UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS
ModernStats World Workshop 2022
27-29 June 2022, Belgrade, Serbia

Towards the dissemination of statistical classifications as Linked Open Data

Christine Laaboudi-Spoiden,
Eurostat, European Commission, Luxembourg, Christine.laaboudi@ec.europa.eu

Abstract
Since the beginning of the millennium, RAMON, Eurostat's Metadata Server has been playing the role of authoritative repository for the nomenclatures and classifications maintained by Eurostat. The classifications and their supplementary information (correspondence tables, translations, list of official classifications decisions) are available for download under an open license (CC BY 4.0) in various open or proprietary formats but are not harmonised in their structure. This paper describes the recent initiative achieve standardisation through the conversion of Eurostat multilingual statistical classifications and their correspondences currently residing in RAMON to 1) Artefacts in the Euro SDMX Registry (the Registry) and 2) as Linked Open Data, with the aim of maximising data Findability, Accessibility, Interoperability and Reusability according to the FAIR Principles. In a first step, a conversion tool that transformed the Ramon classifications into artefacts in the Registry was developed. In a second step, XKOS, an ontology for modelling statistical classifications, developed by DDI (Data Documentation Initiative) was applied on the artefacts in the Registry to convert them into Linked Open Data. This paper details the new opportunities offered by these two future-oriented ways of disseminating statistical classifications.

Keywords: statistical classification; Linked Open Data; interoperability; SDMX annotations; XKOS

Acknowledgement: Thanks to Martin Karlberg (Eurostat) and Luca Gramaglia (Eurostat) for extensively reviewing previous versions of the manuscript.

1. Dissemination of European Statistical Classifications

A statistical classification or nomenclature is an exhaustive and structured set of mutually exclusive categories used to standardise concepts for the collection, compilation and dissemination of statistical data (Hoffman, 1999).

Eurostat has a high level of knowledge and experience in the development of classifications and is the custodian of a number of sectoral and transversal European statistical classifications, and responsible for covering the European dimension for international statistical classifications linked to the European ones under its responsibility (NACE, CPA, PRODCOM, and Combined Nomenclature). Each statistical classifications typically exists in a statistical ecosystem, where it is normally interlinked with other classifications – either structurally, or by means of correspondence tables.

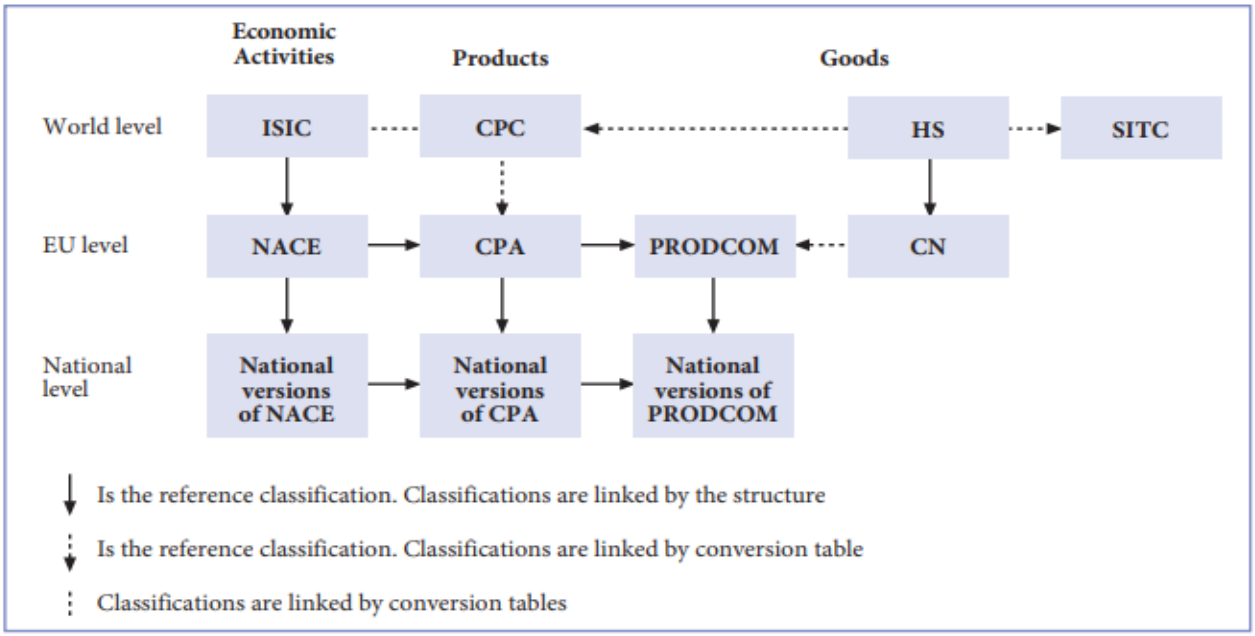

Figure 1: Integrated System of Statistical Classifications (Eurostat, 2017)

