

Distr.
GENERAL

WP.10
15 April 2010

ENGLISH ONLY

**UNITED NATIONS ECONOMIC COMMISSION
FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN
UNION (EUROSTAT)**

**ORGANISATION FOR ECONOMIC COOPERATION
AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE**

Meeting on the Management of Statistical Information Systems (MSIS 2010)
(Daejeon, Republic of Korea, 26-29 April 2010)

Topic (i): Developing common high-level architectures

SDMX architecture for data sharing and interoperability

Prepared by Adam Wronski, Eurostat and Francesco Rizzo, Istat Italy

I. Introduction

1. For several years the SDMX Sponsor Organisations have been developing software and tools in order to facilitate the introduction of SDMX in their countries.
2. Eurostat has been discussing the effectiveness of these tools with the Member States (MSs) during task forces meetings and working groups. Many suggestions from the MSs were collected which led in 2009 to an internal project named "SDMX implementation and support for Member States". The aim of that project was to develop a set of software, organized in re-usable building blocks and tools to help MSs in implementing an SDMX service infrastructure for a more efficient participation to some SDMX projects currently running within the European Statistical System (e.g. European Census Hub project).
3. A prototype was presented during the "Statistics, Telematic Networks & EDI" Working Group (STNE) in June 2009. The participants welcomed with enthusiasm the initiative and encouraged Eurostat to proceed forward.
4. The main objectives of the project can be listed as follows:
 - (a) To support for the Census Hub project and other Eurostat projects;
 - (b) To facilitate SDMX implementations within the MSs with a particular attention to the large PC-AXIS community;
 - (c) To stimulate a "SDMX community of developer".
5. The project deliverables consist essentially of:
 - (a) the SDMX NSI Reference infrastructure document. It represents the syntheses of several experiences worldwide and may be considered not as a strict specification but rather a guide or "best practice" document. The main objective of the document is to provide a

description/specification of a generalized infrastructure that could be re-used partially or entirely by NSIs interested in SDMX projects;

- (b) a set of software building blocks that can be used as APIs to be integrated in already existing statistical dissemination information systems;
- (c) the Mapping Tool software. It is a desktop application that allows mapping of concepts and code lists stored in a "local" dissemination database with concepts and code lists used in a SDMX Data Structure Definition (DSD).

6. In September 2009 a technical workshop on "*From the SDMX Information Model to the development of reusable software components*" was organized by Eurostat. During the workshop all the deliverables mentioned above were presented and discussed with more than 40 IT experts from 25 countries and organisations. A lot of feedbacks had been already received and many participants had been testing the software.

7. For the time being the version 1.0 of the software is available for downloading from CIRCA at the following URL, but version 2.0 is under testing and in few weeks will be available for downloading:
http://circa.europa.eu/Public/irc/dsis/stne/library?l=/x-dis/tools/reference_architecture&vm=detailed&sb=Title

II. Why to introduce SDMX in a NSI

8. There are many reasons why National Statistical Institutes could decide to use the SDMX standards. At the base of all there is the tremendous pressure on there resources of the statistical organisations which face everyday new data demands without a parallel financial allocation. Synergies, standardization and optimization of the processes and infrastructures are the only solution to this challenge. In this context SDMX can help by:

- (a) improving quality and efficiencies in the exchange and dissemination of data and metadata through:
 - harmonisation and coherence of data;
 - preservation of meaning by coupling data with metadata that defines and explains it accurately;
 - open format (XML) rather than a proprietary one;
- (b) reducing national reporting burden to European and international institutions, in fact a data reporting organization publishes data once, and lets their counterparties "pull" data and related metadata as required;
- (c) reducing costs through the re-use of the software;
- (d) facilitating and standardizing the use of new technologies as XML and Web services. Many NSIs are already using, or are planning to use, XML as the basis for their data management and dissemination systems. By choosing SDMX one could avoid the proliferation of many XML grammars.

