STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

COMMISSION OF THE
EUROPEAN COMMUNITIES

EUROSTAT

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

Working Paper No. 13
English only

Topic (i): Infrastructure issues for statistical metadata

METANET: A NETWORK OF EXCELLENCE FOR HARMONIZING AND SYNTHESIZING THE DEVELOPMENT OF STATISTICAL METADATA – PROGRESS REPORT

Submitted by The MetaNet project[1]

Contributed paper

Summary

This paper describes the MetaNet network of excellence, a network funded by the European Union fifth research and development programme. It outlines the background to the network, and describes the way it works. It reports of the achievements to date, specifically with the workshop held in Vienna in October 2001. Finally, it describes the next phase of the project, and identifies how those who are interested can contribute to the work.

I. INTRODUCTION

1. Discussions at the fifth EEA Working Party Meeting in November 1999 made it clear that there is a shared need to harmonize metadata, or more fully, to harmonize the methodology, definitions and models used to describe statistical processing systems. In addition to the funded European Union Framework 5 projects, there is a wealth of expertise in the existing or recent DOSIS (Development of Statistical Information Systems) projects as well as in NSIs (National Statistical Institutes) and independent experts. It was clear from the DOSIS99 meeting discussions that there is a common RTD objective in harmonizing metadata, or more fully, in harmonizing the methodology, definitions and models used to describe statistical processing systems.

2. MetaNet grew out of this realization that there had been significant progress within these projects, and elsewhere, but that co-ordination between the research projects, and with this community and other activities had been lacking. This was a failing recognized by the European Commission with regard to many Fourth Framework projects, and so, within the Fifth Framework it created a number of ways to encourage sharing of experiences between projects. Among the terms used were ‘clusters’, ‘thematic networks’, ‘accompanying measures’ and ‘networks of excellence’. These different types of projects had different conditions attached, and MetaNet was submitted as a ‘network of excellence’ because this was the scheme that would allow the broadest range of participants. However, this particular scheme is quite restrictive on the types of budget headings that it allows. It is important to explain this organizational background in order to understand the strengths and weaknesses of MetaNet: what it can deliver, and what its limitations are.

3. In addition to the R&D focus, there was, and is, expertise within national statistical organizations, international organizations such as UNECE (Geneva), Eurostat (CES), OMG (Object Management Group) and in the national Data Archives. The aim of the network is to bring together experts and users from NSIs, users of official statistics, researchers and developers. It is intended to consolidate the work on

[1] Prepared by Joanne Lamb.

metadata models that has been carried out in NSIs and in Fourth Framework and Eurostat Supcom projects. The network has four main objectives:

- to develop proposals for standards in the methodology used for describing statistical metadata and statistical information systems;
- to develop proposals for recommendations on the metadata objects in a common conceptual model of statistical metadata;
- to disseminate these proposed standards to the relevant user communities and standards bodies;
- 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.

II. THE WORK PLAN

4. In order to reach these objectives, the strategy of the plan for MetaNet was as follows:

- Bring together as wide a constituency of experts and users as possible;
- Supply a modest amount of funding to allow them to communicate and share ideas;
- Produce four deliverables which would inform users and developers of the common thinking and help in implementation;
- Supply a website which permits exchange of information and discussion.

5. More specifically, four Working Groups were identified along with a road map for progress over the 30 months of funding. The Established Working Groups are:

- WG1: Working Group to establish a methodology and tools for communication:
  XML, UML etc
  SELECT, Rational Rose, XML tools etc
  Metadata schemas, RDF, Dublin Core
- WG2: Working Group to establish and unify current practice
  classifying the different types of metadata
  glossary, definitions and relationships
  metadata systems currently in use
  unification of terms and concepts
- WG3: Recommendations:
  Working Group to establish best practices for adopting the outcome of WG1 and WG2
- WG4: Exploitation:
  Working Group to recommend how WG3 results can be implemented in practice.

6. The Working Groups consist of members of the consortium, and volunteers (paid or unpaid) who contribute to the development of the final deliverables. The members were recruited at the Kick-off conference held in April 2001. Since then we have had one joint meeting between Working Groups 1 and 2, and a meeting of all four groups is planned for March 2002 (just after the Metis meeting).

