WP. 8
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 (ii) Metadata concepts, standards, models and registries

REGISTRY FACILITIES FOR SUPPORTING THE EXCHANGE OF STATISTICAL DATA AND METADATA

Submitted by Eurostat [1]

I. INTRODUCTION

1. The SDMX initiative (Statistical data and metadata exchange) is aimed at developing efficient processes for exchange and sharing of statistical data and metadata among international organisations and their member countries. In the years (the initiative started in 2001) SDMX produced a set of technical standards and guidelines to be used for the exchange of aggregate statistical data and metadata by computer systems. The SDMX technical standards are recognised as ISO TS 17369 in the 1.0 version, while version 2.0 is in the process of ISO validation. Full information on the SDMX standards, guidelines and organisation is available at http://www.sdmx.org.

2. The version 2.0 of the technical standard introduced a series of enhancement on the previous version, in particular on metadata management (with the introduction of the "metadata structure definition" to describe the structure of a metadata set) and on the so-called registry architecture, useful for providing visibility to large amounts of data and metadata. Borrowing from the SDMX 2.0 Framework description, "a registry – as understood in web services terminology – is an application which stores metadata for querying, and which can be used by any other application in the network with sufficient access privileges. It can be understood as the index of a distributed database or metadata repository which is made up of all the data provider's data sets and reference metadata sets within a statistical community, located across the Internet or similar network".

3. Eurostat intends to develop the use of SDMX 2.0 within the European Statistical System and to exploit the standard to improve internal collection, production and dissemination processes; this strategy was endorsed by the Statistical Programme Committee (SPC) at its meeting in February 2007.

4. This document reports on the implementation and setup of the SDMX registry at Eurostat, as this is the cornerstone of an architecture for data and metadata exchange aimed at facilitating collection, processing and dissemination of statistics. The paper describes the distinct registry modules and its purposes, also providing a

[1] Prepared by Beng-Ake Lindblad, Marco Pellegrino and Francesco Rizzo.

highlight of the on-going work for enabling end-users and metadata producers to access, analyse and reuse statistical metadata.

II. WHAT THE SDMX REGISTRY IS AND WHAT IS NOT

5. Registries are a generic form of technology which is used in many ways in various applications. For instance, in modern computers, registries are used by local programs to allow other programs to know that they exist on the local machine, and how they can be accessed. Registries are also used to support distributed applications which work on many different computers and servers around a network. These types of registries allow programs on different users' machines to know how to interact with each other: they are used by applications and not directly by end users.

6. An SDMX Registry is an application of this simple idea to statistical exchange for having a central repository which can be accessed by computer programs over the Internet (or an Intranet or Extranet) to locate and access statistics.

7. It is important to stress that registry services are not concerned with the storage of data or reference[2] metadata sets. Data and metadata are stored elsewhere, on the sites of its data providers. The registry is only concerned with providing information needed to access the data and reference metadata. An application which wants a particular data or metadata set would then query the registry for the URL, and then go and retrieve the data from the provider.

8. The Registry is not a single, centralized resource. It can be operated by many different users, each for the purposes of their own statistical community. SDMX assumes that large organizations which help to organize and administer the collection of specific types of statistical data within a statistical community will naturally be the operators of the SDMX Registry technology for the benefit of that community, and for the type of data they collect and disseminate.

A. Functions of the SDMX Registry

9. An SDMX Registry performs a number of tasks:
- It provides information about what data sets and metadata sets are available, and where they are located.
- It provides information about how the data sets and metadata sets are provided: how often they are updated, what their contents are, how they can be accessed, and similar questions.
- It provides information about the structure of data sets and metadata sets, answering questions like: What code lists do they use? What concepts are involved?
- It allows applications to sign up (or subscribe) for notifications, so that when a data set or metadata set of interest becomes available, the application will be automatically alerted.