II. SDMX architectures for data sharing and exchange

9. In order to facilitate the introduction of SDMX within the NSIs, Eurostat and other Sponsor Organizations have been putting in place different initiatives, among which capacity building actions and development of re-usable software and tools. Moreover during the last years Eurostat has developed two SDMX service infrastructures for data collecting that are used in several projects jointly with the Member States:

- (a) Data Repository (warehousing) architecture;
- (b) Data Hub architecture.

The offered SDMX service infrastructure could be used by national statistical authorities as an add-on to the NSI IT architecture and links to national reference or dissemination databases. Therefore it does not require any changes to the national IT architecture. Below the two SDMX architectures are described.

A. Data Repository (warehousing) architecture

10. The Data Repository architecture is implemented by those collecting organisations that periodically collect the data and to load them in their database. In general a batch process is used in order to automate the flow in which a whole or a partial dataset, including incremental updating, is used.

11. The Data Repository architecture supports both “push” and “pull” methods.

12. The push method within Data Repository architecture expects that the data provider sends a file in SDMX format to the data collector. In this case the data provider can:

- (a) Create SDMX format file directly while extracting data from the data warehouse, using a suitable software, or
- (b) convert a data file (generally in CSV) using tools available from the SDMX sponsors.

13. Then the SDMX file is pushed using the appropriate channel (eDAMIS in the ESS).

14. The pull approach within Data Repository architecture includes the following steps based on a provision agreement:

- (a) when data for transmission, the data provider creates an SDMX-ML file containing the to be transmitted data set or provides a Web Service (WS) capable building SDMX-ML messages upon request. Notification to data consumers about the available data and the details on how to obtain them are normally done with an RSS web feed;
- (b) the data collector Pull Requestor reads the arriving RSS feed entry (or receives the information on the new data by other means. He can now retrieve the SDMX-ML file from the specified URL or use the “Query Message” included in the RSS feed to query the data provider’s Web Service.

