

# SODI - Extended interoperability architectures in statistical systems

Maria Glossiotti, Gregory E. Farmakis, Kyriakos A. Kassis,  
Spyros Liapis, Efstratios Nikoloutsos<sup>1</sup>

<sup>1</sup> Agilis SA Statistics & Informatics,  
Akadimias 96-100, Athens 10677, Greece,  
e-mail: {Maria.Glossiotti, 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 modern data exchange scenarios 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 collection and transmission processes.

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.

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 SODI (SDMX Open Data Interchange) architecture, materialises Eurostat's vision for EU-wide data sharing and has been deployed based on SDMX guidelines, web services and web feeds technologies. 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) using web feeds technology. The reference architecture has been graphically depicted using standard UML notation.

**Keywords:** Statistical Information Systems architecture, SDMX, data and metadata exchange and sharing

## 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 (i.e. the consumer retrieves the data), 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.

The following 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 statistical data and metadata in a variety of suitable statistical domains. The *SDMX Registry/Repository project* plays a central role in the data sharing exchange pattern in European statistics and allows organizations to maintain and publish structural statistical data and metadata in known formats such that interested third parties can discover these data (through the submission of properly defined SDMX queries) and interpret them accurately and correctly and within the shortest possible timescale. *SDMX reference metadata project* 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). Finally the *CVD Metadata Handler project* aims to develop an application that will enable the management and retrieval of structural and reference metadata. The application will be accessible from both human operators (domain managers in different Eurostat units) through a sophisticated user friendly GUI as well as other applications in production environment through web services interfaces.

## **2. Statistical Data and Metadata Exchange in the European Statistical System**

### **2.1 SDMX standard and technical specifications**

SDMX consists of technical and statistical standards and guidelines, together with an IT architecture and IT tools, to be used for the efficient exchange and sharing of statistical data and metadata. Seven European and international organisations (the Bank of International Settlements, the European Central Bank, Eurostat, the International Monetary Fund, the Organisation for Economic Co-operation and

Development, the United Nations Statistical Division and the World Bank) act as sponsors of SDMX.

In February 2008, at the 39th Session of the UN Statistical Commission, SDMX was recognized as the preferred standard for exchange and sharing of data and metadata in the global statistical community.

SDMX is built around the SDMX Information Model, which was designed to describe aggregated statistics of the kind typically transmitted from national statistical authorities to Eurostat, to ECB and to international organisations which collect statistics; the SDMX Information Model also fits the kinds of statistics which are disseminated at national or international level. The SDMX Information Model extends the GESMES/TS Information Model, and GESMES/TS - now renamed SDMX-EDI – that is now part of the SDMX standard for backwards compatibility. However, GESMES uses the EDIFACT syntax, whereas full exploitation of the possibilities of SDMX requires the XML syntax (SDMX-ML), which is supported by a much richer set of tools.

The power of SDMX though is not only the XML formatted messages for data and metadata, but the complete package of new artefacts and functionalities it offers. More specifically, SDMX offers, apart from the data/metadata message formatting, XML interface messages in order to achieve data/metadata exchange between organisations and other interested parties.

In the formatting portion of the standard, full specification and XML schemas are provided in order to successfully build structural metadata messages (DSDs, MSDs), data messages (datasets) and reference metadata messages (metadatasets). The SDMX standard offers multiple formatting possibilities especially for the data/metadatasets in order to serve different purposes and requirements and take advantage of the functionality XML offers. For example, there are formats for more convenient transmission (more compact and able to be sent partially – Compact format), for stricter syntax validation (using XML schemas – Utility format), for special non-time series messages (Cross-Sectional format) or for promoting interoperability (self explanatory messages – Generic format). The important aspect of the SDMX standard is that all the aforementioned messages are built upon the same Information Model, thus making conversions between them a trivial task.

