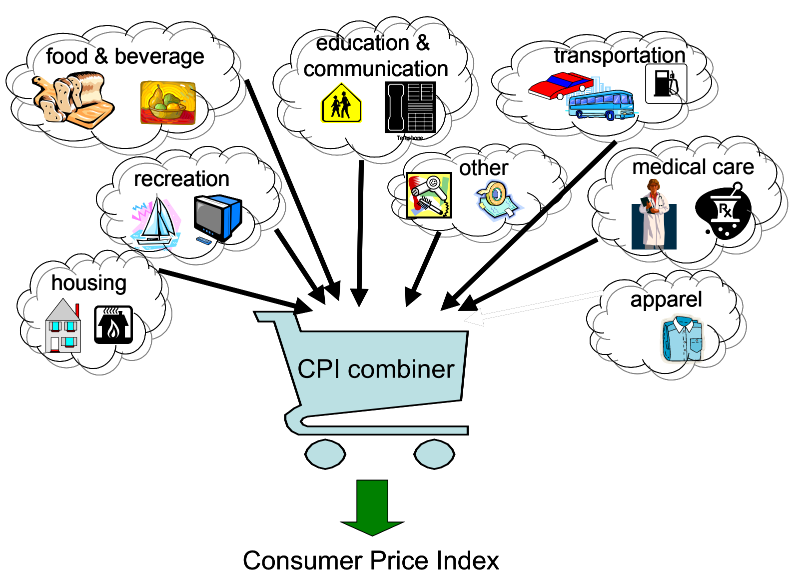
Items are classified into 8 major groups:

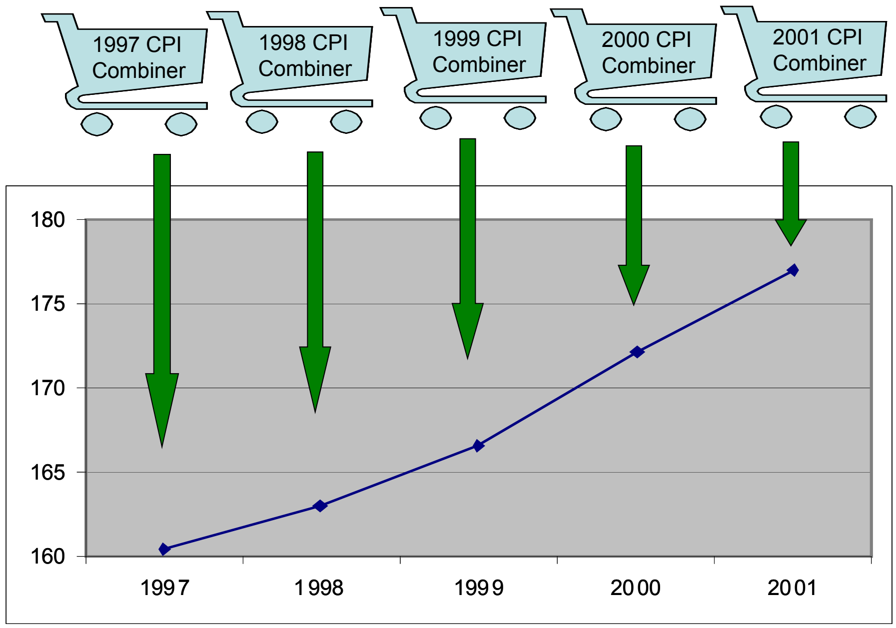
- Food and Beverages
- Housing
- Apparel
- Transportation
- Medical Care
- Recreation
- Education and Communication
- Other

Consumer Price Index - All Urban Consumers

Series Id: CUUR0000SA0
Not Seasonally Adjusted
Area: U.S. city average
Item: All items
Base Period: 1982-84=100

Year	Annual
1993	144.5
1994	148.2
1995	152.4
1996	156.9
1997	160.5
1998	163.0
1999	166.6
2000	172.2
2001	177.1
2002	179.9





The Consumer Price Index has increased since 1997.

Future Work

- Usability testing (summer 2003)
- User controls (summer 2003)
- Increase coverage of statistical vocabulary (ongoing)
- Apply concepts to domain vocabulary, e.g.,
    - full-time, part-time
    - disability
    - region

FP6 AND FUTURE PLANS

eNIPS - eNetwork for Improving Public Statistics

John Charlton, Office for National Statistics, UK

John.Charlton@ons.gsi.gov.uk

This presentation consists of the research activities of the e-NIPS proposal submitted in the FP6.

