A European Strategy for Linked Open Statistics

Luca Gramaglia[1], Christine Kormann-Fromageau[1], Danny Delcambre[1], Jean-Marc Museux[1] and Marta Nagy-Rothengass[1]

[1] Eurostat, Rue Alphonse Weicker 5, Luxembourg

Abstract. Over the past decade, the European Union has promoted the development and deployment of Linked Open Data technologies through the financing of several projects in the framework of the FP7 and Horizon 2020 research and innovation programs. In recent years, Eurostat, the statistical office of the European Union, has started to investigate the application of these technologies in the statistical ecosystem to enable access to new data sources and to improve the cross-domain discoverability and interoperability of official statistical data. This paper provides an overview of the different actions Eurostat is taking in the context of the ESS Vision 2020 modernization program to leverage the statistical metadata assets of the European Statistical System (i.e. concepts, codes, nomenclatures) through the use of semantic technologies. In particular, the paper presents the work launched to build and sustain an LOD community of practice among European statisticians and the ongoing efforts to deepen the links between well-established standards for the description of statistical data and the Linked Open Data world. The paper concludes by outlining some ideas and principles for a reference linked open data architecture in the European Statistical System.

Keywords: Linked Open Data, European Statistical System, Metadata standards

1 The role of standards in the European Statistical System – a brief overview

The European Statistical System (ESS) is the partnership between Eurostat, the statistical authority of the European Union, and the national statistical institutes (NSIs) and other national authorities responsible in each Member State for the development, production and dissemination of European statistics. This Partnership also extends countries belonging to the European Economic Area (EEA) and European Free Trade Association (EFTA).

Eurostat and the ESS have a long history of promoting harmonization and standardization in the field of statistics. This is not surprising, as the process of producing comparable European statistics from several different sources and with the involvement of a large number of stakeholders naturally makes it almost necessary to share a common language for the collection, processing and dissemination of data. The ESS

CEUR-WS.org/Vol-2317/article-11.pdf

has adopted its own definition of a standard that is based on the International Organization for Standardization (ISO) definition and reads as follows:

> *"A normative document, established by consensus among ESS members and approved by a recognized body according to the procedure of ESS standardization, that provides for common and repeated use by several actors in the ESS. This can be rules, guidelines or characteristics for the development, production and dissemination of European Statistics, aimed at the achievement of the optimum degree of order in the context of the implementation of the mission and vision of the ESS". [1]*

Standards cover all aspects of the statistical production process and are present at all stages of the statistical production process. The paragraphs provide a non-exhaustive description of the main areas where the ESS has over the years launched standardization initiatives.

- **Standards for data and metadata exchange**: Eurostat is one of the Sponsors of the SDMX initiative. SDMX[1] (which stands for Statistical Data and Metadata eXchange) is a standard which supports the automated exchange of statistical information. It does so by providing an Information Model for the description of data and metadata assets, statistical guidelines for the modelling of statistical data, as well as an IT architecture and tools. Moreover, using SDMX as a basis, Eurostat and the ESS have developed vocabularies for the documentation of the statistical production process. One such vocabulary, SIMS[2] (Single Integrated Metadata System), has established itself as the standard for quality and metadata reporting in the ESS.

- **Harmonized methods:** This heading encompasses the methodological resources used by the ESS to produce harmonized data, such as classifications and methodological manuals. As a general rule, the ESS always tries to align with global standards (e.g. the International Standard Classification of Occupations (ISCO) developed by the International Labour Organization) and develops its own standards only if it needs to collect more detailed information. Even in these cases, attempts are made to establish mappings between the global and regional standards. These efforts have given rise to an integrated network of statistical classifications at global, regional and national levels. Such an integrated system facilitates comparisons of data sets based on different classification systems.

- **IT tools:** In recent years, the ESS has expressed increasing interest in the creation of standards for the sharing and reuse of IT tools. The ESS has actively participated in the creation of standards such as the Common Statisti-

[1] Official website of the SDMX initiative: https://sdmx.org/
[2] Description of the SIMS 2.0 standard: http://ec.europa.eu/eurostat/documents/64157/4373903/SIMS-2-0-Revised-standards-November-2015-ESSC-final.pdf/47c0b80d-0e19-4777-8f9e-28f89f82ce18

cal Production Architecture[3] (CSPA), whose stated aim is to help create interoperable tools that can be shared within and between statistical organizations.

