

# **A reference architecture for automatic XML data & metadata exchange between public administrations: Eurostat's case study**

Maria Glossioti, Gregory Farmakis, Kyriakos A. Kassis, Spyros Liapis, Efstratios Nikoloutsos

*Agilis SA Statistics & Informatics*

*Akadimias 96-100*

*106 77, Athens, Greece*

*{Maria.Glossioti, Gregory.Farmakis, Kyriakos.Kassis, Spyros.Liapis, Stratos.Nikoloutsos}@agilis-sa.gr*

## **Abstract**

*Interoperability plays a great role in the statistical data life cycle and basically allows businesses and citizens to have easier access to more timely statistical data with well-defined semantics between the figures. In a modern data exchange scenario business experts, organisations at EU and international level (such as Member States, UN, OECD, IMF, ECB etc.) and citizens may equally act either as data providers or as data consumers of statistical information. Their need is to automatically extract statistical figures, which are described using the same metadata which accompany the data during the process of collection and transmission.*

*In such a visionary scenario, aggregate data providers, such as government agencies, statistical authorities and institutes, disseminate their data through publicly accessible web services that can be queried in an ad hoc manner to provide just the requested data at the requested aggregation granularity. The availability of new data is published in web feeds, to which interested data consumers, such as other agencies, international bodies, companies and citizens, subscribe to be regularly notified. Since their needs differ, even for the same kind of data, they automatically submit to the public web services their predefined tailor-made queries, in order to be served with exactly the data and metadata they need.*

*A unified modern architecture, which materialises Eurostat's vision for EU-wide data sharing, has been deployed based on SDMX guidelines and OSS technologies. The guiding design rational behind the platform's architecture is to facilitate interoperability and to allow for extensibility while preserving platform-independence. The architecture is based on a set of web services, which cooperate in order to accomplish the main goal of interoperable exchange of statistical messages (containing both data and metadata). A web feed based publish-subscribe system implementing the pull data exchange scenario has been developed. In this context SDMX datasets are recuperated from the data provider's environment (in response to SDMX query request messages) each time new or updated data are available. Upon reception from Eurostat's operating environment the SDMX dataset is parsed, checked in terms of digital signature and encryption mechanisms and is syntactically validated. Then the normal workflow processing of the organisation may occur such as monitoring, notification of the recipient, storing in databases, and finally publishing (by applying XSL transformations) in Eurostat's web site.*

## **1. Introduction**

Eurostat and the Member States have been gradually increasing their use of standardised messages for the transmission, processing and dissemination of statistical data and metadata, and increasing use of the SDMX standards is expected. The Statistical Data and Metadata Exchange (SDMX) initiative (<http://www.sdmx.org>) sets

standards that can facilitate the exchange of statistical data and metadata using modern information technology, with an emphasis on aggregated data.

The SDMX standards are designed for exchange or sharing of statistical information between two or more partners. Evidently, the SDMX standards have been developed by the sponsors in order to accommodate the constituencies of the sponsoring organisations (national statistical offices, central banks, ministries, etc.). Within and across these constituencies, the standards are intended for reporting (or sharing) statistical data and metadata in the most efficient way. SDMX standards can also be used within a national system for transmitting or sharing statistical data and metadata and by private data providers (such as re-sellers of statistical databases). This is particularly interesting in countries with a federal structure or a fairly decentralised statistical system. In such cases, a close link can be established between the national system for data sharing and the international ones, allowing for additional efficiency gains for the involved organisations. If data are made available for exchange using the pull mode, according to SDMX standards, this could easily evolve to open SDMX-based dissemination; such dissemination may respond well to user demands for well-structured data and metadata in reusable formats, and should be considered as an option for national authorities as well as international organisations.

