

SDMX GLOBAL CONFERENCE

PARIS 2009

EUROSTAT SDMX REGISTRY

(Francesco Rizzo, Bengt-Åke Lindblad - Eurostat)

1. Introduction

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 2001-2008, (the initiative started in 2001) SDMX produced a set of technical standards and guidelines to be used for the exchange of aggregated 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>.

The version 2.0 of the technical standard introduced a series of enhancements 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.

SDMX envisages the promotion of a data-sharing architecture using the pull mode to facilitate low-cost and high-quality statistical data and metadata exchange: a data reporting organization publishes data once, and lets their counterparties "pull" data and related metadata as required. The data-sharing architecture is based on the possibility of discovering easily where data and metadata are available and how to access them.

The SDMX Registry plays an important role in this architecture, in fact it can be seen as a central application which is accessible to other programs over the Internet (or an Intranet or Extranet) to provide information needed to facilitate the reporting, collection and dissemination of statistics.

In its broad terms, the SDMX 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 seen 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.

It is important to stress that registry services are not concerned with the storage of data or reference metadata sets. Data and metadata sets are stored elsewhere, on the sites of the data providers. The registry is only concerned with providing information needed to access the data and reference metadata sets. 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 or metadata set directly from the provider's web server.

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. Moreover it describes the distinct registry modules and its purposes.

2. Functions of the SDMX Registry

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.

These functions form the basis on which an SDMX Registry is organized. There are three layers, which correspond to the first three bullet 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

3. Architecture of the SDMX Registry

In general terms, an 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.

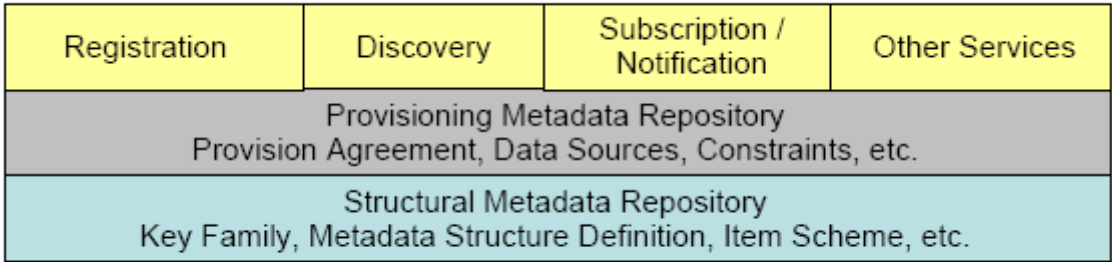


Figure 1: Schematic Architecture of SDMX Registry/Repository

Structural Metadata Repository Layer: the Structural Metadata Repository Layer contains metadata such as Data Structure Definitions, 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.

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.

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.

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 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.

As the objective of an 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.
- 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

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.

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

4. The Eurostat SDMX Registry

Eurostat has developed and put into operation an SDMX Registry which implements the specifications from the SDMX 2.0 standards. The Eurostat SDMX Registry provides a web-based user interface and web services for interacting with the SDMX structural metadata objects in use within Eurostat and with statistical partners (Concept Schemes, Code Lists, Data Structure Definitions, Metadata Structure Definitions, Data Flows, Metadata Flows, Category Schemes, Organization Schemes, Provision Agreements).

Eurostat's SDMX 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 other information systems inside Eurostat. The registry will also enable NSIs and other external organisations to obtain DSDs and other structural metadata, such as the MSD for the Euro SDMX Metadata Structure (ESMS).

The Eurostat SDMX Registry comprises 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 structure definition files.