10. These functions form the basis on which the SDMX Registry is organized. There are three layers, which correspond to the first three points above, while the subscription/notification functionality is available for all of these layers:
- The Data and Metadata Registry
- The Provisioning Metadata Repository
- The Structural Metadata Repository

[2] According to the SDMX Metadata Common Vocabulary, *reference* metadata are metadata describing the contents and the quality of the statistical data, normally including "conceptual" metadata, describing the concepts used and their practical implementation; "methodological" metadata, describing methods used for the generation of the data (e.g. sampling, collection methods, editing processes); and "quality" metadata, describing the different quality dimensions of the resulting statistics (e.g. timeliness, accuracy). These metadata are often stored in a separate metadata repository and they are referenced from the related data element.

B. Registry architecture

11. In general terms, the SDMX Registry is based on a structural metadata repository which supports a provisioning metadata repository which supports the registry services, according to a "layered" architecture as represented in figure 1.

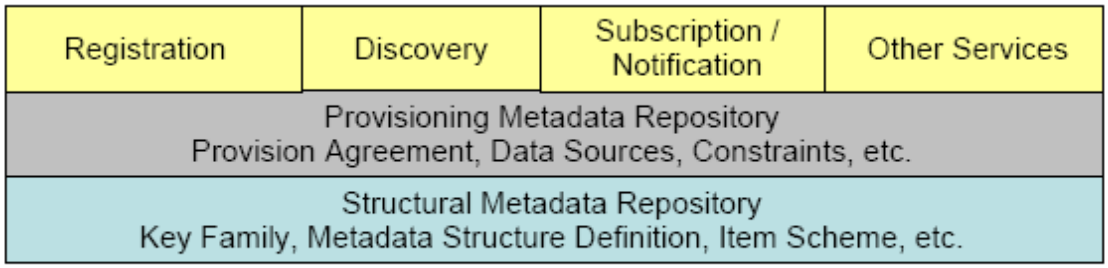

Figure 1: Schematic Architecture of SDMX Registry/Repository

12. **The Structural Metadata Repository Layer**. The Structural Metadata Repository Layer contains metadata such as Data Structure Definitions (previously called "key family" in Gesmes), Metadata Structure Definitions, Maintenance Agencies, etc. This layer must allow structural definitions to be created, modified and removed in a controlled way, also allowing the structural metadata to be queried and retrieved either in part or as a whole. Structural metadata is information about how data sets and metadata sets are structured. This type of information is needed by applications to process the data and metadata sets. Thus, once an application has discovered and retrieved a data set, it can then query for the structural metadata which goes along with that data set. In addition to concepts and code lists, the structural metadata repository contains many other pieces of needed information, including categorization and classification schemes, lists of organizations, and so on.

13. **The Provisioning Metadata Repository Layer**. Provisioning metadata is information about how data and metadata sets are made available by data providers. This is analogous to a "service level agreement" whereby a data provider commits to publishing a dataflow or metadataflow according to an agreed schedule. This layer includes details about the online mechanism for getting data (e.g., a queryable online database or a simple URL) as well as information about the release calendar, sources and contents of the data and metadata sets. This information is stored in the SDMX Registry, which is why this layer is termed a "repository". All of its information is accessible over the Internet using SDMX-ML messages, just as for all communications with the SDMX Registry.

14. **The Data and Metadata Registry Layer**. This portion of the SDMX Registry acts like a catalogue or a phone book, allowing applications to look up and see which data and metadata are available. Data and metadata sets are categorised to facilitate searches. Although there is a recommended high-level categorisation for statistical data in SDMX, each Registry can have a tailored categorisation which matches the statistics within the statistical community that the registry serves.

15. **Subscription/Notification**. A user may wish to receive updates regarding a specific part of the contents of any of the layers of the Registry, for instance when a new data set is published or when a list of organizations is updated. There are two ways to receive such updates: the first is the subscription/notification mechanism, using SDMX-ML messages. Another mechanism is the use of RSS feeds[3] which is typically used for updates to data. In either case, the update can serve as a trigger for the receiving application – to go out and get the updated or new data set, or to perform some other automated process.