Strategic Objectives of e-NIPS:

- Bring high quality statistical information to the fingertips of the European citizen, businesses and governments.
- Improve the impact of ICT's on the quality of public statistical knowledge and the community.
- Create and sustain a multidisciplinary network of excellence supporting research in technologies for public statistical processing, and disseminate the resulting information - a virtual research institute for public statistics.

eNIPS will achieve its objectives through...

1. Developing ICT's supporting organisational networking, process integration, and the sharing of resources.
2. Integrating visionary European and international research communities to build up new knowledge;
3. Bringing together all stages and aspects involved in statistical information production and dissemination processes.

The eNIPS Virtual Research Institute will:

1. Develop a strategic, collaborative, interdisciplinary vision of future requirements, leading to shared understanding and a joint programme of activities;
2. Disseminate technological developments arising from joint research;
3. Develop and maintain open technological standards leading to high quality pan-European information;
4. Develop theory with resulting algorithms implemented in integrated open-source solutions

5. Build a knowledge management system, and promote of a KM approach to statistical production and dissemination;
6. Bridge gaps between types of institution (NSIs, universities, SMEs), and specialists - computer scientists and statisticians;
7. Encourage techniques for on-line dissemination that safeguard the privacy of individuals and businesses;
8. Bring together experts to develop inter-operable state-of-the-art e-Government solutions to existing and evolving needs.

Knowledge management objectives

i) Understand, prioritise and continuously revise the knowledge needs of the network;

ii) Establish clear contribution guidelines;

iii) Increase the breadth, depth, currency and coherence of the research knowledge base;

iv) Provide ready access to up-to-date information about the capabilities and activities of the network;

v) Disseminate knowledge gaps and expertise;

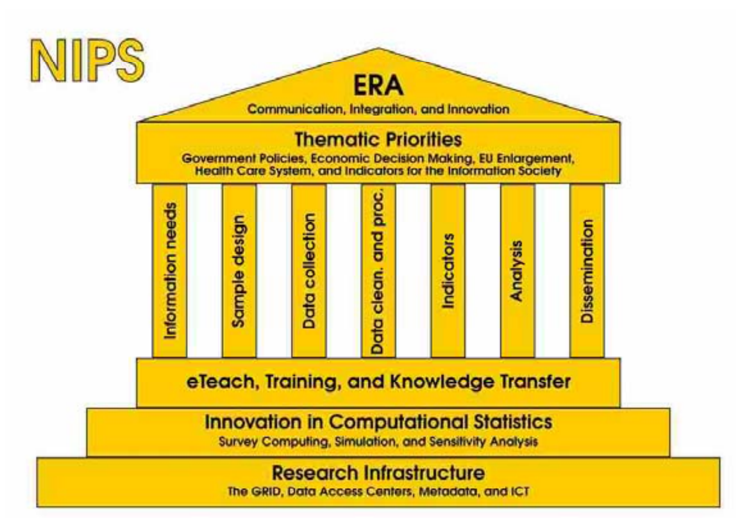
vi) Reduce duplication and re-invention;

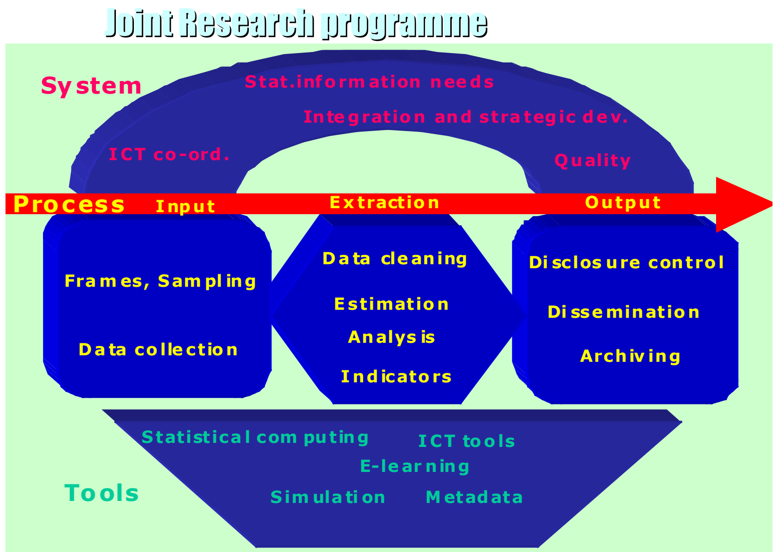
vii) Increase sharing and re-use of existing knowledge;

viii) Facilitate collaboration and foster innovation;

ix) Enhance learning and feedback.