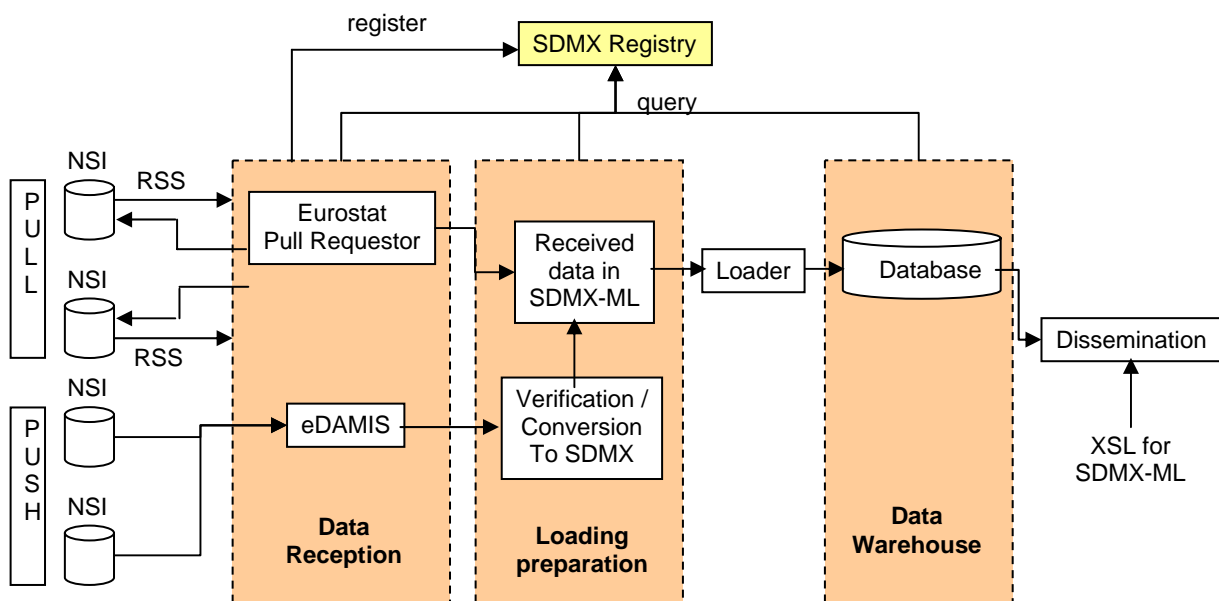


Fig. 1 Data Repository (warehousing) architecture

B. Data Hub architecture

15. The Data Hub architecture supports the “pull” method only i.e., a group of partners agree on providing access to their data directly from their database according to standard processes, formats and technologies (web service).

16. From the data management point of view, the hub is also based on a pre-specified datasets, which are - contrary to the database driven architecture - not kept locally at the central hub system. Instead the following process operates as follows:

- (a) A user identifies a dataset through the Web Graphical User Interface of the Data hub using the structural metadata, and requests it;
- (b) The Data Hub translates the user request in one or more queries and sends them to the related data providers’ systems;
- (c) Data Providers’ systems process the query and send the result to the Data Hub in standard format (SDMX-ML);
- (d) The Data Hub puts together all the results originated in all implicated Data Providers’ systems and presents them in a human readable format.

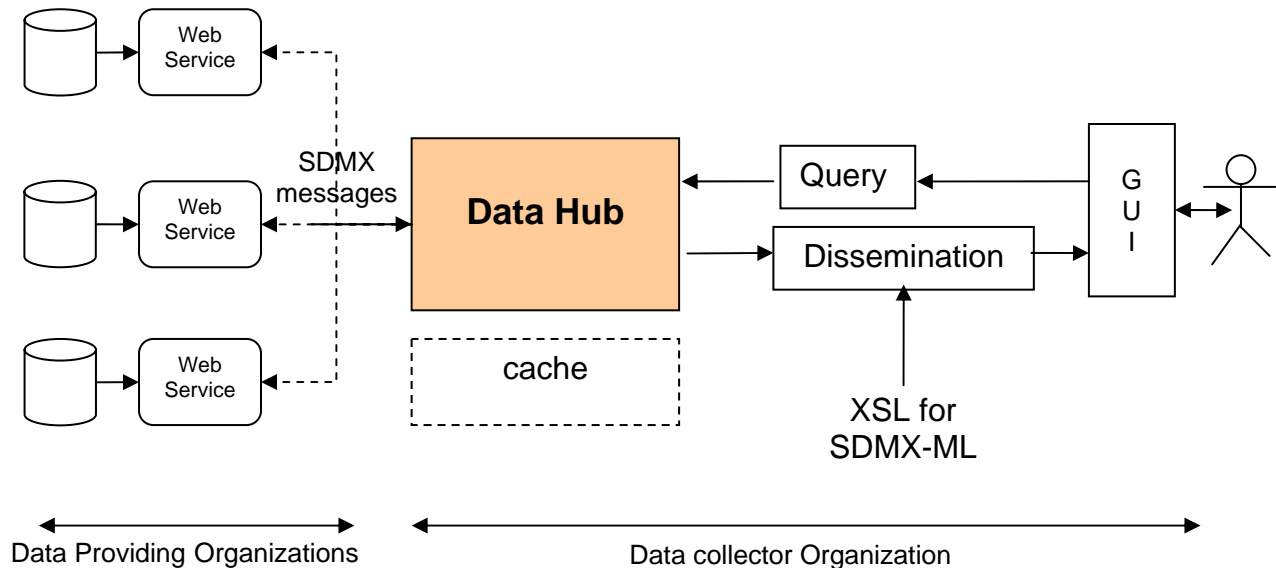


Fig. 2 Data Hub architecture

III. Strategies to foster SDMX implementations within NSIs

17. SDMX was born with improving quality and reducing costs ideas in mind. Eurostat for some time has been developing re-usable software in order to facilitate the introduction of SDMX within NSIs. These software can generally be freely downloaded (open source) from the SDMX website. The source code for these tools is available so that they can be used as components for building own IT systems in statistical organisations.