Additional messages (SDMX Query messages) are specified within the SDMX standard for querying SDMX enabled web services for data and metadata. This kind of messages use logical operators ‘AND’ and ‘OR’ in order to combine the search terms (values of dimensions, attributes or other SDMX artefacts) in order to specify better the request. The SDMX Query message can be used as a means of interoperable querying/exchange of statistical data. Data Producers (e.g. NSIs at European level or national organisations at National level) can implement a Web Service able to respond to such a request (SDMX Query message) in order to serve data/metadata to interested Data Consumers (e.g. Eurostat).

In order to support data/metadata exchange, SDMX provides full specifications for a central (or distributed) structural/provisioning repository and dataset/metadataset registry (SDMX Registry). These specifications describe the content of the SDMX Registry as well as the interface messages used for communicating with this Registry. The purpose of this SDMX Registry, apart from providing its content to human users (via a graphical user interface), is to support applications that need SDMX structural metadata in order to participate in an SDMX data/metadata exchange. Thus, the SDMX Registry as described in the SDMX standard provides a Web Service interface that makes available metadata in SDMX-ML format.

To conclude, it should be stressed that SDMX is not yet another data/metadata formatting standard. SDMX is a full set of specifications for formatting, sharing and exchanging data/metadata in a statistical IT context.

## **2.2 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 a Web Service for retrieving dynamically data by the consumer. The data provider offers a web feed where the newly published data are registered, described by a standard SDMX query. The data consumer then fetches the data on his own initiative, after checking the feed for data he is interested in. In this case, more than one data consumers may be allowed to take the pieces of data needed by each one. 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.

## **2.3 SODI web feed technology and related 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.

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

#### 3.1. 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 rationale 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.

#### 3.2 Context of the SODI processing environment architecture

The SODI processing environment was developed in the context of the SODI project. Its aim is to provide a sharing method that will make national data and metadata available more quickly and more easily accessible. SODI supports transmission of both data and reference metadata.

In the following context diagram the SODI processing environment is shown as high-level component (system) that interacts with external to this system actors, which may be human users or other systems.