Statistical classifications and their correspondence tables used for European Statistics are disseminated via RAMON[1], the Eurostat **R**eference **A**nd **M**anagement **o**f **N**omenclatures platform. Since the early 2000s, RAMON has been hosting the Classifications developed by Eurostat as well as some external classifications not managed by Eurostat, but reused for collecting statistics or for developing correspondence tables to their reference classifications. In RAMON, the representation of the classifications is following the basic principles for Standard Statistical Classification*s* (Hoffman, 1999) (Hancock, 2013) (ESS, 2019) although this representation is not formalised according to commonly used community standards. One example of this is that there are properties that do not have the same designation across the classifications or the additional files available in various open (XMLS, CSV) or proprietary (MDB, XLS) formats.

The classifications are human-readable, accessible by browsing their structure via a succession of HTML pages. A search box enables to search by code item, returning the structure of the code item, with the contextual details being available in a second web page. Finally, the general description of the classification is available in a separate HTML web page.

Classifications may have a number of correspondence tables available for download or via a link to a third-party website. Classifications and their correspondence tables are presented in different sections of RAMON and are not interoperable. For instance, a classification item offers no direct links to its correspondences and a correspondence table lists the codes (sources, targets) but does not display their semantic context. Moreover, in some cases, correspondences are based on the official code of a classification, while the same classification may be presented in RAMON organised by a different sorting key.

As the platform hosting RAMON has to be phased out by the end of 2022, we seized the opportunity to reflect on what and how we want to disseminate our classifications in order to make them widely reusable by the in the European Statistical System (ESS). In this context, we have followed two approaches:

1) Converting all the statistical classifications hosted in RAMON in SDMX/XML and storing them into the SDMX Euro Registry, as presented in Section 2,

2) Converting the statistical classifications developed by Eurostat into RDF and exposing them as Linked Open Data as presented in Section 3.

In Section 4, we present the advantages of these two future-oriented ways of disseminating statistical classifications in relation to RAMON, as well as their relative advantages in relation to each other. Finally,

[1] RAMON website: https://ec.europa.eu/eurostat/ramon

Section 5 sets out future possible improvements concerning the dissemination and reuse of statistical classifications.

2. Statistical classifications in SDMX

Eurostat has a high level of knowledge and experience in data modelling with SDMX[2] (Statistical Data and Metadata eXchange), an international standard whose aim is to make it easier to exchange and share statistical data and metadata, supported and owned by seven international organisations, including Eurostat. The SDMX standard (SDMX, 2021) defines several objects or artefacts that can be used to model and automate data and metadata exchange (the Information Model), as well as specifications of Registries that allow for the storage and retrieval of these objects. The Euro SDMX Registry[3] is Eurostat's implementation of an SDMX Registry and is used to store all SDMX artefacts needed to share and interpret statistical data and metadata exchanged in the European Statistical System (ESS).

In this first approach, we developed a script for converting the properties of each statistical classification from RAMON into an SDMX Code List (SDMX, 2018), a container of codes that associates an identifier (the Classification Sorting Key) with a name (the Classification Code), a description (the Classification Label) and a Parent Code setting up the structure. From a data modelling perspective, the Classification Code is stored in the Name property, since in some classifications there is a unique Key different from the classification coding, preventing the Key property from being used.

In the Euro SDMX Registry, the 'Classification' Code Lists are prefixed by CLS_ followed by the identifier of the classification (CLS_CPA_2.1 for the CPA version 2.1) in order to differentiate them from the default SDMX Code List Artefacts. The content is accessible as a flat list, ordered by ID or Name, or in a hierarchy following the structure defined by the Parent Code. Each classification is searchable by Code or Expression part of the Description. The full classification is downloadable in SDMX/XML format directly from the Registry or can be retrieved via a query to the SDMX-Registry Rest API (Eurostat, 2020).