Three main projects actively use and exploit the implementation of SDMX standards in various statistical domains in order to ensure harmonisation with current business processes. SODI project is aimed at implementing SDMX in the collection and dissemination of the Principal European Economic Indicators (PEEI) and at a later stage in other suitable statistical domains. The SDMX Registry/Repository plays a central role in the data sharing exchange pattern in European statistics and allows organisations to maintain and publish structural statistical data (DSD – Data Structure Definition) and metadata (MSD – Metadata Structure Definitions) in known formats such that interested third parties can discover these data (through the submission of properly defined SDMX-ML messages) and interpret them accurately and correctly and within the shortest possible timescale. Finally, SDMX reference metadata relies on the SDMX registry infrastructure and facilitates the creation, management, querying and publishing of reference metadata (that is metadata regarding data quality, methodology, data provisioning or any other user-configurable concepts that require reporting).

## **1.1. Statistical data exchange in the European Statistical System (ESS)**

### **1.1.1. SDMX Overview and Web Services Guidelines**

In 2001, the Bank for International Settlements (BIS), the European Central Bank (ECB), the Statistical Office of European Communities (Eurostat), the International Monetary Fund (IMF), the Organisation for Economic Cooperation and Development (OECD) and the United Nations joined together to develop more efficient processes for the exchange and sharing of data and metadata within the current scope of their collective activities. The World Bank joined the initial group of sponsor organizations in 2003.

The goal of the Statistical Data and Metadata Exchange (SDMX) initiative is to foster standards and guidelines that allow national and international organizations to gain efficiencies and avoid duplication of work in the area of data and metadata exchange through the use of modern technology. Sponsor organizations have been making progress over the past few years, especially through the increasing involvement of international and national statistical agencies. SDMX builds on existing and emerging technical exchange protocols and on the content-oriented efforts of statisticians that have worked on these long-standing issues in various domains and fora. More details on the SDMX standard can be found at the SDMX website ([www.sdmx.org](http://www.sdmx.org)).

The SDMX initiative developed a guideline for using Web Services with SDMX-ML. The aim is to exchange SDMX-ML messages over the Internet. SDMX-ML is seen as the standard for representing statistical data and metadata whereas SOAP is seen as the standard for transporting the SDMX-ML messages. The SDMX Web Service should respect the WS-I Profile 1.1 standard for achieving the inter-operability between the Web Services.

Both the Web Services and SDMX-ML are well suited to be combined for achieving the inter-operability:

- Via their interfaces (set of functions) the Web Services should model the concepts and businesses using a standard abstraction. SDMX-ML achieves this because it is based on a common information model, which can serve as basis for defining the Web Service interfaces.

- The Web Services exchange data using XML. This allows the use a common way to serialize and de-serialize the data. SDMX-ML is, by essence, in XML.

The SDMX initiative defined a set of SDMX messages for different purposes. One of them is called the 'Query message' and is used to invoke an SDMX Web Service. It should expose functions that allow the users to query for information specific to SDMX (data, concepts, code lists, key families, meta data structure).

The SDMX initiative defined a multiple exchange pattern for the SDMX Web Services. The exchange pattern is expressed as a succession of functions. These functions are specific SDMX-ML messages that are self-descriptive. Thus, no parameters need to be defined within the SDMX Web Service interface (WSDL) since the SDMX standard proposes message based web service implementations.

### **1.1.2. Exchange of SDMX-ML messages using the push and pull approach**

Messages can be exchanged in two different modes, the push mode and the pull mode:

Push mode means that the data provider takes action to send the data to the party collecting the data. This can take place using different means, such as e-mail or file transfer, and in some cases the transfer can be supported by systems such as Eurostat's Stadium and Statel. These are the "traditional" modes of data collection, carried out by international organisations for many years.

Pull mode implies that the data provider makes the data available via the Internet. This may be as simple as placing a structured (SDMX-ML) file on a website or it may involve the use of web feed notification mechanisms, available via the web and capable of processing a standard SDMX query. The data collector then fetches the data on his own initiative. In this case, more than one data collector may be allowed to take the pieces of data needed by each collector. This mode also resembles dissemination in the sense that access might be given to final users of information, who will then, according to their needs, access multiple web sites all using the same formats.

While all combinations of the modes above are supported by SDMX standards, it is the aim of the SDMX initiative to further promote data sharing & exchange using the pull mode.