2 The opportunities of Linked Open Data for European statistics

Against the background of this long-standing commitment to standardization in the ESS, semantic technologies have created new opportunities and expectations with respect to metadata and data discovery and analysis. In order to retain its position as leading providers of statistical information and to contribute to the modernization of public administration by engaging more with citizen and businesses, National Statistical Institutes (NSIs) in charge of collecting and disseminating official statistics in the EU must upgrade their data accessibility, discovery and analytics capabilities and follow a proactive approach in meeting the needs of their key users.

Firstly, end users of EU statistics are looking for better discoverability of open/public data. Semantic technologies can help improve discoverability and use of official statistics data by leveraging the rich European Statistical System (ESS) statistical metadata assets (concepts, codes, nomenclatures …). Statistical metadata needs to be made machine-readable, discoverable and active in data querying and data integration jobs. This will contribute, on one side, to easing access to EU statistics and, on the other side, create the conditions for a wider semantic interoperability and accessibility of statistical data produced by public administrations in general. External data brokers (collecting and reusing data) will also benefit from well-established standard semantic metadata encouraging them to align with these standards and thus boosting the (re)use of statistics published by government agencies.

Secondly, policy analysts in charge of the design and evaluation of government policies have to embrace the complexity of societal and economic changes. This requires analyzing increasingly detailed official statistics and enriching them with available external information. Currently the official data are publicly available but their usage is rather tedious as they require quite some technical skills to access and combine them and a good knowledge of the concepts involved. Further action is needed to provide an intuitive and integrated data analytics workbench and metadata driven services to explore, combine and analyze a broad variety of cross-border data sources, leveraging the wealth of official statistics and the richness of the metadata assets compiled in the ESS.

[3] Official CSPA specifications:
https://statswiki.unece.org/display/CSPA/Common+Statistical+Production+Architecture

3 First implementation actions

In recognition of the potential offered by semantic technologies, Eurostat, the statistical office of the European Union, has started investigating the application of these technologies for EU statistics. In line with the subsidiarity principle enshrined in the EU founding treaties and in Regulation on EU statistics, interoperability is the preferred mode of collaboration of NSIs and Eurostat to produce and disseminate EU statistics.

The main vehicle for these first steps towards Linked Open Statistics in the European Statistical System is the DIGICOM project[4], which is one of the flagship projects of the ESS Vision 2020 modernization program. In the context of this project, a number of actions were set in motion: a study was carried out in 2016 to provide an overview of LOD experiments in the European Statistical System, a joint framework for action in the field of LOD was adopted and a set of collaborative cross-country pilots were launched.

3.1 2016 study on LOD activities and use cases

A study [2] was conducted by Price Waterhouse Coopers on practices in NSIs which have experimented with Linked Open Data. The study looked in particular at the state of play in the National Statistical Institutes of France (INSEE), Italy (ISTAT), Ireland (CSO Ireland), the United Kingdom (ONS) and Switzerland (Federal Statistical Office of Switzerland). It was complemented by an ESS Workshop on LOD held on 18 and 19 January 2017 in Malta. The study and the workshop led to a shared understanding of the key benefits of LOD for official statistics. These include more flexible means of data dissemination, enhanced data exploration between datasets and the ability to link with other sources (e.g. within a national statistical system) while keeping the information on data provenance. Indirect benefits include fostering internal coherence of data and metadata, reinforcing the role of NSI as standard setters and stimulating partnerships.

However, the study highlighted that LOD is an area in which NSIs are still largely experimenting. Considering scarce resources and the need for specific skills, the study concluded that these experiments would benefit from a joint approach in the ESS. The study identified some key priorities, such as: capacity building at NSI- and ESS-level, including multidisciplinary teams (IT, dissemination, content and classifications), joint governance for LOD, evaluation of existing technologies, development of a community of practice in the ESS, and cooperation with experiments outside the EU.

[4] A detailed overview of the DIGICOM project can be found at the following link: http://ec.europa.eu/eurostat/web/ess/digicom

3.2 Joint Framework for action (2017-2018)

Based on the results of the Malta workshop and on the findings of the study, a joint framework for action for the ESS in the field of LOD was developed and endorsed by the ESS members in February 2017 [3]. The framework for action takes a holistic approach to the development of statistical LOD and covers five complementary dimensions: strategy and policy; people and capacities; data and metadata; technology and infrastructure; and governance.

The activities foreseen by the framework include:

- Common investment in capacity-building and knowledge sharing across ESS members. The objective is to assist National Statistical Institutes in the in-house development of core LOD skills and expertise. This has led to the inclusion of a training course on Linked Open Data as part of the European Statistical Training Program.