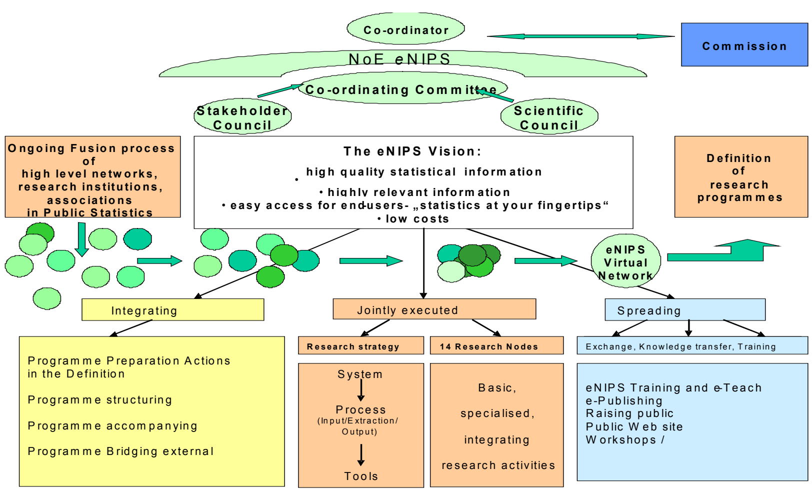
Integration through eNIPS





Networks Incorporated into eNIPS

- METANET - metadata
- CODACMOS - data capture
- AMRADS - fostering take-up of CBM
- European Forecasting Network
- CASC - disclosure control
- FP5 projects