III. PROGRESS TO DATE

7. The Kick-off conference was very successful, with nearly seventy participants coming from a wide variety of backgrounds. These included NSIs, academic and research institutions, the software industry, data archives, banks and health organizations. The geographical distribution covered the European Union, the USA, Canada, Israel, Poland, Slovakia, Switzerland, Norway, Slovenia, and international organizations. This number was more than had been expected (the anticipated number was fifty) and we had more than the anticipated number expressing interest in becoming members; in fact we needed to develop a system where organizations could participate without payment. In addition we agreed to create a semi-public website to receive contributions.

8. Since the Kick-off conference, Working Groups 1 and 2 have been working on their tasks. It was decided that there was some overlap between Working Groups 1 and 2, and it was agreed that Working Group 1 would concentrate on collecting information, taking a broad-brush approach, and Working Group 2 on analysing it, and developing ideas further. Consequently, Working Group 1 developed the outline of a report, due to be available in draft form in March 2002. The outline of the report is given below.

- Section 1: Introduction
- Section 2: Dimensions of statistical metadata
  2.1. Life cycle
  2.2. Usage type
  2.3. Usage level
  2.4. User functionality
  2.5. Software functionality
- Section 3: Tools
  3.1. Levels of abstraction
  3.2. Modelling languages
  3.3. Communication and storage languages
  3.4. Software supporting modelling and communication
- Section 4: Metadata models
  4.1. Possible typologies
  4.2. Description of existing models.

9. Working Group 2 has a duration of 12 months, with an expected delivery date for the report being October 2002. The activities of this Working Group comprise a sequence of 9 phases as listed in Table 1 below.

Table 1. WG2 Activity overview

Phase	Activity	Begin–End (approx.)
1	Collection of metadata material (standards, models in use) together with WG 1	May – October 2001
2	Development of a reference metadata structure	June – October 2001
3	Working Group meeting: Final version for the conceptual model of the reference metadata structure	October 2001
4	Mapping of metadata models into the reference platform	August – December 2001
5	Detailed elaboration of the reference metadata model	October – December 2001
6	Detailed description of mapping procedures	January – April 2002
7	Description standards for metadata models	March – June 2002
8	Development of a terminology database (list of synonyms and related terms)	March – September 2002
9	Deliverable of WG2	April – October 2002

10. The Working Group meeting in Vienna in October 2001 produced a first conceptual model for describing and classifying statistical metadata. Metadata can be described according to a number of dimensions, and the two Working Groups identified similar dimensions. Since the Vienna meeting, the two sets of definitions have been harmonized.

11. During the meeting presentations were made of thirty-four models, tools, standards, and statistical software systems. The subjects of these presentations have been classified according to their type and status (proposal, commercially available etc). The meeting discussed these presentations, and concluded with a review of the products, standards and models. The range of the presentation material was wide, in order to identify as many aspects of statistical metadata as possible.

12. Eight standards were reviewed: METIS, ISO 11179, XML, Dublin Core, DDI, GDDS/SDDS, OECD, GESMES/CB. Of these XML and DDI were considered to have a model, and XML was also a tool. Four other informal or 'near standards' were also identified: the ECE recommendations, the Infological model (αβγτ), PC-Axis and Triple S. The αβγτ is also a model, and PC-Axis and Triple S also have tools associated with them.

13. Fourteen other models were considered: CWM, relational model, spreadsheets, UML, FASTER, CRISTAL, ComeIn, MIMAMED/MAMEOB, Banca d'Italia, CMR, IDARESA, Blaise, terminology model and UNIDIS. Of these spreadsheets, UML, FASTER, CRISTAL and ComeIn were identified as having tools attached, and spreadsheets were also described as a system. Additionally, Banca d'Italia, CMR, IDARESA, Blaise, terminology model and UNIDIS were identified as systems.

14. Three other tools were identified: CODED, TADEQ, OLAP, and four other systems: ISMIS, Beyond 20/20, StatLine/StatBase and "system files".

15. The items that have been identified have been discussed quite fully, in order to demonstrate the range of models and standards under consideration. We shall not reproduce the results of the session reviews here, with the exception of the fifth session, which was concerned with usage. The session focused on the one hand on a discussion of metadata usage aspects, highlighting the complexity of the issue and identifying a set of dimensions useful in structuring the metadata domain. We also saw a practical demonstration of software prototype (based on UNIDO's industrial statistics database) that gave rise to lively discussion on the merits of a formal management of (annotative) metadata, and suggested some ways of extending the framework towards process integration.