- Definition and implementation of standards. Common vocabularies and ontologies for describing the statistical data and metadata assets must be adopted. The Open data community has already relevant standards for such metadata (STAT DCAT for catalogue, Data Cube Vocabulary for describing statistical datasets, SKOS for modelling metadata concepts). The ambition is to close the gap between these standards and the corresponding standards in use in the statistical community (e.g. SDMX) and, when needed, to develop new ontologies to express the already rich set of statistical data and metadata assets (e.g. code lists and classifications) and to link them to key other resources on the web (e.g. DBpedia).

- Selection and development of common, reusable IT tools. In order to create a highest take-up within the ESS a common approach to access mechanisms such as APIs, SPARQL endpoints and direct URI resolution is required. Eurostat and NSIs should work together on testing and selecting tools. Moreover, special attention will be given to the development of intuitive interfaces for users to query European statistics. The objective is to overcome the relative difficulty of exploiting the power of SPARQL end points or graph databases and to enrich the experience of non-specialist users.

3.3 ESSnet on Linked Open Statistics

One of the main principles of the joint framework for action is that the various building blocks it foresees should be created through practical experimentation, and in particular through collaborative cross-border projects.

In order to jumpstart the realization of proofs of concept that can identify and meter the potential of Linked Open Data for the dissemination of EU statistics, at the

end of 2017 Eurostat launched an ESSnet[5] involving NSI Bulgaria, INSEE, ISTAT, CSO Ireland and academic partners. An "ESSnet" is a multi-beneficiary grant given to a network of several ESS organizations and which is aimed at providing results that will be beneficial to the whole ESS. ESSnet projects are one of the primary instruments through which Eurostat fosters harmonization, collaboration and the sharing of best practices within the ESS.

The ESSnet will deliver two main outputs, pilots and a collaboration platform. The pilots will focus on linking data and metadata at local, national and international level as well as on linking statistical data and metadata with non-statistical web-based data – for example, wiki, georeferenced data, etc. The pilots will demonstrate the benefits of this approach for users, but also help to assess risks and costs. The ESSnet also aims to stimulate cooperation and provide capacity building for the whole ESS. For this purpose, it will set up a platform for collaborative work among ESS experts on LOD topics and will provide training material and webinars, guidance and toolkit on how to design and build an LOD portal and to use it for standards.

For its part, Eurostat is also developing pilots in close coordination with the ESSnet. The pilots focus on exploring how Eurostat data and metadata can be modelled as linked open data. One point of particular interest is the exploration of how the existing standards and vocabularies used to describe data, metadata and classifications (e.g. the SDMX and SIMS standards described in chapter 1) can be expressed in a Linked Open Data environment.

4 Towards a reference architecture for Linked Open Statistics in the ESS

The knowledge acquired via the experiences and pilot projects outlined above will allow the ESS to identify common principles and best practices for the use of Linked Open Data in the context of European Statistics. The intention is that these principles and best practices should form the basis for the creation of a reference architecture for Linked Open Statistics in the ESS.

The purpose of a reference architecture is to provide a template for statistical organizations in the development of their own LOD capabilities and infrastructure. It shows what capabilities organizations should acquire and how they should organize and structure systems to disseminate statistical information based on LOD concepts/standard. It does not have a binding character but helps to spread common views and definitions and to foster collaboration between partners. It can provide a basis for mapping ESS activities/Proofs of Concept, for kick-starting LOD activities

[5] Homepage of the Linked Open Statistics ESSnet project: https://ec.europa.eu/eurostat/cros/content/linked-open-statistics_en

in Member States and for the identification of shared/common building blocks/solutions.

The reference architecture under development will identify the major use cases and contain a list of definitions, required capabilities, recommended standards, building blocks, principles and governance recommendations for statistical organizations who would like to invest in Linked Open Data and maintain a high degree of interoperability with other ESS partners.

References

1. Nadezda Fursova, Jurate Petrauskiene, Vocabulary of standardization related concepts, https://ec.europa.eu/eurostat/cros/system/files/ESSnet%20on%20Standardisation_SGA_2_WP1_Del1.pdf, last accessed 2018/06/15
2. Price Waterhouse Cooper, Study on LOD strategies and use cases in the ESS, https://ec.europa.eu/eurostat/cros/content/task-2-use-cases-and-analysis_en, last accessed 2018/06/15
3. A joint framework for action on linked open data at ESS level, https://ec.europa.eu/eurostat/cros/content/item-2-draft-ess-strategy-linked-open-data_en, last accessed 2018/06/05