16. As the objective of the SDMX Registry is to allow organisations to publish statistical data and metadata in known formats such that interested third parties can discover and interpret them accurately and correctly and within the shortest possible timescale, the setup of structural metadata and the exchange context (referred to as "data provisioning") is a key issue, which involves a series of steps for maintenance agencies:

- Agreeing and creating a specification of the structure of the data (called "data structure definition, DSD) which defines the dimensions, measures and attributes of a dataset and their valid value set

[3] RSS stands for "really simple syndication".

- Defining a subset or view of a DSD which allows some restriction of content (called a “dataflow definition”)
- Agreeing and creating a specification of the structure of metadata (metadata structure definition, MSD) which defines the attributes and presentational arrangement of a metadata set and their valid values and content
- Defining a subset or view of an MSD which allows some restriction of content (called a “metadataflow definition”)
- Defining which subject matter domains are related to the dataflow and metadataflow definitions to enable browsing
- Defining one or more lists of data providers (which includes metadata providers)
- Defining which data providers have agreed to publish a given dataflow and/or metadataflow definition - this is called a provision agreement

17. Publishing the data and metadata involves the following steps for a data provider:
- Making the metadata and data available in SDMX-ML conformant data files or databases (which respond to an SDMX-ML query with SDMX-ML data) - the data and metadata files or databases must be web-accessible, and must conform to an agreed dataflow or metadataflow definition (data structure or metadata structure subset)
- Registering the published metadata and data files or databases with one or more SDMX Registries
- Notifying interested parties of newly published or re-published data, metadata or changes in structural metadata. The Registry can optionally support a subscription-based notification service which sends an email announcing all published data that meets the criteria contained in the subscription request.

18. Discovering published data and metadata involves the following steps:
- Optionally browsing a subject matter domain category scheme to find dataflow definitions (and hence DSD) and metadataflows which structure the type of data and/or metadata being sought
- Build a query, in terms of the selected data structure or metadata structure definition, which specifies what data are required
- Submit the query to an SDMX Registry which will return a list of (URLs of) data and metadata files and databases which satisfy the query
- Processing the query result set and retrieving data and/or metadata from the supplied URLs

III. DEVELOPMENT OF THE EUROSTAT SDMX REGISTRY

19. Eurostat's Registry has been developed in 2006-2007, implementing the specifications from the SDMX 2.0 standards. Eurostat has also developed and deployed the Data Structure Wizard application, which is designed to work with the Eurostat SDMX Registry for the creation, editing and viewing of Data Structure Definitions (DSD).

20. Eurostat’s registry will be used as a "back-office application" for internal access to data and metadata structure definitions by eDAMIS (the single entry point), the SODI infrastructure and by any other information system inside Eurostat. The registry will also enable national statistical institutes to obtain SDMX structure definitions and dataflow agreements.

21. As of March 2008, the first version of the registry is already installed in the European Commission Data Centre test environment and is currently under the stress test procedure. Once this procedure is completed, the registry is going to be transferred to the production environment where it will be accessible by external organisations The registry software has been published as Open Source Software under the EU Public Licence; it can be downloaded from http://sdmx.org/index.php?page_id=52 or from CIRCA (see links under VI. References) .

22. The registry comprises of three major blocks:
- The Database (DB) which is the storage of all the data maintained within the Registry.

- The Web Service (WS) which exposes the registry interface via Simple Object Access Protocol (SOAP).
- The Graphical User Interface (GUI), a web interface for human interaction with the registry. The GUI offers a user-friendly web interface for adding/deleting/updating structural information, as well as import/export features for interaction with SDMX-ML and GESMES structural files.

23. The Eurostat SDMX Registry will become the core of a general-purpose "Metadata Handler" environment at Eurostat. The Metadata Handler will ultimately provide a common environment for managing all kinds of structural and reference metadata, providing services to users (via a GUI or via associated applications such as the Data Structure Wizard, see below) and to other applications, via the web service. In principle the Metadata Handler web service will be accessible for any application inside or outside Eurostat (subject to appropriate security mechanisms) providing metadata services currently available through several existing metadata applications.

