Official Statistics meets the Semantic Web:
How SDMX and RDF can live together

Raffaella Maria Aracri\(^1\), Stefano De Francisci, Andrea Pagano, Monica Scannapieco
{name.surname}@istat.it

Keywords: Semantic Web languages, Statistical data, Open Data, SDMX.

1. INTRODUCTION

In the Official Statistical (OS) domain, the issue of data interoperability has been present since decades: both National and International exchanges of data resulting from statistical processes are made possible only by adopting common metadata models and formats. Semantic Web technologies, and in particular the Linked Data initiative (see linkeddata.org) are more and more affirming as the principal mean for data interoperability, by permitting to create and interlink arbitrary volumes of structured data across the Web. In particular, the Linked Data initiative is made possible by the widespread adoption of Web standards for publishing data according to the Resource Description Framework (RDF) model. In this paper, we describe a project to integrate the SDMX [1] dissemination architecture with the Semantic Web Standards.

This work has been carried out by Istat (Italian National Institute of Statistics) within Eurostat grant “Horizontal and vertical integration: implementing technical and statistical standards in ESS”. In particular, the paper shows the design and implementation of an extension of the SDMX.Source.NET [2] to support the model and format translation of RDF. It is very much important to note that the translation step from SDMX to RDF is not only a “format” translation, but it involves a “model” translation, i.e. the specification of a set of rules to obtain RDF model’s constructs starting from SDMX model’s ones. Details of such a translation will be provided in the paper.

1.1. Background

In this Section, some background information on used technologies is described.

**RDF, RDF Schema, RDF QB:** The RDF (Resource Description Framework) [3] is a standard W3C data model that has features facilitating data integration even if the underlying schemas differ. All objects are represented by URIs and URIs are linked by a simple subject-predicate-object (triple) structure. RDF data can be represented with one of the following serialization formats: (i) RDF/XML, (ii) N-triples [4]; (iii) Notation3 [5]; (iv) Turtle [6]. RDF Schema is an RDF standard [7] for ontology definition. The specialization of RDF protocol to represent statistical data is RDF Data Cube Vocabulary (RDF-QB) [8]. RDF-QB is based on the SDMX Information Model.

**SDMX – Reference Infrastructure (RI) and SDMXSource.NET:** SDMX-RI [9] has been developed by Eurostat and consists of software and tools that facilitate the production of SDMX data and their exposure via Web Services technologies. SDMX-RI features include: (i) Serving SDMX v2.0 v2.1 data and structural metadata (SOAP/REST

---

\(^1\) Istat – Istituto Nazionale di Statistica
National Statistical Institutes Web Services: NSI Web service (WS) is the component of SDMX-RI that provides data to connected consumers in various formats, according to their requirements. Using the HTTP Content Negotiation [10] mechanism, the client specifies the desired format and version of the resource using the Accept HTTP header. The NSI Web Service of SDMX-RI can make data available by: (i) SDMX-ML Generic Data Format, version 2.1; (ii) SDMX-ML Structure Specific Data Format, version 2.1; (iii) SDMX-ML Structure Format, version 2.1. The specific objectives of the present work are to: (i) design a translation mechanism from SDMX data model to RDF Data Cube Vocabulary, (ii) design and develop the RDF extension of the NSI WS, that supports RDF-XML Data Format, version 2.1 and RDF-XML Structure Format, version 2.1.

2. METHODS

In the first part of the work, we describe the design solution to map elements of the SDMX data model to elements of the RDF-QB data model [11]. In the second part, we describe the design and the development of the software to extend the SdmxSource.NET to the new RDF-QB output.

2.1. Modeling: from SDMX to RDF Data Cube Vocabulary

Given the space constraints, we cannot provide the detail of the rules, but we highlight the SDMX-RDF mapping complexity. Let us consider the component REF_AREA stated in the key family section of the DSD related to “Separate collection indicators” of the Istat data on Environment and Energy Waste Figure 1. In general the component is more verbose when described in RDF-QB as it must be declared not only as a dimension of the Data Cube Vocabulary but also as an SDMX DimensionProperty and as an SDMX CodedProperty as it refers to a specific code list (in this case “territory” code list).

![Figure 1. SDMX-RDF mapping of REF_AREA component](image-url)
2.2. Adding structure and data format to SdmxSource.NET

In this section we show how we design and develop the code to extend the SdmxSource.NET to the new RDF-QB output. In doing this we leverage the existing SdmxSource.NET source code wherever possible and we develop specific add-ons that can be included as separate libraries in future SdmxSource.NET releases.

![Sequence diagrams for request RDF Structure (left) and Data (right)](image)

Figure 2: Sequence diagrams for request RDF Structure (left) and Data (right)

The sequence diagram in Fig.2 shows how the newly implemented “Writer” for structural metadata and data in RDF format works. As shown in the left picture when a client requests structural metadata in RDF format to the NSI WS a Controller module processes the request instantiating a Generic Writer containing the output format type and version (RDF V.2.1) the Structure Retrieve module connects to the Mapping Store DB and gets back the structural metadata. At this point the Controller specializes the Generic Writer into a RDF-Writer that returns the metadata into the required format to the client. Similar process is applied for data requests (right picture in Fig. 2).

3. RESULTS

To validate the results we leverage the “Sdmx-RI Test Client”. The Test Client is a tool to test the SDMX Reference Infrastructure building blocks and to expose/browse the dissemination environment of an NSI. First of all we deploy a new NSI WS on MS-IIS (Microsoft© Internet Information Services), then we use the Test Client (Fig. 3) to inquiry it and validate the RDF returned messages.
In the Test Client the user can check the “REST compatible query” option in order to build a REST query for the SDMX-ML 2.1 version. Then the user clicks the “Execute SDMX Query with REST” button, to display the “REST Service Test” window: The application provides the REST query, starting from the selected DataFlow of the main form, and adding the http Header “application/rdf+xml”. Pressing “Send REST Request” button the Test Client invokes the NSI WS available at the “Base Endpoint” and returns the DataFlow in RDF format.

4. CONCLUSIONS

The paper describes the solution implemented by Istat to integrate RDF-based technologies with the SDMX-based dissemination architecture promoted by Eurostat.

The major outcome of this work is the proof that the two world can live together: indeed, a direct translation from the SDMX NSI WS is possible. On the other side, we highlight that such a solution can be also integrated with a Linked Open Data dissemination channel based on a SPARQL Endpoint. Indeed, so far the URIs produced by the SDMX translation steps are not actual URIs: they need a “deployment” on a SPARQL endpoint. If such data were actually “deployed”, instead, the result of the SDMX translation could pass from the document/dataset state to the one of singly deployed data values.
REFERENCES