WP. 14
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Joint UNECE/Eurostat/OECD work session on statistical metadata (METIS)
(Luxembourg, 9-11 April 2008)

Topic 2 (iii) Metadata and the statistical cycle and Implementation

METADATA MANAGEMENT AT THE FSO - INTRODUCING AN END-TO-END APPROACH TO SUSTAIN MODERNIZATION PROJECTS

Submitted by Switzerland[1]

I. INTRODUCTION

1. Metadata management at the FSO has for a long time been a task that was under the direct supervision of decentralized production units. As a consequence, many subject matter specialists have developed various solutions to cover the short term needs of their respective units. While some of those metadata management systems are quite advanced in the features and content they provide to FSO statisticians, none of them offer the basis for an exhaustive centralized system that would contain metadata designed for internal specialists as well as external users (metadata for publication). The same statement applies to metadata processing workflows, which even as of 2008 are still handled differently in the many statistical activities that produce data in the FSO.

2. The introduction in 2005 of a central metadata system and a central data warehouse marked the first substantial step to structure and increase the coherence of statistical data and metadata used for data analysis.

3. The FSO is currently reengineering data production workflows in line with the concept of integrated statistical production systems. The paper describes how modernization projects will affect the production of metadata and make it necessary to adopt an end-to-end approach to cover the extensive metadata needs of integrated production systems in the whole statistical cycle. Organizational measures are presented in detail and motivations behind the development of a new metadata system are outlined as well.

II. METADATA CENTRALIZATION AT THE FSO: THE CODAM PROJECT

4. Centralization in a Federal State is considered at best as a necessary evil and is meant to remain some sort of exception restricted to selected aspects of the affairs of the State. The same remark could apply to metadata centralization at the FSO. Centralization in metadata and data management was originally conceived

[1] Prepared by Fabien Perrot (fabien.perrot@bfs.admin.ch).

in the late nineties in what would become the CODAM[2] project. A conservative approach to centralization was adopted in the sense that centralization was restricted to the analysis and distribution phases of FSO data. Previous steps of data processing were largely left out of the scope of the project.

5. The main metadata assets brought by the project are a central metadata system[3] used as catalogue of data collections stored in the central data warehouse and a classification server (shown in yellow in Illustration 1).