24. The first release of the Metadata Handler (planned for 2008) will provide a common user interface for the management and retrieval of structural metadata, and the web service for access to structural metadata for applications. It will replace the existing GESMES Structural Metadatabases used by Eurostat and ECB, which are the reference source for the GESMES Data Structure Definitions used for data transmission from Member States.

Figure 1 – welcome screen of the Eurostat SDMX registry

IV. REGISTRY CONTENT AND IMPLEMENTATION

25. The Registry provides a web-based user interface and also a mechanism for the maintenance of the SDMX objects in use within Eurostat and with statistical partners (Concepts Schemes, Code Lists, Data

Structure Definitions, Metadata Structure Definitions, Data flows, Metadata flows, Category Schemes, Organization Schemes, Provision agreements).

26. The Registry is complemented with other off-line and on-line applications and specific modules being developed for handling specific data or metadata objects.

A. The Data Structure Wizard

27. The Data Structure Wizard is a desktop application that is able to convert/edit commonly used metadata formats into SDMX-ML formats. It contains an interface that allows the user to select a given Data Structure and to complete the data according to the requirements. The Data Structure Wizard is a Java standalone application that supports version 2 of the SDMX standards. It can be used both off-line and on-line, depending on user choice and access rights.

28. The off-line mode is intended to be used for the maintenance of the following SDMX artefacts: Data Structure Definitions, Code Lists, Concept Schemes, Data Flows, Hierarchical Code lists, Category Schemes and Organization Schemes. A local repository is created by default in the application installation folder in order to store the XML files with artefacts.

29. In the on-line mode, users can perform the same operations as in off-line mode plus the possibility to interact with any standard compliant SDMX Registry.

B. The use of the SDMX registry architecture for promoting the exchange and management of reference metadata

30. The use of the registry for facilitating the exchange and re-use of metadata from several sources can be demonstrated by a pilot initiative being conducted within the SODI[4] project for data sharing and exchange within the European Statistical System. SODI started in 2005 with a pilot exercise involving five National Statistical Institutes (Germany, France, Netherlands, Sweden and United Kingdom). The statistical institutes of Denmark, Italy, Norway and Slovenia joined the pilot in 2006, while Finland and Ireland joined the exercise in 2007.

31. Eurostat, in coordination with Member States, intends to make use of the SDMX standards for complementing data with a consistent and standardised layer of metadata which could be easily exchanged, read and processed by computers. The progress made on the identification of commonalities in the existing metadata systems and on the standardisation of terminology and concepts used (see SDMX cross-domain concepts and metadata common vocabulary) will help reducing the metadata reporting burden of national institutes and, at the same time, will improve the quality and consistency of metadata descriptions across countries. For these purposes, the registry architecture proposed by SDMX was considered to be the best option.

32. In 2006-2007, Eurostat therefore launched a specific project (“SDMX-compliant Metadata for use within Eurostat's Web Site”) aimed at demonstrating how SDMX technical standards and content guidelines can support the exchange and web dissemination of reference metadata. To realise this goal, it has been necessary developing:

- tools and standard formats for the creation and management of SDMX-compliant Metadata Structure Definitions (MSD);
- tools allowing the creation, transfer and management of reference metadata in SDMX-ML format, using as much as possible information already available in existing metadata repositories;
- registry-based architecture for transferring reference metadata to external users and to the web-site.

33. The project, based on the information model of SDMX 2.0, has been organised into four main tasks, as follows:

[4] SODI stands for "SDMX Open Data Interchange".

- Architecture analysis of the metadata environment; design and development of metadata structure definitions (MSD) and SDMX-ML schemas.
- Metadata converter/editor for transferring and retrieving SDMX-ML reference metadata from national and international sources.
- Loading metadata into Eurostat's database; formulating queries for extracting the information.
- Dispatching SDMX-compliant metadata to Eurostat's website.

34. A context diagram that depicts the different software components that are being implemented in the framework of the project is provided in Chart 1.