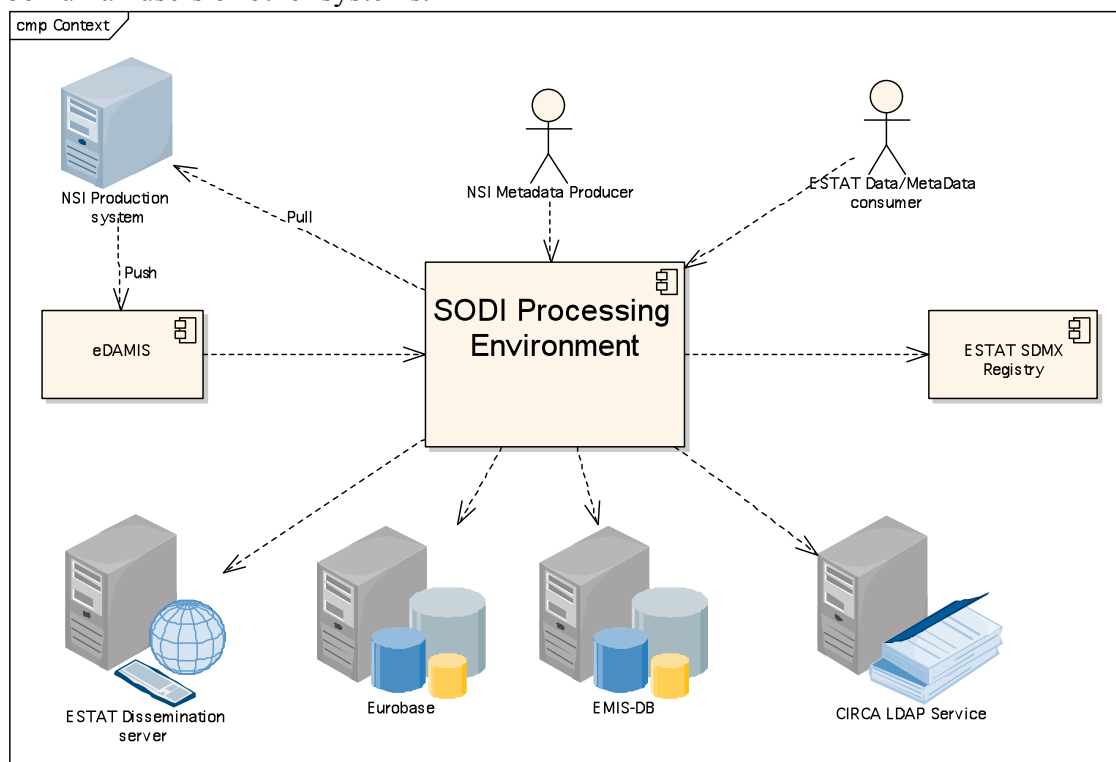


Figure 1. SODI processing environment – Context Diagram

The main external actors:

- The NSI production system where data compiled in the NSI and can be received by SODI either with ‘Pull’ or ‘Push’.
- The NSI Metadata Producer is human user responsible for compiling reference metadata.
- eDAMIS is an existing system in Eurostat that is used for receiving data from the Member States. It offers the Single Entry Point (SEP) for Member States to ‘Push’ data.
- ESTAT dissemination server is the place where data processed by SODI are sent for dissemination in a human readable format (“Publishing”).
- Eurobase is a production database where the received data (not reference metadata) are stored.
- EMIS is a production database where the received reference metadata are stored.
- ESTAT data/metadata consumer is human user (Eurostat personnel) that has the right to extract data from the production databases through the SODI processing environment.
- CIRCA LDAP Service is a web service offered by CIRCA available in the European Commission network that authenticates users.
- ESTAT SDMX Registry is the SDMX Registry of Eurostat that is responsible for storing and providing all structural/provisioning metadata used in the transmission process.

### **3.3 SODI component view**

The SODI component view gives an insight of the implementation of the system since it provides all the components with their interfaces and interactions thus formulating high-level specifications of the implemented components.

The component view is separated in three separate component diagrams for clarity:

- Transmission diagram representing components involved in the push and pull transmission
- Production diagram illustrating component in the “back-end” of the SODI processing environment related with dataset extraction and publication.
- Authorisation diagram illustrating the components realising the authentication and authorisation mechanism in SODI and the components using them.

In all diagrams the modules residing inside the SODI processing environment are put into a boundary in order to be differentiated from external systems.

SODI processing environment provides two transmission methods i.e. pull and push. Push method is based on the existing transmission scheme where member states prepare their data and send them directly to the Single Entry Point of Eurostat. Since eDAMIS is not capable of processing SDMX-ML datasets, in the push method, when it received datasets in SDMX-ML, it forwards them to the SODI processing environment.

The pull transmission is an innovative method in data exchange systems. In this scheme data provider doesn’t has to send data to the data consumer Eurostat. Data provider publishes the data produced in a feed. The feed contains the metadata of the data published e.g. dataflow, reporting period, data provider, location that reside. Eurostat is responsible to check the feed, for not processed entries that need to be acquired according to provisioning agreements stored in the SDMX Registry and retrieves them directly from data provider’s infrastructure (URL or a Web Service).