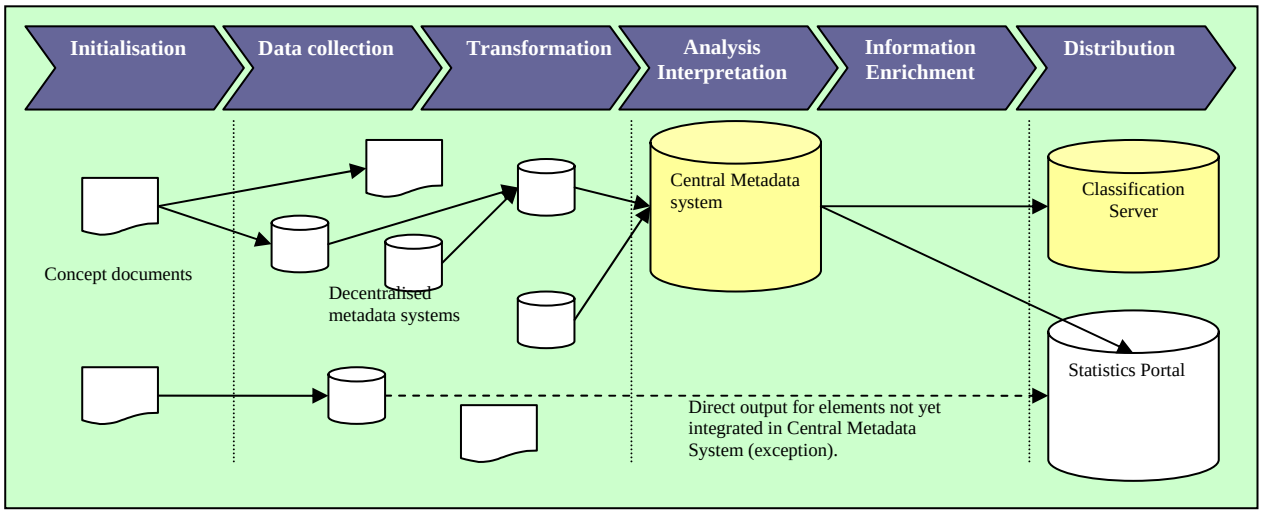

Illustration 1 - CODAM Project - Metadata workflow and systems

6. Key lessons learned from the CODAM project can be summarized as follows:

a. A unique, cross-domain metadata model can be used to capture most of the descriptive elements related to statistical activities, data collections, variables, value domains and classifications[4].
b. Provision has to be made for additional, content specific attributes in the model which can be made available to users. This will ensure that statisticians can input most of the information they need in the central system and not have to rely on information stored in external documents. These attributes may typically not be cross-domain elements.
c. A central metadata system should contain information meant for internal users (not meant for dissemination) and external users (specific for publication). Metadata may have different target audiences and this distinction needs to be reflected in metadata modeling.
d. Metadata centralization and metadata harmonization are two different concepts even though they were seen as overlapping concepts at first. Thus, centralization can be done with no or very little harmonization, although this approach is clearly not recommended in order to avoid redundancy in metadata content. Harmonization, on the other hand, cannot be achieved without centralization.
e. Redundant metadata capture can only be avoided if the central system entirely replaces previous decentralized systems. This means that an end-to-end approach to metadata management ought to be favoured in the conception of a central metadata system.
f. The central metadata system has to be able to interface to most data production systems to maximize its usefulness.
g. Older metadata present in decentralized systems may never be migrated to a new system if the cost of migration cannot be justified. This is problematic if older systems can no longer be maintained. Metadata may then be lost or stored in a way that makes their retrieval difficult.

[2] CODAM is the abbreviation of “Corporate Data Management”.
[3] The system has been in production since 2003 for classifications and was extended in 2005 to contain statistical activities, contextual variable definitions and value domains.
[4] Statistical metadata terms used in this paper are extracted from the Neuchâtel Terminology (variables and classifications).

7. Work introduced as part of the CODAM project proved that centralization is a viable option for metadata at the FSO, but the vision was somehow too restrictive by not providing an end-to-end strategy for metadata management. As a consequence:

a. Double capture of metadata could not be entirely avoided. Statisticians were obliged to keep part of their metadata in two different systems which were not interconnected.
b. Harmonization of codifications, variable names, etc. only occurred in the central system and was not always reported back in other systems, producing in effect two largely similar and overlapping metadata sets instead of one.
c. The central metadata system was meant to be the source of official metadata at the FSO but was until recently considered by statisticians to be a target system. In practice, up to date metadata were often to be found in decentralized systems and were only later, once content was frozen, migrated to the central system.

8. The central metadata system as part of the CODAM project can be said to have partially fulfilled its goals by enabling central storing of metadata related to data collections used in the central data warehouse. However, due to latencies in the metadata preparation workflows, content is brought very late in the system and statisticians cannot always refer to metadata for their own work in a timely manner.

9. Further optimizations are needed to ensure the success of metadata centralization at the FSO, especially when taking into account the needs of ongoing modernization projects.

III. IMPACT OF FSO MODERNIZATION PROJECTS ON METADATA

10. The FSO is currently reengineering data production workflows with the definition of integrated statistical production systems. Those modernization projects[5] will bring a change in the way data are collected and processed at the FSO. It will mark the transition from a business model where each unit collects and processes its own data (stove pipes) to a model where data are collected once and reused in different statistical products. The use of administrative sources will also be favoured in order to reduce the burden imposed on respondents and increase the overall quality of source data.

11. Integrated statistical production systems typically focus on the observation of a particular statistical unit type[6] or are thematically related in the sense that they tend to reuse the same conceptual variables[7]. New production systems do not necessarily imply that statistical activities are merged and their output statistical products dramatically changed. They introduce however new methodologies and work processes, and facilitate harmonization of specific subject matter.

12. The introduction of modernization projects also bring the need to centralize even more metadata types and make those accessible to statisticians involved in statistical activities regrouped in an integrated statistical production system. This trend towards further centralization can be explained by the following factors:

a. Statisticians who collect data are potentially no longer those who will analyze or comment on data in statistical outputs. Transmission of knowledge using metadata is vital to ensure that data are properly handled in the whole statistical cycle. Emphasis has to be made for instance on indicators about quality of data after each processing step.
b. Reuse of existing material means that fewer data items are collected, but particular attention must be paid to methodological aspects to ensure that results can be used in several contexts. A questionnaire

[5] A global description of modernization projects, which regroup integrated statistical production systems as well as other projects, is available on the FSO internet page http://www.bfs.admin.ch/bfs/portal/en/index/news/00.html .
[6] Categorization of observed objects like persons and households, or enterprises and establishments.
[7] Like employment or income, i.e. variables which are not meant to be specific to a context as given by the statistical activity.

might contain data that will be processed separately in the context of several statistical activities. Also, the frequency of data capture will differ from that of data publication and the need may arise to merge data collected in different reference periods.

c. Statistical activities may need to rely on more than one data capture method. Data may come from registers or questionnaires, or may be reused from other data collections. Those different data paths have to be documented, and the complexity of data supplier management increases as well.
d. Process definition and monitoring are needed to ensure that each actor in the statistical cycle is aware of work to be done and can plan its own activities accordingly.