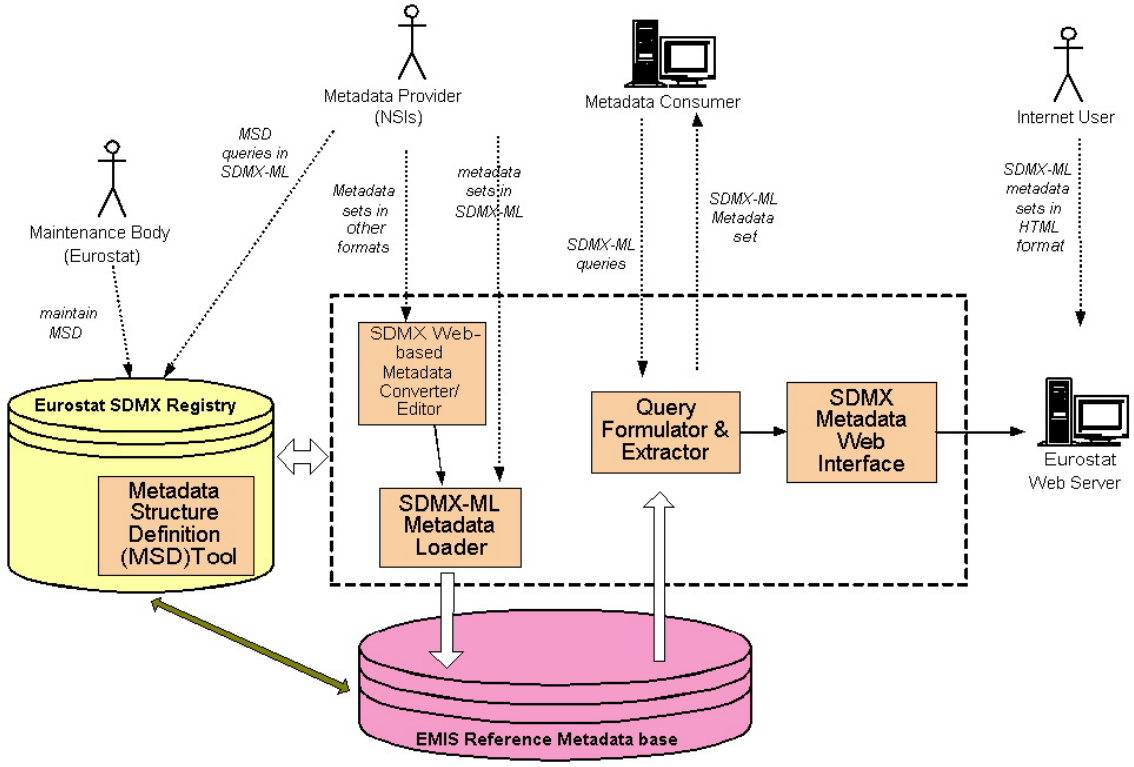

Chart 1*: Architectural Overview of the Software Components*

35. The first task of the project concerned the design, development and testing of a web service module for the definition and management of metadata structure definitions identifying which metadata concepts are to be reported; the identity; format and representation (textual or coded); the role in its usage (e.g., mandatory or conditional); the data structures to which metadata can be attached. A prototype of MSD tool is at present being tested and will be further enhanced during 2008 with a view to improving its user-friendliness and its functionalities. The tool will be populated with MSDs and corresponding XML schemas. The web-based application able to store and manage MSD will use information that resides in the SDMX registry.

36. The development and testing of a web service module for editing/converting metadata into SDMX compliant format included the possibility of transferring metadata from national systems to Eurostat using SDMX technical specifications. A web form was constructed to assist countries which are not currently in an advanced state for generating SDMX-ML metadata sets. An SDMX "Reference Metadata Converter/Editor" is part of a metadata processing environment which provides a new approach for collecting and disseminating metadata data from EU Member States. The tool is intended to be used over the Internet environment by countries that cannot currently generate SDMX-ML metadata sets. This way, the transmission/submission process of this kind of information into Eurostat's operating environment will be facilitated. The Metadata Converter/Editor introduces an efficient way for handling and editing reference metadata using a standard web

questionnaire/template based on metadata concepts and report structures. The template will be dynamically created from Metadata Structure Definitions existing in the SDMX Registry application.