### **1.1.3. SODI Web Feed Scenarios**

The following distinct configurations for the SODI syndication mechanism emerge during the pull process between Eurostat and National Statistical Institutes (NSIs):

- The static scenario: NSIs prepare new available data in the form of SDMX files, which are maintained on a specific URL (such as an http or ftp server or a data delivery web service). Subscribers are notified by a simple feed on the availability and location - URI - of new or updated data files and fetch them. Only entire datasets (as maintained by the publisher) can be downloaded. This scenario requires just the provision of the location (URI) for the web feeds.

- The dynamic scenario: NSIs publish data on a dissemination data warehouse. When new data is loaded or updated, a notification containing a description of the new or updated data is automatically formulated in terms of statistical concepts and the SDMX standard (i.e. an SDMX query). The notification is included in the corresponding web feed. Subscribers can then parse this notification, and act accordingly (such as submitting a query for the entire dataset or restricting it). The notification can be digitally signed and encrypted if required, while subscriber aggregators can authenticate and validate it. This scenario requires apart from the provision of the location of the published data, the metadata that describe the published datasets.

The case of the real SODI pull scenario includes transmission of metadata within the web feed. These metadata describe the newly available datasets. Thus, a way for the effective transmission of this information within a web feed is essential in the SODI case. A possible way to achieve this is the transmission of a proper SDMX-ML query that would facilitate the SODI environment to "understand" the content of the newly published dataset.

## **2. Eurostat's SODI processing environment and Web Services Architecture**

### **2.1. Main Architectural Considerations**

A unified modern architecture which materialises Eurostat's vision for EU-wide data sharing has been deployed based on SDMX guidelines and OSS technologies. The guiding design rational behind the platform's architecture is to facilitate interoperability and to allow for extensibility while preserving platform-independence. The architecture is based on a set of web services, which cooperate in order to accomplish the main goal of interoperable exchange of

statistical messages (containing both data and metadata). A web feed based publish subscribe system implementing the pull data exchange scenario has been developed. In this context SDMX datasets are recuperated from the data provider's environment (in response to SDMX query request messages) each time new or updated data are available. Upon reception from Eurostat's operating environment the SDMX dataset is parsed, checked in terms of digital signature and encryption mechanisms and is syntactically validated. Then the normal workflow processing of the organisation may occur such as monitoring, notification of the recipient, storing in databases, and finally publishing (by applying XSL transformations) in Eurostat's web site.

All web services modules communicate through standard SOAP interfaces. Widely accepted as well as emerging standards and the use of Open Source Software have been adopted. Industry standards and technologies such as SQL, XML, Java, JDBC, SOA, SOAP, Web Services, Web feeds (ATOM and RSS), WSDL, J2EE, have been utilized to bring forth the desired objective.

## 2.2. SODI Reference Architecture

The component diagram depicted in Figure 1 presents the different components that participate in the SODI reference architecture. These components are grouped into subsystems or external systems, annotated with different colour for reasons of clarity. The SODI architecture consists of the following basic sub-systems:

The Pull Requestor includes the following components:

- SDMX Web Feed Client, which is responsible for checking, at periodic intervals, the feeds of the data providers, determining if the feed contain available data for download. These data can be either new or revisions. Moreover it checks if there is a provision agreement with the data provider for the dataflow specified in the feed. This is determined by accessing the SDMX Registry Web Service. The query of the Registry is based on the Registry Interface that is consisted by XML request/response messages (SDMXRegistry.xsd).
- Request Formulator, which is responsible for formulating the SDMX-ML request to be sent to the NSI's web service.
- The SDMX Retriever, which is responsible for retrieving the available data in SDMX-ML format by sending the previously created SDMX-ML request.