13. Although integrated statistical production systems do not exclusively focus on metadata, their introduction gives additional emphasis to the importance of having up to date metadata available centrally. This will bring more incentive to compile metadata and will demonstrate that metadata are not only designed to help the external end-user of statistics, but are relevant to producers as well.

14. In this context, the central metadata system is assigned a new task. If metadata in the CODAM project are essentially static and descriptive[8], integrated statistical production systems introduce the necessity to have process documentation within those metadata. Consequently, new or expanded metadata types to be added as part of the central system include:

a. Process definitions which indicate how data collections transit in the statistical cycle (from data collection phase to distribution);
b. Variable derivation to indicate transformation at the level of each contextual variable;
c. Populations and sampling techniques;
d. Questionnaires and questions;
e. Thesaurus to assist in the harmonization of term definitions;
f. Event logs to indicate history changes and to help with time shift of data collections mapped to classifications (mostly geographic classifications);
g. Standard quality indicators at the level of data collections and contextual variables.

15. Equally important is the need to redefine metadata capture workflows and provide assistance to metadata producers in the context of an organizational framework which links metadata stakeholders and allows them to reduce the burden of metadata creation and maintenance by sharing work.

IV. DEFINITION OF AN ORGANIZATIONAL FRAMEWORK FOR METADATA

16. Due to lack of centralization in metadata creation and management at the FSO so far, there can be no absolute certainty that metadata items only have one maintenance unit. This stems from the fact that the same metadata have to be present in several decentralized and unconnected metadata systems. In some rare instances, the need to customize standard metadata (classifications and value domains) has also lead to the creation of officially endorsed derivatives, which are not maintained by the original maintenance unit. Possible problems in coherence are detected when metadata are aggregated in the central system from different sources.

17. In order to better support metadata producers and deal with issues related to harmonization, an organizational concept has been developed to identify metadata stakeholders in the FSO and define overall workflows in the creation and validation of metadata[9]. The organizational system that is promoted takes into consideration that most metadata elements produced in the FSO are created in decentralized units. These efforts however need to be coordinated by a unit in charge of the central metadata system and statistical standards.

[8] Although they are also used as arguments in (semi-) automated processes, such as data load, validation and transformation processes.
[9] The organizational concept was adopted by FSO board of Directors in November 2007 as part of the MetaStat@FSO project, the goal of which is the revision of the current central metadata management system.

IV.I Centralization vs. decentralization and the creation of areas of harmonization

18. As work progressed on the organizational concept in 2007, it became clear that a strictly bipolar structure (centralized vs. decentralized) would not be sufficiently flexible to address coordination needs for the following reasons:

a. There are more than 200 statistical activities being carried out at the FSO. The central unit cannot have structured and direct links with all statisticians who are responsible for an activity.
b. Most issues related to metadata production do not need to be discussed with a central authority, but would benefit from some harmonization structure under the coordination of a subject matter manager.
c. Discussion about metadata content in production and its validation is only remotely a responsibility of the central unit dealing with the supervision of the central metadata system. Some involvement has to take place with statisticians, but not on a daily basis.
d. Enquiries submitted to the central metadata unit are likely to introduce high latencies in production workflows and this should be avoided as much as possible.

19. An alternative organizational model was proposed with the creation of intermediate units[10] in charge of harmonization at the level of integrated statistical production systems. In this context, the newly introduced production systems are seen as areas where metadata issues related to harmonization of content can be discussed (Illustration 2).