37. The following functionalities are provided:
- to create metadata in SDMX-ML format;
- to select the domain for which metadata should be reported;
- to preview metadata reports submitted in the past;
- to preview the reference period of the metadata report;
- to preview the filled metadata template and be informed if there are parts that have not been completed
- to save submitted responses and continue/upload unfinished parts of the template/questionnaire at a later stage;
- to manage versions of the metadata reports;
- to import a Metadata Template from an existing Microsoft Office 2003 XML Word document;

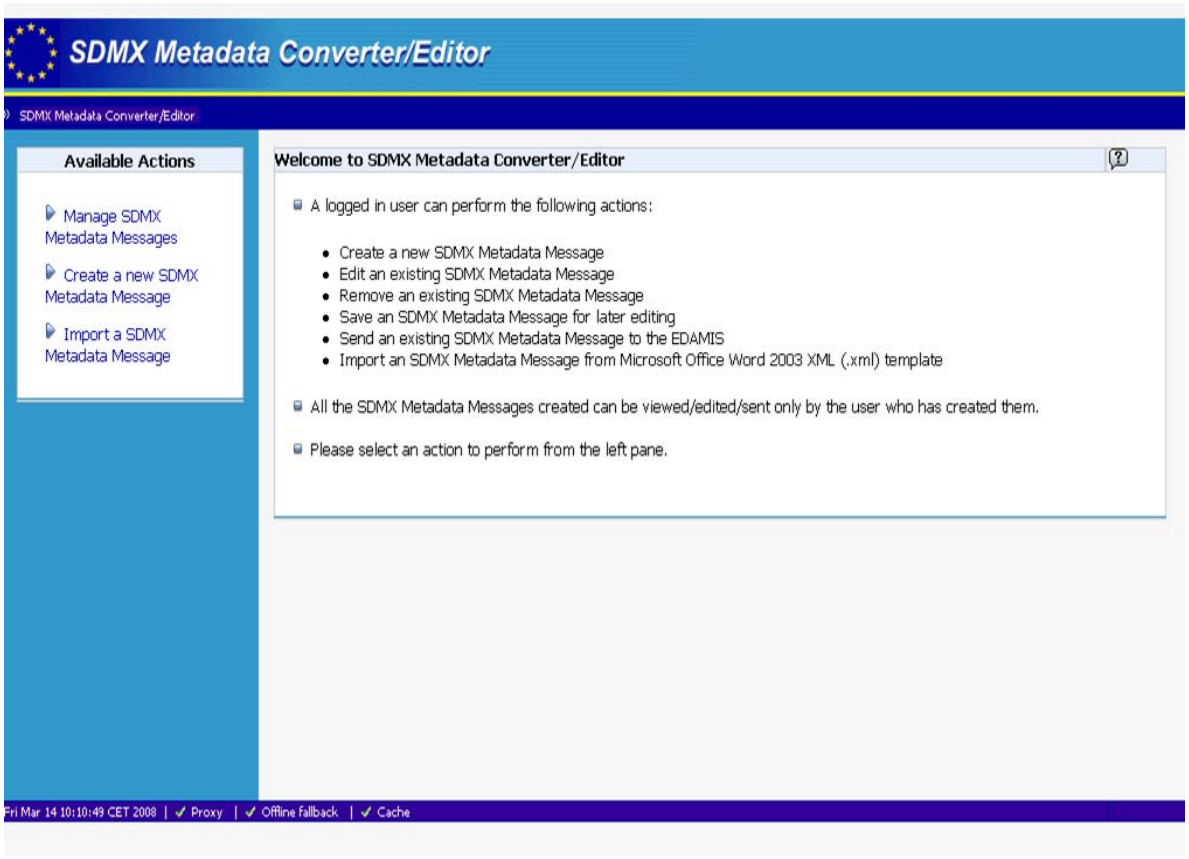

Figure 2 – SDMX Metadata Converter/Editor

38. The third task of the project consisted of the loading of SDMX-ML metadata, coming either from the SDMX-ML metadata converter module or directly in such format, into Eurostat's reference metadata base (EMIS) using a metadata loader web service. In addition, a web service extracting metadata from the database via dynamic formulation of SDMX queries (Query Formulator and Extractor) is being implemented. This tool will be able to respond to structured SDMX-ML queries forwarded by either web services (metadata consumers) or human operators. In this, it will also serve the exchange with other SDMX partners interested in re-using Eurostat's metadata.

39. Finally, the SDMX metadata publisher web service will be the interface with Eurostat’s web site. This tool will be responsible for transforming the received SDMX-ML metadata sets into publishing formats (via XSLT). This tool will enable interaction with usual Internet users.

C. Implementation strategy for reference metadata

40. Eurostat's SDMX implementation strategy for reference metadata in the European Statistical System foresees, as one of its main areas of intervention, the application of SDMX to statistical domains. The physical implementation of reference metadata will be assisted by Metadata Structure Definitions, SDMX-ML metadata reports, including schema files for their different SDMX report formats where needed. These structures and schema files should of course be registered in the SDMX registry (within the future Metadata Handler environment, see "III. Development of the Eurostat SDMX Registry") where they can be accessed by any SDMX-enabled system via web service in order to locate the metadata reports.

V. STANDARDISATION OF METADATA CONCEPTS