The SDMX Web service includes the following components:

- SDMX Web Service Operator, which is responsible for accepting SDMX-ML messages, calling the validator to validate the incoming SDMX-ML, using the logger for logging the occurring activities and finally for calling the Loader in order to load the data to the Eurobase. Additionally, this component dispatches the SDMX-ML to the eDAMIS PULL input directory using SCP connection, as well as to the publisher component for publishing the data to the Eurostat's web site.
- Metadata Requestor, which is responsible for retrieving the necessary metadata (e.g. XML Schemas) used by the Validator. The metadata are retrieved from the SDMX-Registry Web Service.
- Validator, which is responsible for validating the SDMX-ML in terms of syntax and code lists, using the corresponding XML Schemas.
- Logger, which is responsible for logging the occurring activities, like incoming SDMX-ML, success or failures of the rest of the SDMX Web Service components, dispatched SDMX-ML, etc.

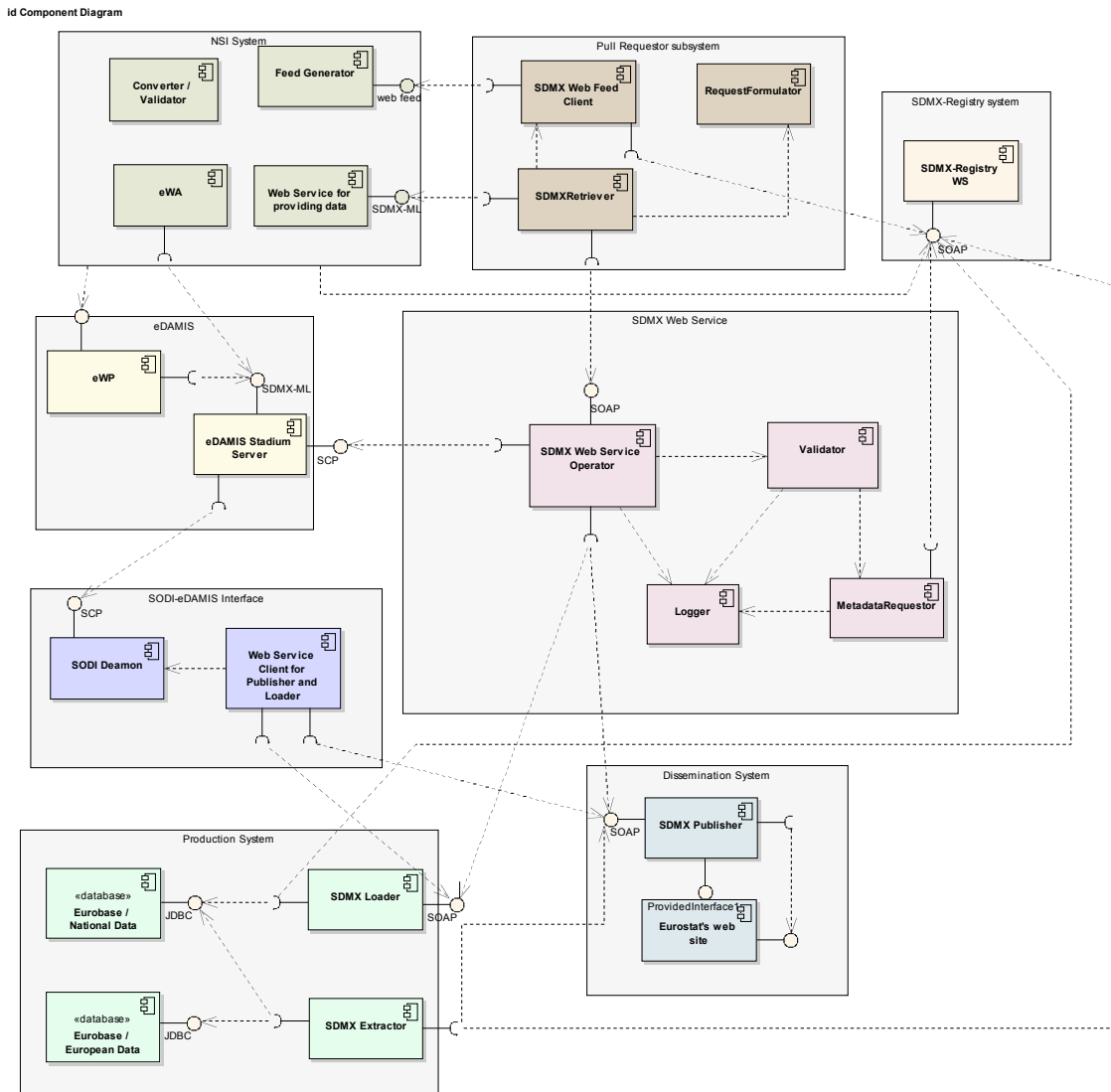
The SDMX Dissemination system includes the following components:

- SDMX Publisher, which is responsible for publishing the SDMX-ML to the Eurostat's web site by first transforming the XML to appropriate formats (e.g. HTML) using XSL. This component is itself a web service, and can be called either from the SDMX Web service exactly when incoming data are available, or at a later time by the SDMX extractor. SDMX Publisher accesses the SDMX Registry Web Service for acquiring structure metadata of the message to process.
- Eurostat's web site, where data will finally be disseminated.

The Production system includes the following components:

- Eurobase / National Data, which is the database where national data are stored
- Eurobase / European Data, which is the database where European data are stored
- The SDMX Loader, which is responsible for loading SDMX-ML files into the Eurobase. It will be implemented as web service. It accesses the SDMX Registry Web Service for acquiring structure metadata of the message to process.

- SDMX Extractor, which is responsible for extracting data from Eurobase and transforming them to SDMX-ML for further publishing.



**Figure 1. Component Diagram of the SODI processing environment**

The National Statistical Institute (NSI) System includes the following components:

- Feed generator, which is responsible for generating feeds, informing the interested parties (e.g. Eurostat) that there are data available over the Web. See “D3.1.a SODI Web Feed Feasibility Study” for more information.
- A Web service for providing the data. This web service is called by the Pull requestor. Call to the SDMXRegistry WS will be needed for the NSI Web service to have access to structural metadata.
- Optionally the eWA for pushing data to Eurostat.
- A conversion / validation module, used to convert dataset from CSV or GESMES to SDMX-ML and to validate it against corresponding XML Schemas before being sent (pushed) via eWA or eWP to eDAMIS stadium server. Call to the SDMXRegistry WS will be needed for the NSI Web service to have access to structural metadata.

The eDAMIS system includes:

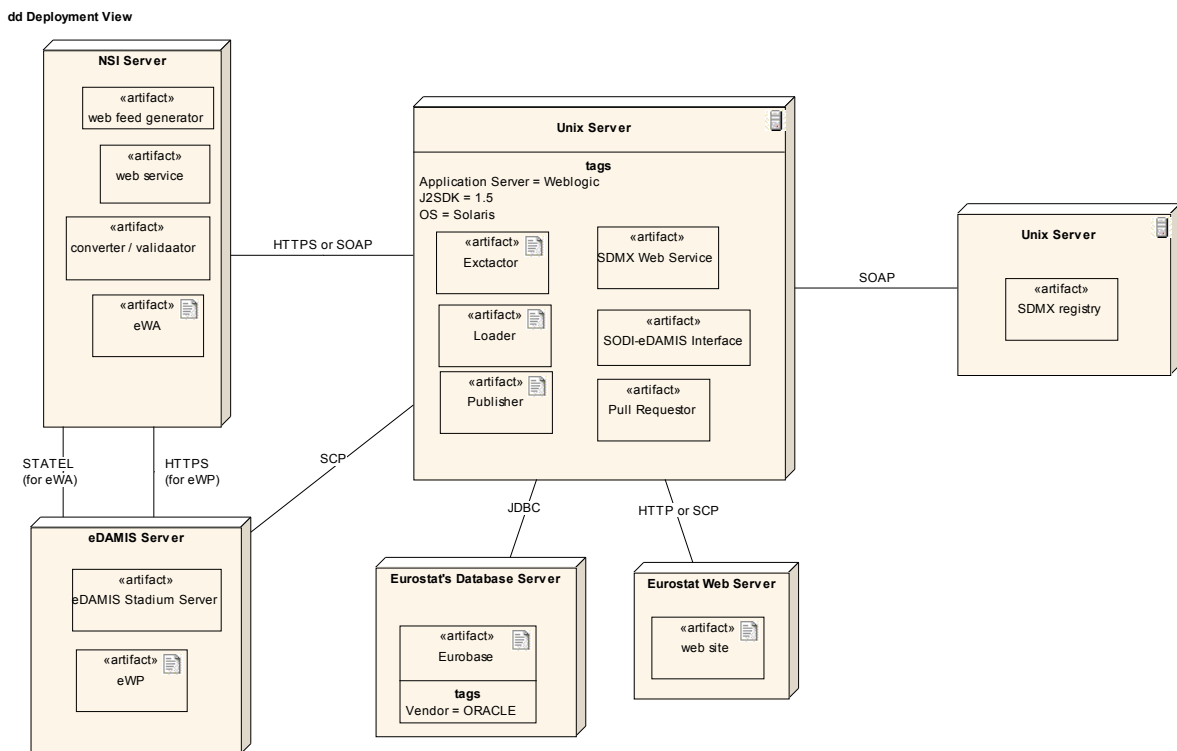
- The eDAMIS Stadium Server
- The eDAMIS Web Portal