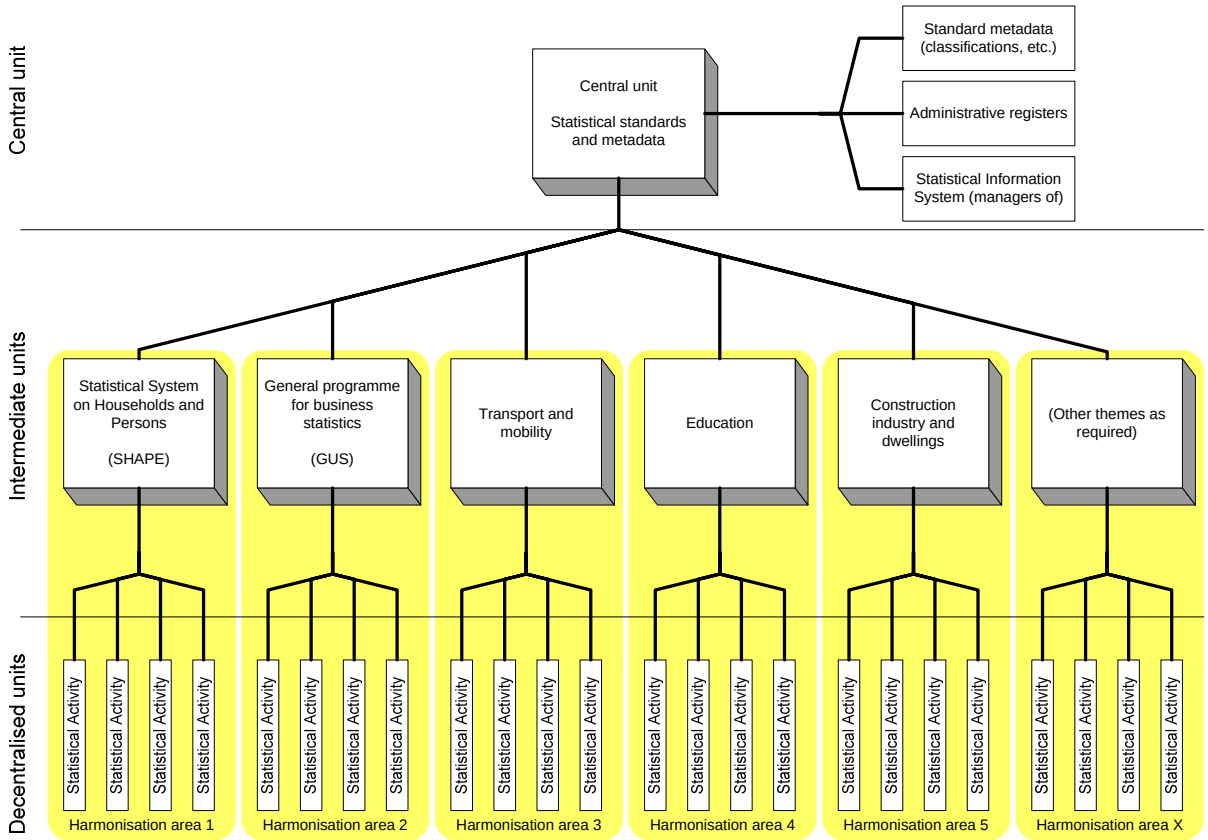

Illustration 2 - Organizational model with intermediate units

[10] Units as expressed here have to be understood in the context of a matrix organization. These units do not necessarily appear on the official organization chart of the FSO.

20. The introduction of a three layer organizational model is possibly more complex than the first model that was envisioned, but the role model of actors is much clearer with the introduction of intermediate units, as it allows a better distinction between the tasks of the three entity types in the organization:

a. The central unit is concerned about the central metadata system and its evolution. It is not primarily a metadata content owner. It has the role of metadata editor, not of metadata producer. Tasks of the central unit include:

   i. Definition and overseeing of harmonization strategy of the FSO;
   ii. Definition of system rules and policies of the central metadata system;
   iii. Updating the metadata model and interfaces with other IT systems[11];
   iv. Creation of training programme of users interacting with the system;
   v. Quality audits of metadata produced by decentralized authors;
   vi. Representation of FSO in national and international standardization groups;
   vii. Direct overseeing of standard cross-domain metadata, such as main classifications and administrative registers. Also acts as router for requests which cannot be handled by one particular intermediate unit;
   viii. Definition and maintenance of the central list of conceptual variables and thesaurus terms.

b. Intermediate units are responsible for quality and harmonization of metadata produced in their area. They also have the role of metadata editor, though more focused on metadata content than on other aspects which are handled by the central unit. Tasks of the intermediate units include:

   i. Formal validation of metadata produced by decentralized authors;
   ii. Direct support of central metadata users;
   iii. Capture of metadata in the central system if this is not done by decentralized units[12].

c. Decentralized units regroup statisticians who are in charge of one or more statistical activities. This is where most of metadata content is designed and created in the central system. Decentralized units are the owners of metadata that are produced in their unit and are as such responsible for their quality. Tasks of the decentralized units include:

   i. Original metadata conception and maintenance in the central system;
   ii. Management of requests of metadata end-users, who deliver input for metadata change or versioning over time.