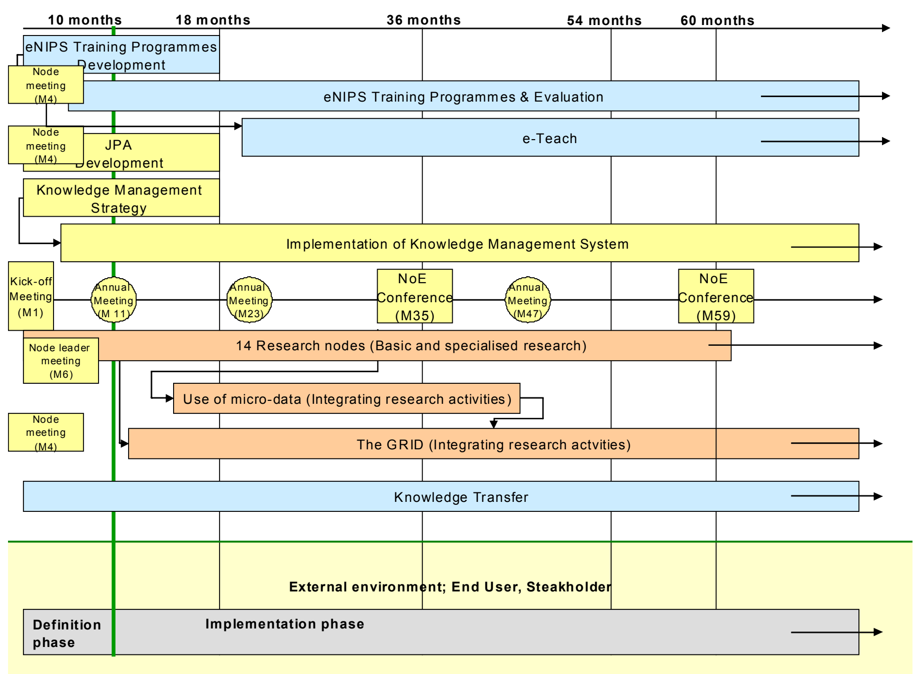


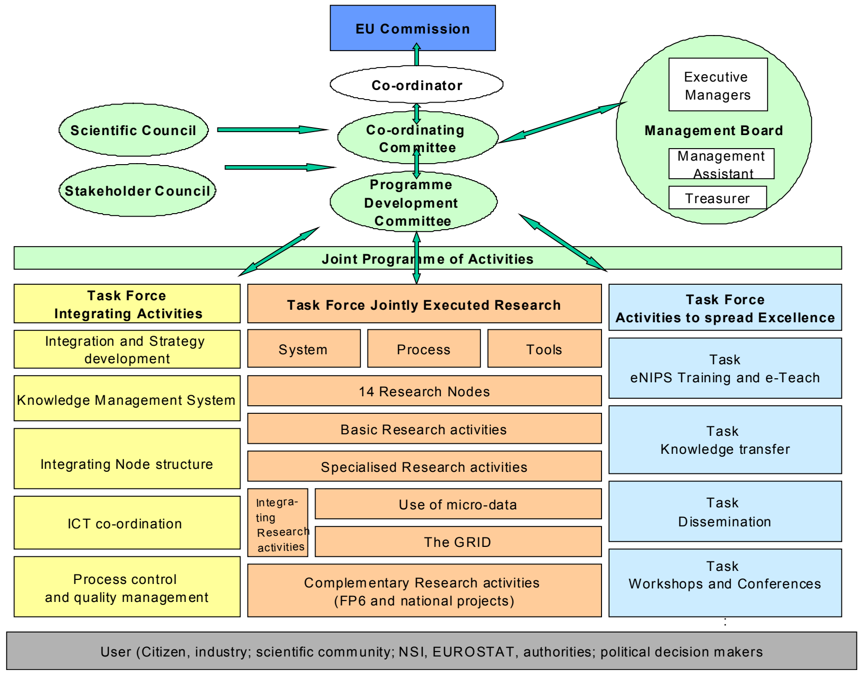
E-NIPS overview

Jointly executed research activities (nodes)

0. Integration and strategic development
1. Statistical information needs
2. Economic and social indicators and forecasts
3. Survey frames, sampling, and estimation
4. Analysis
5. Data collection
6. Data cleaning and processing
7. Quality measurement and benchmarking of NSIs
8. Dissemination and archiving
9. Metadata
10. Statistical disclosure control
11. Simulation and statistical computing
12. Information and communication technologies for public statistics
13. Training and eLearning

Roadmap





eNIPS will provide a common integrating framework for research, including...

- A common interactive R&D website,including an e-library
- A web-based helpdesk/ mail-base
- A glossary of statistical terms
- Training including e-teaching (also for those outside the network)
- A software component library (open-source)
- Support for bidding for R&D funds
- Annual European Public Statistics Methodology Conference (c.f. JSM in USA - Europe lacks such meetings)
- Workshops on important topics
- A Project Office supporting the network administratively

The e-NIPS nutshell:

Build an integrated Network of Excellence for improving Public Statistics, with an appropriate common integrating framework

Develop a joint programme of activities

Participate jointly in bids for agreed R&D activities in FP6 (e.g. STRPs)

http://www.publicstatistics.net

The following questionnaire was distributed to the participants of the Conference for completion.

METANET: Evaluation Questionnaire

Please complete the following sections, as appropriate:

1. The activities of the Metanet Network of Excellence (Partners, Members & Associates)
2. Metanet: What Next? (All, as applicable)
3. The Metanet Final Conference (All)

Please indicate your Metanet membership status:	Partner	Member	Associate	Other
	☐	☐	☐	☐

Section 1: Evaluation of the MetaNet Network of Excellence
What did you hope to achieve through your involvement with the Metanet Network?
What have you gained through Network participation? Can you provide an example of how you have benefited or your work has been influenced?
Which aspect of Network participation has been the most disappointing?

Please describe ways in which you believe that the Network has already contributed or has the potential to contribute to any of the following:	
Product or Process Innovations	New Services or Methods
Technical Standards	EU/International Regulations

To what extent have the following criteria been met?

Please rank where: **5= Strongly Agree 3= Agree 1= Strongly Disagree**

	5	4	3	2	1
The Network has met my expectations.	☐	☐	☐	☐	☐
The Network has influenced the languages, tools and methodologies used to describe statistical metadata.	☐	☐	☐	☐	☐
The Network website(s) is(are) informative on the activities of the Network.	☐	☐	☐	☐	☐
The Network website(s) is(are) informative on the activities of other relevant bodies.	☐	☐	☐	☐	☐
The Network serves as a reliable source of workable solutions and best practice.	☐	☐	☐	☐	☐
Information retrieved is easy to apply in my work.	☐	☐	☐	☐	☐

Section 2: Metanet - What Next?	Yes	No	N/A
Overall, has Network participation proved worthwhile?	☐	☐	☐
Would you like to see the Network continue in some form?	☐	☐	☐
If continued, is the Network purpose still valid?	☐	☐	☐
If continued, would you recommend the Network to colleagues?	☐	☐	☐
Please provide any additional comments or suggestions regarding the future of the Network			

Section 3: Evaluation of the Metanet Final Conference
What did you hope to achieve by attending the Final Conference?

Please rate the following Conference components. Where **5=Excellent 4=Good 3= Acceptable 2=Needs some improvement 1=Poor**					
	5	4	3	2	1
Registration process	☐	☐	☐	☐	☐
Information available through the Conference website	☐	☐	☐	☐	☐
Overall Conference programme content	☐	☐	☐	☐	☐
Variety of topics and discussions	☐	☐	☐	☐	☐
Quality of presentations	☐	☐	☐	☐	☐
Insight gained into the presented topics	☐	☐	☐	☐	☐
Usefulness of the knowledge gained for everyday work	☐	☐	☐	☐	☐
In what ways did the Conference meet or fail to meet your expectations?					

Thank you for completing this questionnaire. Your feedback is much appreciated.