16. One of the most useful outcomes of the meeting was the presentation by Working Group 2 of a conceptual framework for describing statistical metadata, identifying five canonical dimensions for describing statistical metadata. Figure 1 illustrates four of these five dimensions: structure, stage, role, form (intentional and extensional), the fifth dimension being function.

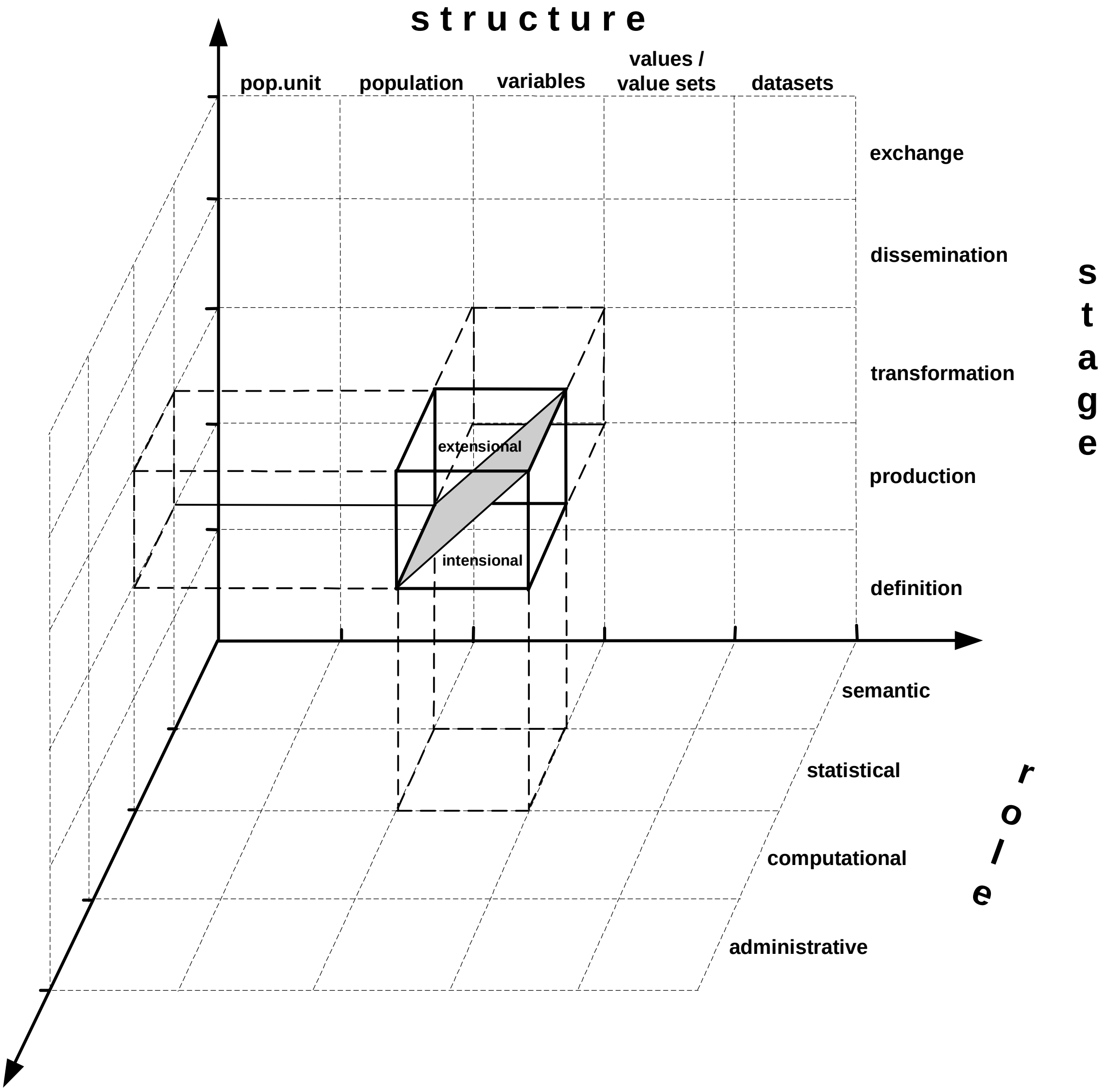

Figure 1

17. The group then mapped the existing models and standards against this framework, and discovered that the space identified by these five dimensions was not uniformly covered. For example, the DDI model was rather strong in the semantic and statistical aspects of datasets, but rather weak in administrative terms. In contrast, the CMR model covers the relevant administrative aspects in great detail. With respect to processing, a number of metadata models cover the requirements of specific processes in detail (for example GESMES for exchange, the model of Banca d'Italia for modification, or TADEQ for production) but no single model embraces all types of processing.