21. The global splitting of tasks among actors of the organization has been designed so that decentralized units do normally not have to rely on the central unit when designing and using metadata. Most harmonization issues dealing with content of metadata can be discussed at the intermediate level, and if dealing with cross-domain metadata, by the central unit.

22. The overall success of this splitting of tasks will be measured once the organization is in place. Ultimately, the success of the organization will rely on its ability to detect what metadata are to be defined as cross-domain metadata and should therefore be discussed in a wider circle rather than in a particular harmonization area. A risk exists that such questions will not be handled equally well in all harmonization areas and corrective measures will have to be taken.

[11] Other IT systems might include standard generic platforms defined in the Statistical Information System of the FSO (project SIS@FSO) or any other production application.

[12] This exception has been added for metadata writers who do not have enough interaction with the system to justify their training, e.g. classification specialists who produce new versions every 5 to 7 years and who otherwise have no interaction with the system.

IV.II Metadata validation workflow

23. The other organizational measure that will be introduced is a formalized metadata validation workflow. The goal of this formalization is to even the quality of metadata in the whole organization and ensure that all published metadata meet minimal quality requirements. Also, this should decrease the need to proceed to metadata correction at a later stage in the statistical cycle, which is currently an issue at the FSO. Some subject matter managers have already implemented this kind of "four eyes" control for metadata, so this will be a generalization of these measures at the level of the FSO[13].

24. In order to reduce as much as possible interference with production, metadata validation will occur at three predefined moments in the statistical cycle (Illustration 3):

a. End of initialization phase: this will be to ensure that statisticians make use of standard metadata in the conception of their statistical activity. This is also meant to ascertain that statisticians have a clear picture of methodological aspects they intend to use, and to identify external metadata dependencies with other activities. This validation is a formalization of what has been done until now at the FSO outside of the metadata system, mostly in concept documents. An early introduction of metadata in the central system allows the automation of several controls which would otherwise not be possible.
b. Early data analysis phase: this is mostly to ensure that standard codification practice has been respected, and that descriptive and quality metadata are there to guarantee a proper interpretation of data. This second validation is particularly important for statistical activities which are analyzed by statisticians who were not involved in data collection phase.
c. End of information enrichment phase: this is the final control before publication. Descriptive metadata produced in later phases, mostly based on basic metadata elaborated earlier in the statistical cycle, are reviewed here.

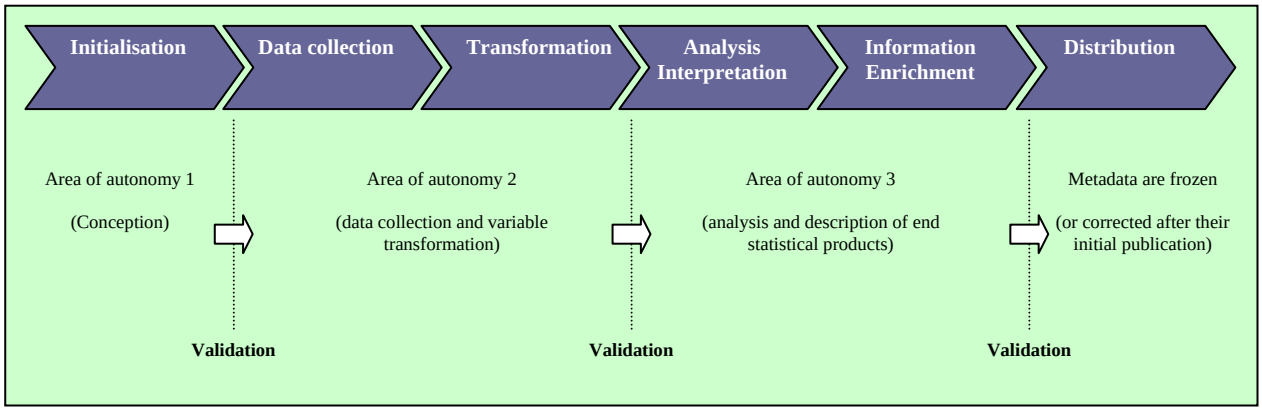

Illustration 3 - Metadata validation in the statistical cycle

25. Except when validation occurs, metadata producers can work with great autonomy in the central metadata system.

26. Metadata validation will be organized so that a given metadata item is validated at least once in the statistical cycle, before it is made available to its users (internal or external users). The proper time for validation is given by the phase in the cycle in which metadata will be used. Obviously, the sooner it is used, the sooner it has to be validated.