In addition to the basic structural elements (identifier, code, labels and parent code), statistical classifications may include additional components such as:

- *Explanatory Notes,* outlining the conceptual basis for interpreting the classification or explains what a classification item is about and how it should be used for categorising statistical units;
- *Case law* (classification opinions, decisions, ruling), incorporated in administrative or legal interpretations and providing rules about how to classify new situations or responses until the next revision of a classification takes place;
- *Classification levels,* aggregating the classifications items by levels, each level following a code patterns reflecting the level of details;
- *Unit of measure,* referencing the measurement unit for this heading.

We represented these components in *SDMX Annotations*, a flexible extension mechanism allowing organisation-specific metadata to be added to a SDMX Structural Artefact with the aim of improving machine interactions by proposing a controlled vocabulary for the type property of the SDMX Annotation and a recommended usage (SDMX, 2021). Finally, all the correspondence tables, downloadable from Ramon in XML, have been converted into SDMX Structure Set Artefacts, mapping the SDMX Code from a Source Code List to a Target Code List, both stored in the Registry. The context of the source and target Code List is available in the downloaded XML file, which includes the mapping as well as the complete Source and Target Code Lists.

[2] https://sdmx.org/

[3] https://webgate.ec.europa.eu/sdmxregistry/

3. Statistical classifications from a Linked Open Data perspective

We based our second approach on the SDMX terminology introduced in Section 2. (Information Model, SDMX Annotations) that we reinterpreted in the context of Linked Open Data (LOD). While there is no single formal RDF ontology that provides a full one-to-one equivalent for the SDMX Information Model, the most relevant ontology that can cover the modelling of statistical classifications is XKOS (Extended Knowledge Organization System) (DDI, 2019), an SKOS (Simple Knowledge Organization System) (W3C, 2019) extension for representing statistical classifications developed by DDI (Data Documentation Initiative) that meet domain-relevant community standards and best practices. In relation to the SDMX artefacts, XKOS has the added advantage of being is compliant with the semantic web technologies and allowing a richer description of the resources as LOD. In this second stage, we only considered the classifications developed and maintained by Eurostat as well as their correspondence tables, provided that the target classifications are available in RDF.

3.1 Crosswalk between SDMX and XKOS

For the storage and dissemination of our Classifications in RDF, we used a suite of corporate semantic tools offered by the Publications Office of the EU. For the purpose of Linked Data Modelling, we used Vocbench[4], a RDF web-based multilingual collaborative platform for managing controlled vocabularies and generic RDF datasets. Based on the mapping between SDMX and XKOS properties, we achieved the transformation of Eurostat statistical classifications via an integrated tool that enables to convert structured data into RDF triples and to inject it directly in the RDF graph database (a Triplestore).

Table 1: Crosswalk between SDMX and XKOS objects

Concept	**SDMX Artefact**	**XKOS Class**
Statistical Classification	Code List	skos:ConceptScheme
Classification item	Code Item	skos:Concept
Correspondence Table	Structure Set	xkos:Correspondence
Code Mapping	Structure Map	xkos:ConceptAssociation
Classification Level	N/A	xkos:ClassificationLevel

3.2 Dissemination

Once uploaded in Vocbench, Eurostat statistical classifications are going to be officially disseminated in the EU Vocabularies website[5], allowing users to search, browse the content or download the Classifications and their Correspondence Tables in different formats for reuse (SDMX/XML, RDF). The statistical classifications are richly described with a multitude of accurate and relevant attributes, released with a clear and accessible data usage license (reuse policy of the European Commission)[6] , and will be aggregated under the 'Eurostat Statistical Classification Business Collection' in EU Vocabularies. The content of Eurostat statistical classifications is accessible in ShowVoc[7], an online platform[8] built to facilitate the visualisation of RDF-based vocabularies maintained in Vocbench. One of the advantage of ShowVoc is the extra feature that allows browsing or searching the correspondence tables.

---

[4] Vocbench https://op.europa.eu/en/web/eu-vocabularies/vocbench

[5] EU Vocabularies https://op.europa.eu/en/web/eu-vocabularies

[6] The reuse policy of the European Commission is implemented by a Decision of 12 December 2011.

[7] Access to ShowVoc https://showvoc.op.europa.eu/#/home

[8] ShowVoc: Online tool description https://op.europa.eu/en/web/eu-vocabularies/showvoc

4. FAIR principles applied to statistical classifications

In this Section, we present the advantages of these two future-oriented ways of disseminating statistical classifications in relation to RAMON. We emphasise how we applied to the four FAIR principles (GO FAIR, s.d.) (Cox SJD, 2021) to European statistical classification in order to enhance their **F***indability* and **A***ccessibility*, make them **I***nteroperable* with other referenced and derived statistical classifications and finally optimise their **R***euse* inside the European Statistical System (ESS).