18. Sharing free software can have various forms: the distribution of tools developed by one member of the community for the benefit of the others or the joint development in a collaborative way such that each partner contributes to the final product. Eurostat is currently supporting both approaches with:

- (a) project aiming to design an SDMX service infrastructure for NSIs and developing related building blocks;

- (b) support, through SDMX ESSnet¹, a group of Member States that have joined their resources in order to develop SDMX re-usable software.

19. The main deliverable of the above approaches is the SDMX service infrastructure composed by several Building Blocks that can be re-used entirely or as single blocks to be integrated in an existing statistical information system.

IV. SDMX NSI Reference infrastructure

20. The infrastructure represents the syntheses of several experiences (in several statistical offices) and can be considered not a strict specification rather than a guide or “good practice” document.

21. The main objective is to provide a description / specification of a generalized service infrastructure that can be re-used partially or as whole by NSIs interesting in starting SDMX projects. To that end Eurostat have been developing software and tools that facilitate the production of SDMX data and their exposure via Web Services technologies.

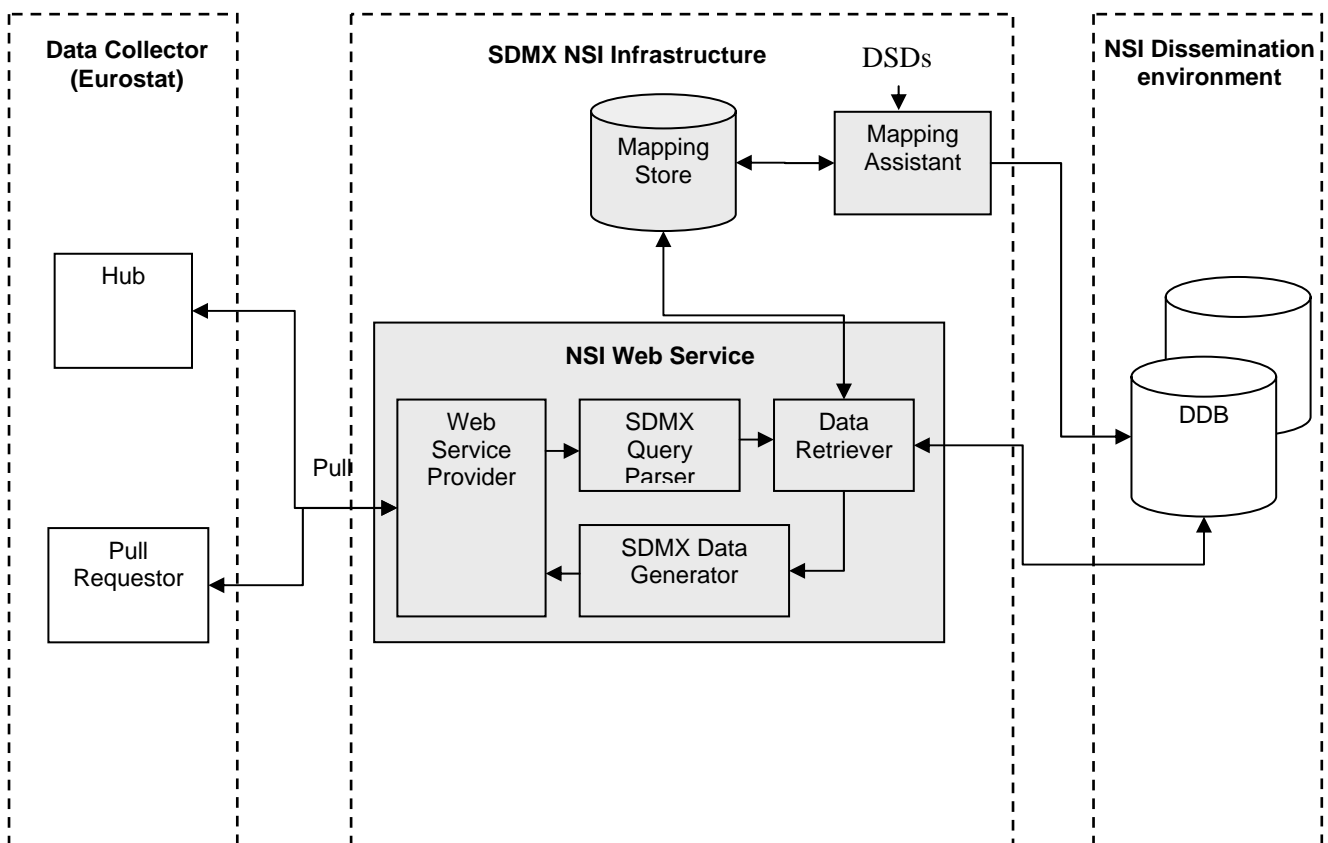