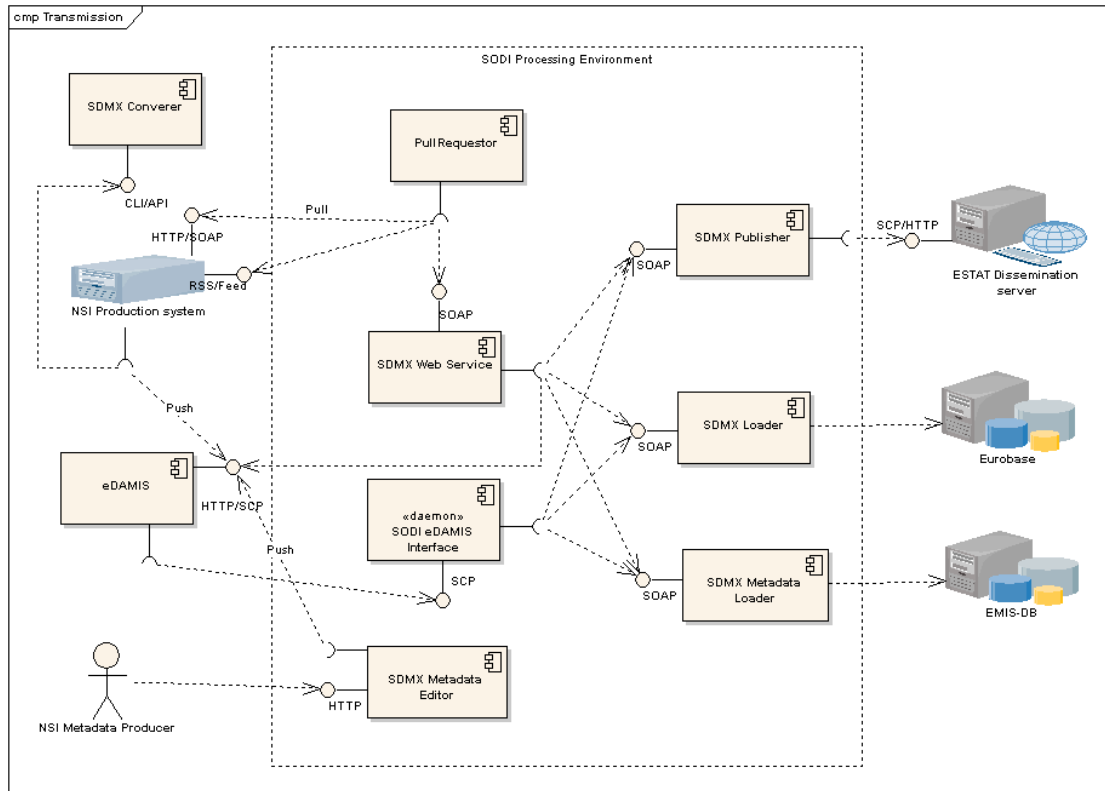


Figure 2. Transmission Component Diagram.

This new transmission scheme allows for the data provider to publish its data only once. Multiple data consumers can register to the feed and acquire data provider's data according to their needs. It is an open and extendable architecture where the data provider can be a data consumer for his data providers. The data provider can have unlimited data consumers without having to do something more than publishing data. Therefore this scheme facilitates data exchange utilising the SDMX standard.

The responsibilities of the SODI processing environment for the received data (either from push or pull) is to load them in the production databases (Eurobase, EMIS) for future processing, and disseminate them in a human readable format (HTML tabular layout) in Eurostat's dissemination server (e.g. New Cronos).

The Production component diagram (Figure 3) represents the components in the "back-end" of the SODI processing environment related with dataset extraction and publication. The ESTAT data/metadata Consumers query SDMX Extractor/QF and SDMX Metadata Extractor/QF respectively for acquiring data and reference metadata respectively from the production databases. These extractors provide the possibility to publish extracted data thus call the SDMX Publisher Web Service (SOAP).

The SDMX Publisher component includes a sub-component called Scheduler that is a daemon. This daemon allows for scheduling of automated publication from data (not reference metadata) extracted from SDMX Extractor/QF web service. The user SODI Admin configures the scheduling of the publication tasks.