The SODI-eDAMIS Interface module includes:

- The SODI Daemon, which is responsible for checking, periodically, the Loading / Publishing Delivery Directory for new data. This directory includes datasets that were sent using the PUSH mechanism.
- a web service client, which is called by the SODI Daemon when new data are to be published and loaded to Eurobase and is responsible delivering them to both the publisher and loader modules via SOAP.

### 2.3. Deployment of the SODI modules in Eurostat’s operating environment

The current installation diagram in the Data Center of the Commission is shown in Figure 2.



**Figure 2. Deployment Diagram of the SODI processing environment**

### 3. Conclusions

The objective of this paper was to present the minimum requirements in order to deploy a reference architecture for efficient, interoperable data exchange between statistical organizations using modern technology and well established standards.

This architecture has been already put in place and real data exchange in a pilot basis and with particular emphasis on the Principal European Economic Indicators (PEEIs) will occur in 2007, while further extension to other statistical domains (national accounts, purchasing pricing parities, demographic, agricultural, and education statistics) is planned for 2008.

## 4. Acknowledgments

Work presented in this paper has been performed under a framework contract from Eurostat (Statistical information technologies: Lot 1 – Implementation and support of standardised data formats for statistical data). We would like to thank Eurostat Unit B-3 personnel (J. Allen, L. Maqua, B.A Lindblad, G. Sindoni) for their guidance and kind cooperation throughout the execution of this contract.

## 5. References

- [1] IDABC - Content Interoperability Strategy, Working paper, September 2005
- [2] IDABC - European Interoperability Framework For Pan-European Government Services (version 1.0), EC 2004.
- [3] SDMX Technical standards, version 2.0 (November 2005 - [http://www.sdmx.org/standards/standards\\_package\\_2\\_0.aspx](http://www.sdmx.org/standards/standards_package_2_0.aspx))
- [4] Bass L., Clements P., Kazman R. Software Architecture in Practice, Addison Wesley, April 2003.
- [5] Bass L., Clements P., et al., Documenting Software Architectures, Addison Wesley, May 2003.
- [6] The Web Services-Interoperability Organization (WS-I), Basic Security Profile version 1.0, Working Group Draft (17-08-2006) available at <http://www.ws-i.org/Profiles/BasicSecurityProfile-1.0.html>
- [7] Rich Site Summary <http://www.rssboard.org/rss-0-9-1-netscape>
- [8] RDF Site Summary (RSS) 1.0 Official Specification <http://web.resource.org/rss/1.0/spec>
- [9] RSS 2.0 Specification <http://www.rssboard.org/rss-2-0>
- [10] AtomEnabled <http://www.atomenabled.org/>, <http://www.intertwingly.net/wiki/pie/Rss20AndAtom10Compared>
- [11] Atom Publishing Format and Protocol (atompub) <http://www.ietf.org/html.charters/atompub-charter.html> and The Atom Syndication Format <http://www.ietf.org/rfc/rfc4287.txt>