41. The registry architecture is a valuable resource and a key element for managing data and metadata flows, but this cannot substitute the need of a content harmonisation in the production of metadata. The SDMX Metadata project involves providing users and producers with a standard layer of metadata concepts and definitions, together with an open architecture and common tools. The Metadata Concept Scheme on which the project is based (called Euro-SDMX Metadata Structure, see METIS 2008, WP 10, "Structural and reference metadata in the European Statistical System") is maintained by Eurostat and makes use of the set of cross-domain concepts released by SDMX, covering both data descriptions and quality assessment elements.

42. The key idea is that, through the alignment to a common set of concepts linked to a standard terminology, there is a concrete possibility of setting up an exchange of reference metadata among countries and international organisations using web services to navigate, find and process the information, reducing redundancies and increasing comparability. Alignment does not necessarily entail the direct adoption of precisely the same format by each agency in its internal workflow; although such adoption would facilitate the capacity of exchanging metadata between agencies, it is sufficient for organisations to be able to map their own granular concepts (developed to meet their own and needs) to the list of cross-domain concepts specified in the SDMX list and published by a maintenance agency, in this case Eurostat. This would facilitate direct access to metadata on the web instead of the current transmission by national agencies of different metadata to different international organisations. And this would also help achieve a reduction of effort at the national level. Using a common XML format and standardised web tools, one organisation would be able to identify and retrieve those metadata which are relevant for its own framework, avoiding duplications.

43. Eurostat also intends to make use of the new set of SDMX tools for the transfer of metadata between the platforms respectively run by Eurostat, European Central Bank and International Monetary Fund in the new "Euro area" page of the Dissemination Standards Bulletin Board. Eurostat, as the central statistical institute for the Euro area, in cooperation with the European Central Bank, can coordinate Euro-area requirements and metadata flows while interconnecting national metadata systems. European countries, on their hand, should be able to provide metadata to more organisations at the same time, for the same SDMX concept, using as much as possible information extracted from their original metadata systems, reducing manual interventions, double work and inconsistencies.

VI. REFERENCES

SDMX Standards, Version 2, November 2005: http://sdmx.org/index.php?page_id=16#package
- Registry Specifications: Logical interfaces
- Implementor's Guide for SDMX standards

SDMX User Guide, release 2007.1 (http://sdmx.org/index.php?page_id=38)

Eurostat Registry user guide (on the STNE/X-DIS library, http://circa.europa.eu/Public/irc/dsis/stne/library)