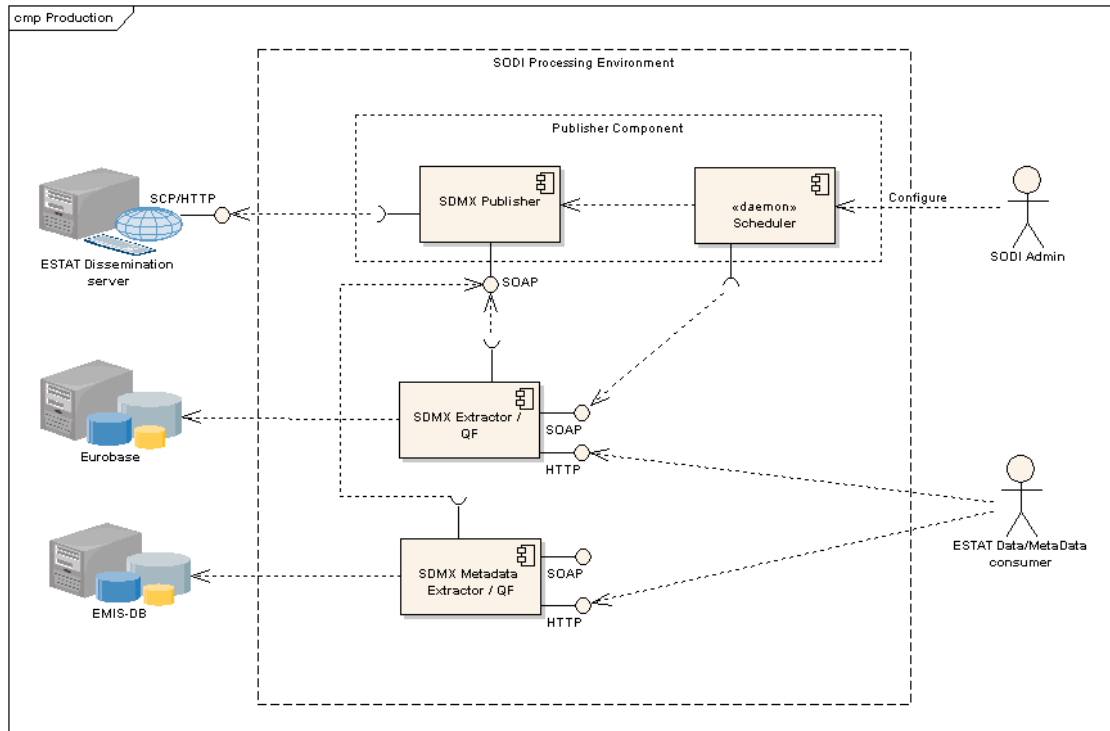


Figure 3. Production Component Diagram.

The Authorisation component diagram (Figure 4) illustrates the components realising the authentication and authorisation mechanism in SODI and the components using them. The Auth WS component performs the authentication and authorisation by using the CIRCA LDAP Service and User/Roles artefacts (stored as an XML file) respectively. For this purpose the Auth WS is called by the SDMX Metadata Editor and the Extractor/QF modules.

The SODI Admin and NSI Admin users manages these through the SODI User Management component the Users/Roles artefact.