Eurostat's SDMX Registry was developed during 2006-2008. As of September 2008, the first version of the registry was installed and running in three different environments:

- Test environment, accessible only internally at Eurostat and used for test purpose

- Production environment accessible at the following URL:
<https://webgate.ec.europa.eu/sdmxregistry>
- Training environment which can be used as a "sandbox" for training courses and presentations, without the risk of modifying the real Registry. Access to the training environment can be provided on request to Eurostat. It is accessible at the following URL:
<https://webgate.training.ec.europa.eu/sdmxregistry>

The production and training installations are accessible using the GUI via a CIRCA account in read-only mode. Web services are accessible, at the moment, only by internal applications; external applications will be able to access the registry web services in during 2009.

The registry already contains all the DSDs used by Eurostat and ECB, previously stored in the existing GESMES Structural Metadatabases. Eurostat envisages adding further content, with the aim of including all harmonised structural metadata and the ESMS (Euro SDMX Metadata Structure) MSD.

The Eurostat SDMX Registry software and the Data Structure Wizard (see below) have been published as Open Source Software under the EU Public Licence; it can be downloaded via the tools page of the SDMX website¹.

5. The Data Structure Wizard

Alongside the Registry, Eurostat has also developed and deployed the Data Structure Wizard application, which is a desktop application designed to work with SDMX-compliant registries for editing and viewing SDMX structural metadata objects. The Data Structure Wizard is a Java standalone application that can be used both off-line and on-line, depending on user choice and access rights.

The off-line mode is intended to be used for the creation and maintenance of the following SDMX objects: 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.

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.

6. How to use the Eurostat SDMX Registry and the DSW

As mentioned above, the Eurostat SDMX Registry is accessible to the SDMX community, both for human users (via the GUI) and for applications, via the registry web services. In addition, it is possible to download the Data Structure Wizard application and use it in off-line and on-line modes.

Most of the functions provided by the Eurostat SDMX Registry can be performed both via GUI and web service, but the GUI presents added functionalities particularly regarding creation, viewing and editing of Maintainable Artefacts². Below are listed the main functions that a user can perform through the GUI.

¹ SDMX website tools page: http://www.sdmx.org/index.php?page_id=13

² Maintainable Artefacts are: ItemSchemas (i.e. CodeLists, ConceptSchemes, OrganisationSchemes, CategorySchemes, HierarchicalCodeLists, etc.), StructureUsage (i.e. DataflowDefinitions, MetadataflowDefinitions), Structures (i.e. DataStructureDefinitions, MetadataStructureDefinitions).

Using a CIRCA account, an authenticated user is allowed to view the artefacts stored in the Registry, whereas an authorized user (that belongs to an Agency and has the appropriate permissions) can insert, edit, or delete artefacts belonging to the related Agency.

An authenticated user can view and download all the SDMX artefacts currently stored in the Registry. In particular, for a given DSD, he can download the related SDMX Structure file in both SDMX and GESMES format. Moreover he can extract the related XML Schema (.xsd) and the Message Implementation Guide³.

An authorized user, besides performing the same functions of an authenticated user, can create new artefacts and can upload files in GESMES and CSV format.

Statistical organisations could also use the portable version of the Eurostat SDMX Registry application as a basis for setting up their own in-house SDMX Registries.

The Data Structure Wizard application could be used by statistical organisations in the following ways:

- In on-line mode, as a tool to view DSDs and other structural metadata held in the Eurostat SDMX Registry. It should be noted that for certain purposes the Data Structure Wizard may offer more functionality or have a more user-friendly interface than the GUI of the Registry;
- In off-line mode, as a tool to support learning about SDMX, for example by allowing users to experiment with the creation of DSDs. Of course, locally-created DSDs will not be loaded into the Eurostat SDMX Registry.

7. References

- SDMX Standards, Version 2, November 2005 - Registry Specifications: Logical interfaces; Implementor's Guide for SDMX standards:
http://sdmx.org/index.php?page_id=16#package
- 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>)
- Registry facilities for supporting the exchange of statistical data and metadata (Lindblad, Pellegrino, Rizzo) – METIS April 2009

³ The Message Implementation Guide is a document in "rich text file" format that helps a user, not skilled in XML, in producing or even in consuming an SDMX data set based on a specific DSD.