[13] Systematic controls are already in place for end statistical products as part of the "good output practice" policy introduced some years ago at the FSO, but they do not focus on metadata validation or assessment of their adequacy for a particular target audience.

V. THE END-TO-END VISION IMPLEMENTED IN A NEW METADATA SYSTEM

27. All of the above outlined processes and organizational measures ultimately rely on the existence of a central metadata system implementation which enables an end-to-end processing of metadata. The current central system, part of the CODAM project, only partially fulfils this need. Work on a new metadata system was thus initiated in 2007 as part of the MetaStat@FSO project, which is to become the backbone of the statistical information system (SIS@FSO) currently being developed.

28. The new central metadata system defined in the MetaStat@FSO project is meant to be a second iteration based on experience made with the current central system, Bridge^{NA}[14]. While essentially maintaining the same paradigm in metadata management, many changes are planned on both conceptual and technical elements which will lead to the creation of a new implementation. These changes will simplify metadata management or bring new functionalities required to enable an end-to-end management of metadata in a single central system. Changes focus on the following aspects:

a. Conceptual aspects

   i. MetaStat@FSO still relies on the Neuchâtel Terminology Model (classifications and variables) and the MetaNet Reference Model[15]. Whenever differences may appear in naming conventions between the two models, names used in the Neuchâtel Terminology Model will be favoured.
   ii. Many parts of the BridgeNA model which were not used so far will be used in production to cover the needs of the integrated statistical production systems[16].
   iii. An overall generalization of the model has been made to ensure that future needs may be accommodated without breaking the rules imposed by the model. This is for instance seen in the way several constraints on relations between concepts have been lessened to allow for more flexibility, e.g. by transforming “one to many” relations in “many to many” relations.
   iv. Several new concepts have been added, such as the notion of target audiences, which allow a textual metadata attribute to be instantiated differently for different groups of people or for different outputs as needed by specific applications.
   v. Several new attributes have been added for FSO specific needs. These are mostly textual attributes which have no overall incidence on the model.

b. Technical aspects

   i. Emphasis has been given to changes in the way metadata versioning is to be handled by the system. The Swiss statistical system uses many classifications which can be characterized as slowly changing dimensions[17]. In order to avoid massive metadata redundancy when minor changes have to be documented, change management has been implemented on a very molecular level to describe only actual changes in the definition of classifications, value domains and contextual variables.
   ii. Metadata will be migrated to a relational database. This has always been a major requirement from IT people that was not met by the Bridge^{NA} system[18]. This change in implementation will

[14] Bridge^{NA} is developed by run Software Werkstatt GmbH and is available as a commercial application. The FSO uses a customised version with specific attributes added to the core metadata model as well as some additional import and export tools.

[15] Both models are essentially based on the same modelling paradigm. The MetaNet Reference Model contains additional objects and attributes which are (as of yet) not part of the Neuchâtel Terminology Model. These typically include metadata types which are used to link conceptual metadata definitions with the actual data, such as where the data are physically stored.

[16] Metadata types considered are mentioned in section III of this paper.

[17] This means that structural changes occur frequently, but are usually limited to a small number of items of the classification. This is seen for instance in Swiss geographical classifications which tend to be revised several times in a year to reflect structural changes such as merge of political communes.

[18] BridgeNA makes use of a proprietary object oriented database which can be accessed using specific interfaces.

provide the ability to better index metadata for search purposes and simplify interconnection with other databases based on the same database engine.

iii. Access to metadata will be done through the definition of a service layer as part of the SOA[19] governance defined in the statistical information system (SIS@FSO). External applications will then be able to take advantage of reusable services to produce new content or access existing metadata while making use of business rules already implemented in the services.

29. The central metadata system defined in the MetaStat@FSO project is this time meant to gradually replace all decentralized systems. This will address most problems identified in the current metadata strategy of the FSO and enable an end-to-end management of metadata across the whole statistical cycle.

[19] SOA is the abbreviation of “Service oriented architecture”.