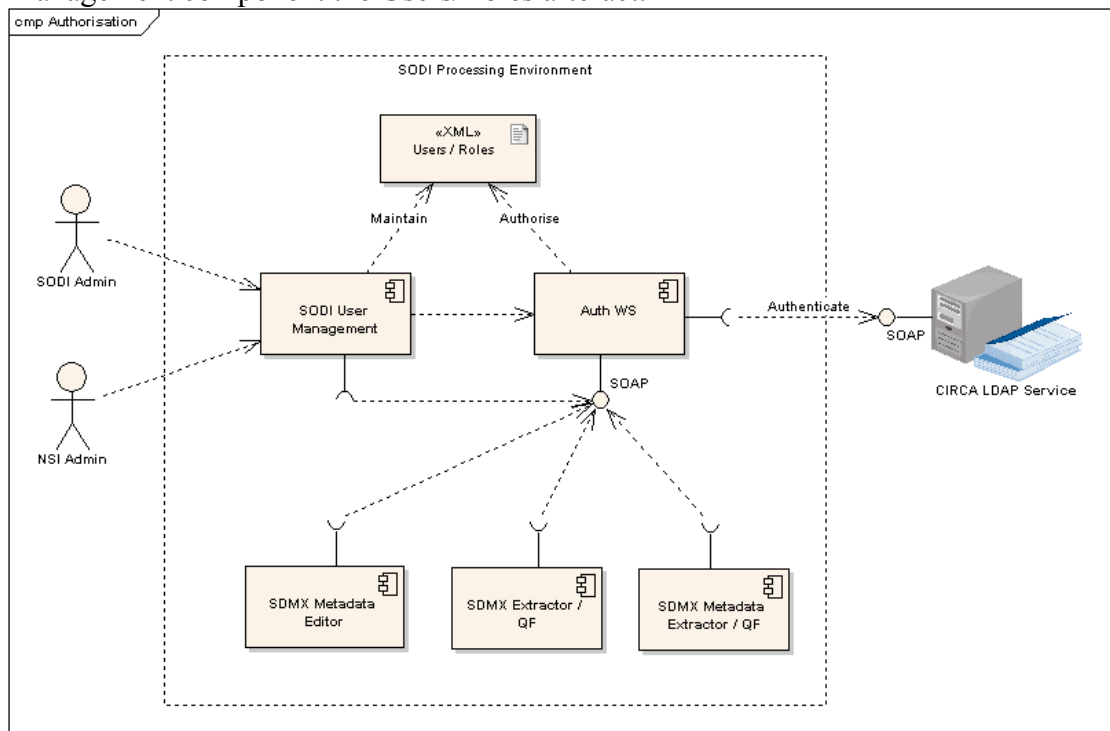


Figure 4. Authorisation Component Diagram.

### 3.4. Deployment of the SODI modules in Eurostat's operating environment

The SODI processing environment development has been based in Java and the Web Logic application server has been used for deploying its components in the Eurostat's infrastructure. The interfaces between its components are based on the SDMX-ML and all web services modules communicate through standard SOAP interfaces. Widely accepted as well as emerging standards and the use of Open Source Software been adopted in order to ensure interoperability and platform independent solutions. 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.

This section describes the physical network configuration on which the software is deployed and run. Each node represents either hardware or a software element, and the associations between nodes indicate a line of communication, while the name of the association indicates the protocol used for this communication. The deployment view of the SODI processing environment is depicted in Figure 5.