18. Consequently, it was decided that the identified gaps in the mapping should be covered by identifying five tracks, and requesting members of the Working group to address these issues. The five tracks are:

- Track 1: Development of metadata repositories in intentional (machine processable) form.
  This track concentrates mainly on the *structure* and *role* dimensions of the framework and its aim is to develop a reliable terminology model.
- Track 2: Comparison of existing "small area" models
  For a number of important areas there exist competing models. These should be evaluated in order to obtain a clear picture about both strong and weak points of the

approaches. As a result of the discussions of the meeting, the following areas can be identified:

a) Data capture
b) Retrieval of data sets (e.g. tools such as PC-Axis, Beyond 20/20)
c) Retrieval of documentation about data sets

- Track 3: Analysis of existing process models inside statistical offices
  The main task of this track is identification of the requirements of process models from the perspective of national statistical offices.
- Track 4: Development of a structural metadata model
  The goal of this development is a metadata model useful for *active* use of metadata in statistical processing. The model should, on the one hand, be more specific than the ECE guidelines for metadata modeling, on the other hand sufficiently flexible for adaptation to different production environments. Starting point will be the following two task forces:
  a) Comparison and unification of the models of Banca d'Italia and IDARESA/ISMIS
  b) Comparison and unification of CMR and DDI
- Track 5: Analysis of Usage Scenarios
  Different users of statistical data have views of metadata. In particular the production viewpoint and the dissemination viewpoint can be very different. This track will examine these two views and identify the different usages of metadata, and the type of metadata captured and used.

IV. FUTURE PLANS

19. We have described the work of the MetaNet network to date, focusing on the outcomes of the Vienna workshop, and in particular on Working Groups 1 and 2. It is planned to have a draft of the Working Group 1 report and reports from the five tracks of Working Group 2 by the next Project meeting in March 2002. Shortly after that, the Working Group 1 report will be finalized and published, while Working Group 2 will continue according to the following timetable.

- Detailed description of mapping procedures: January – April 2002
- Description standards for metadata models: March – June 2002
- Development of a terminology database (list of synonyms and related terms): March – September 2002
- Deliverable of WG2: October 2002

20. The meeting in March will be attended by members of all four Working Groups. It will review progress to date, and the work of Working Group 3, i.e. recommendations on best practices for adapting to recommended metadata models and standards, will begin. This group will have a joint meeting with Working Group 4 in the autumn of 2002. The remit of Working Group 4 is to address the practical issues faced in moving to the recommended best practices, and will develop a training manual. A final conference will be held in the spring of 2003.

V. PARTICIPATION AND ACCESS

21. MetaNet is an open access network. There are a number of different levels at which it is possible to participate, and some of these are still open to those who are interested.

- Level 1: the project is coordinated by the University of Edinburgh, who are responsible for the administration, interaction with the European Commission (via Eurostat) and the maintenance of the websites.
- Level 2: there are six partners, four of whom lead Working Groups, and two have responsibility for the proceedings of the conferences. These partners are:

  WG2: Statistics Netherlands
  WG2: University of Vienna
  WG3: Statistics Sweden
  WG4: Statistics Norway

Kick-off Conference: Survey & Statistical Computing
Final Conference: University of Athens.

- Level 3: Members and associates were recruited, mainly from the Kick-off Conference. The status is the same, except that members receive some money from the project, and associates are self-financing. Members and associates participate in the project meetings, and contribute to the reports of the Working Groups. The numbers in the Working Groups has been kept reasonably small, but there is room for a few extra people, if they are interested.
- Level 4: We maintain an 'interested parties' web site and a mailing list. Anybody interested can subscribe himself or herself to the mailing list, and receive emails about the project. In addition, the password to the website may be requested, giving access to most of the work in progress. Further, he can upload documents of his own, and take part in a discussion forum. The reason for personal contact before granting access to the web site is because we give the facility to submit information as well as to read it, and therefore we wish to maintain some control over who has access.
- Level 5: The public website holds the published reports of the network, and also a resource section giving access to a terminology database, publications and other reference material.

22. Further details can be found on the website at http://www.epros.ed.ac.uk/metanet