Fig. 3 A simplified view of the SDMX NSI Reference service infrastructure

22. In the Figure 3, three areas can be identified (bordered by dashed lines). The left-hand side area concerns the Data Collector, e.g., Eurostat. It contains the modules "pulling" SDMX data from a Data Producer, e.g., NSI. The right-hand side area concerns the Data Producer. The only part of Data Producer IT environment concerned the dissemination environment is presented in the figure 3. The dissemination

¹ ESSnet - European Statistical System Centres and Networks of Excellence is an instrument created by Eurostat in order to find synergies (from cooperation between partners), harmonization and dissemination of best practices in the ESS.

responsibility, inter alia, is to provide data to Data Collectors. The central area represents the software developed by Eurostat that acts like an interface between the Data Collector and the dissemination databases in Data Providers' environment

23. A NSI can decide to use the SDMX service infrastructure as a whole, can extend the infrastructure adding new modules, can modify some modules, or can integrate some building block within its existing dissemination environment.

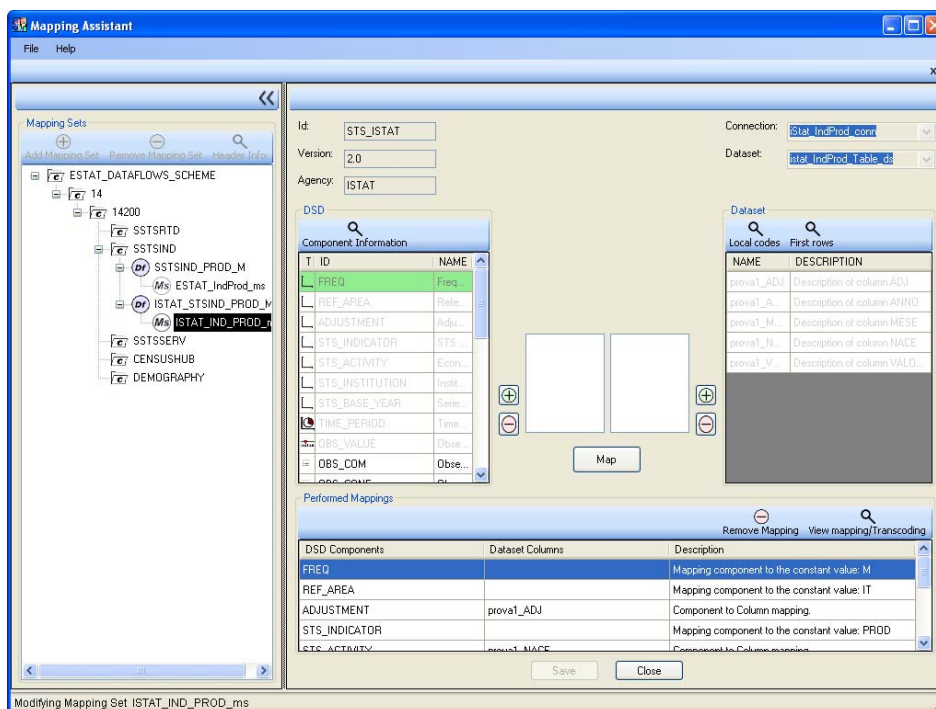
24. Details on the modules participating in this infrastructure are included in the following sections. More specific information can be found on CIRCA website at the following URL:

http://circa.europa.eu/Public/irc/dsis/stne/library?l=/x-dis/tools/reference_architecture/reference_architecture/EN_1.0_&a=d

25. Dissemination Database (DDB) is the storage data warehouse (or database) of the Data Provider dissemination environment maintained to store data ready for publication / dissemination to potential Data Collector. In some cases, the DDB may consist of files, e.g., PC-Axis files.

26. **Web Service Provider** is responsible for receiving an SDMX Query message and responding with an SDMX-ML data messages. It concerns the dynamic pull scenario. It also co-ordinates building blocks used when producing the response. This component exposes the underlying functionality using a SOAP interface.

27. The Mapping Assistant (MA) is a desktop tool allows user to create through a Graphical User Interface a mapping between the structure metadata provided by an SDMX-ML Data Structure Definition (DSD) and those that reside in a Dissemination Database of an NSI dissemination environment.



28. The Mapping Assistant is designed to edit and store the mapping information in a DBMS called Mapping Store (MS), and communicates with both the Mapping Store and the Dissemination databases in standard SQL.

29. The Mapping Store contains the mappings between the SDMX and the native format (a file or a DB schema). It is a database maintained by the "Mapping Assistant" in order to provide these mappings to the "Data Retriever" module.

30. A user creates with the help of Mapping Assistant:

- (a) DATASET;
- (b) MAPPINGSET;
- (c) TRANSCODING.

31. The DATASET defines a physical mapping of storage schemas from Dissemination databases or PC-AXIS files, to a DSD related schema in such a way that DSD component information resides in one or more DATASET columns: represented by SQL queries. E.g., one can map one or more columns of a Dissemination database to one or more dimensions of a DSD.

32. Alternatively a user can write a custom "select" SQL queries for the DATASET. SQL query belongs to four predefined query types relating columns and tables and joining tables to the parameters of the DSD (e.g., codes, dimensions, measures).

33. The MAPPINGSET contains the logical mapping between a DATASET and a DSD. It allows user defining relationships when the mapping of concepts used in the Dissemination database to concepts described in the DSD is not one to one.

34. For example the "local" concept in the Dissemination database named "Unit" could be mapped to two concepts in the DSD: "Unit of measure" and "Unit multiply".

35. The TRANSCODING relates codes from code lists in the Dissemination database to those in the DSD. This operation can be achieved directly through the GUI or importing the transcoding rules from a CSV file. E.g., transcoding on the Frequency concept:

"local" codes	DSD codes
1	A
4	Q
12	M

36. For the time dimension, the tool allows relating several kinds of time formats. E.g., if the Dissemination database time dimension is YYYY:MM:DD, the tool allows to map as YYYY for annual data, or YYYY-MM for monthly data

VII. Software maintenance and governance

37. The design of the SDMX NSI Reference Architecture and the development of the related building blocks were conceived with the goal of offering everything as open source package under the EUPL licence.

38. Up to now Eurostat is managing both the evolutive and adaptative maintenance. In the future the governance of the versioning could be difficult to achieve, because the re-using of the building blocks and their improvement by NSIs could bring to a scenarios with different versions of the same software created by different subjects.

39. For the time being there are very few experience of "open source" software development within the statistical community, so it is hard to learn from the experience. The experience developed worldwide in many "open source" communities could help, but they could be not adapted directly to the statistical community.

40. Know-how could come from NSIs participating in two ESSnet projects.