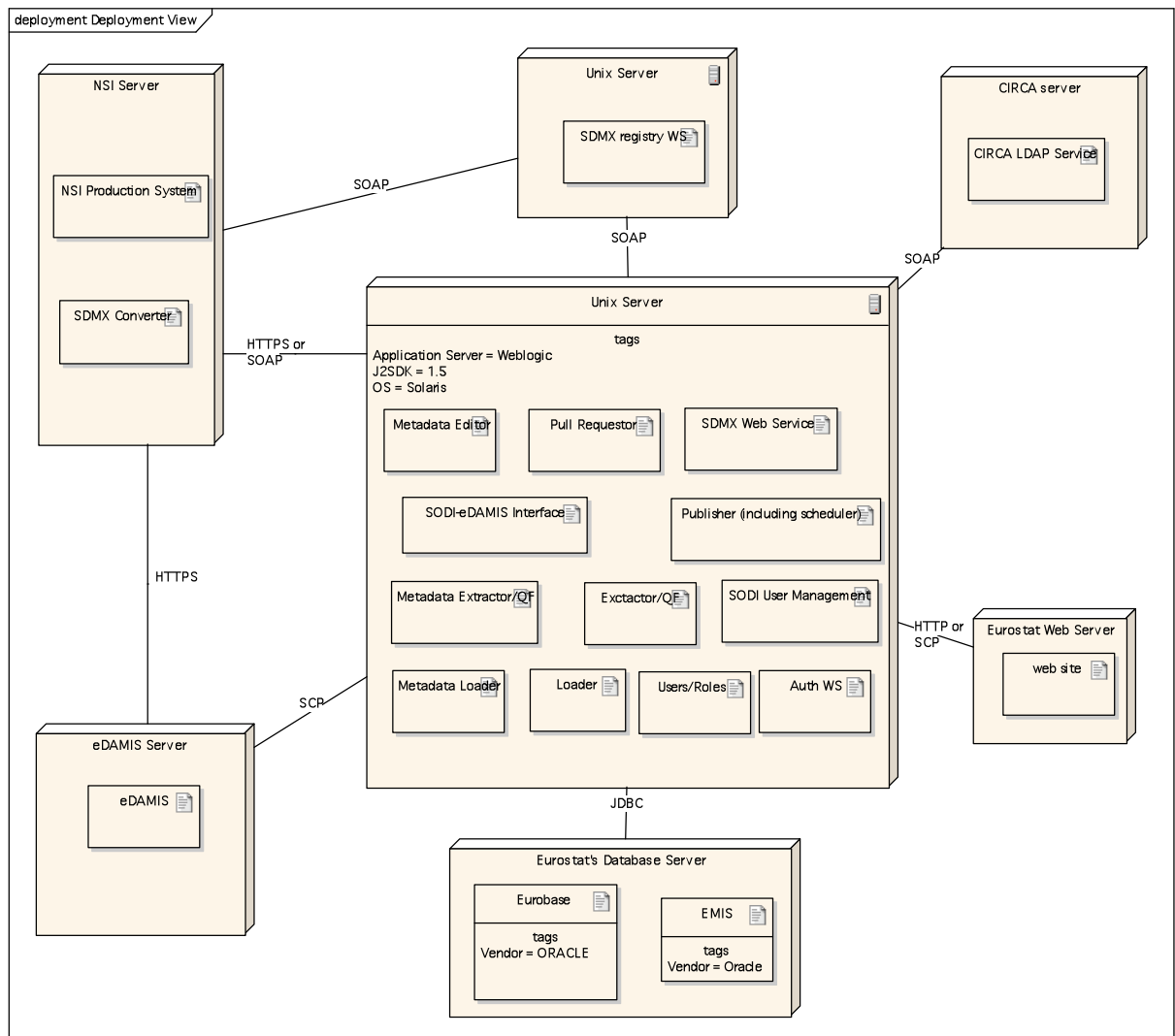


Figure 5. Deployment diagram of the SODI processing environment.

The following hardware elements can be identified in the following diagram:

- A Unix Server, where all web services will be deployed, as well as the Pull Requestor and the SODI-eDAMIS Interface modules. An application server installed

in the Unix Server will host all the web services and rest SODI processing environment modules, e.g. SDMX Web Service, Publisher, Loader and Extractor web services. Note that it is not necessary that all web services be deployed in the same physical machine. This deployment decision of the SODI processing environment has advantages and drawbacks. The advantage of deploying all the SODI processing environment in one server, is the use of common infrastructure thus maintenance of one server. However this could worsen the performance of the system when multiple messages arrive for processing. On the other hand scattering services into different servers would lead to a networking overhead. At the current phase SODI processing environment has been deployed in one server. If performance problems arise during usage then the deployment of a service to a different server will be considered. A possible service would be the SDMX Loader that needs to be close to the RDBMS.

- A Unix Server, where the SDMX registry resides and is available to the SODI processing environment via SOAP.
- Eurostats's Database Server, where the Eurobase and EMIS resides
- CIRCA server where the CIRCA LDAP service resides
- eDAMIS Server where the eDAMIS system is deployed
- Eurostat's Web Server, where the data will be finally disseminated
- the NSI server

#### **4. Conclusions**

A reference architecture for the efficient exchange, sharing and dissemination of statistical data and metadata has been presented. Apart from the obvious benefits of cost-effective extensibility, scalability and maintainability, this architecture ensures in an inherent way information quality, harmonisation of data and metadata through the use of widely used technical standards.

#### **References**

- IDABC (2005) Content Interoperability Strategy, Working paper  
IDABC (2004) European Interoperability Framework For Pan-European Government Services (version 1.0)  
SDMX consortium (2005) *SDMX Technical standards version 2.0*, SDMX.ORG website  
Bass L., Clements P., Kazman R. (2003) *Software Architecture in Practice*, Addison Wesley.  
Bass L., Clements P., et al., (2003) *Documenting Software Architectures*, Addison Wesley.