4.1 Findability and Reuse

Eurostat classifications will be defined in the domain 'data.europa.eu', with one namespace by Classification Family (for example, http://data.europa.eu/xsp/ will be identifying the Combined Nomenclature). In the ongoing implementation, a persistent URI (Unique Resource Identifier) will be assigned to each instance of the following resource types:

- Classification (or Classification item forming part of a Classification),
- Correspondence Table (or Concept association part of a Correspondence Table),
- Classification Level.

Eurostat statistical classifications are described with rich metadata providing an accurate description of their content, ensuring the users of being well-informed about the context as well as references to other resources such as:

- The legal basis of the classification that refers to the European Legislation Identifier (ELI) in EUR-Lex, the EU Legislation Database;
- The URI(s) of the successor or predecessor classifications;
- The number of Levels and a reference to a description of the Classification Levels;
- The languages, custodian(s) or the coverage expressed by reusing standardised resources from Corporate Code Lists.

The description of a Classification item includes the structural elements, multilingual labels, Explanatory Notes as well as, for the Combined Nomenclature, qualified references to the relevant classification implementation decisions published in the Official Journal of the European Union, based on the European Legislation Identifier (ELI).

The description of a Classification level provides details about the name of the level, the notation pattern, and list all the classification items aggregated under the level. The Classification Level is typically a construct of XKOS, represented in a textual Annotation in SDMX.

4.2 Accessibility of statistical classifications

In addition to the SDMX XML format offer in the Euro SDMX Registry, the datasets (Classifications, Correspondence Table) can be downloaded in SKOS/RDF from the EU Vocabularies website or via a SPARQL endpoint that enables to query the classifications and extract sub-set of data on demand based on reusable SPARQL query templates.

4.3 Interoperability of statistical classifications

RDF metadata use a formal, accessible and shared language for knowledge representation, facilitating the integration with other data and enhancing the interoperability between Statistical Classifications.

One illustration is the 'PRODCOM List' and the 'Combined Nomenclature' that describe an additional indication of 'volume physical unit' quantifying the units being classified. In SDMX, this property is managed in an Annotation that references the textual description of the Unit, repeated as many times as the unit is associated with a code of the classification. In XKOS, the Units of Measurements are assigned Persistent

Identifiers and maintained in a 'Unit of Measurement' standardised Code List, referenced by both classifications.

The best example to illustrate the interoperability between classifications is the management of the correspondences. In XKOS, a Correspondence Table is considered as a dataset that contains a set of concept associations referencing the mapping properties between a source and the target resource(s). A Concept Association may have one target (one to one), more than one target (one to many) or any target if there is no correspondence in the target classification.

Correspondence tables are established between Eurostat classifications stored in RDF in Vocbench but also to some international classifications (ISIC, CPC), available remotely in RDF in the Caliper platform, a project run by the Food and Agriculture Organization of the UN (FAO). Machine-actionable metadata enables to dereference the context of the target resources and as a result, present it in a human-readable view, facilitating the assessment of the alignment.

Table 2: Comparison between SDMX and XKOS w.r.t. their compliance with the FAIR Principles

Legend: ●=fulfilled criterion, ○ = partially fulfilled criterion, – not fulfilled

FAIR Principles[9]		**SDMX**	**XKOS**
F1	(Meta)data are assigned globally unique identifiers	●	●
F2	Data are described with rich metadata	●	●
F3	Metadata clearly and explicitly include identifier of the data they describe[10]	●	–
F4	(Meta)data are registered or indexed in a searchable resource[11]	○	●
A1	Metadata are retrievable by their identifier using a standardised communication protocol[12]	○	●
A2	Metadata should be accessible even when the data is no longer available	●	●
I1	(Meta)data use a formal, accessible, shared and broadly applicable language for knowledge representation	●	●
I2	(Meta)data use vocabularies that follow the FAIR principles[13]	–	●
I3	(Meta)data include qualified references to other (Meta)data[13]	○	●
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes	●	●

Overall, the two approaches have different strengths and weaknesses, as illustrated in Table 2 above, motivating the parallel dissemination of statistical classifications in both formats.

[9] See: https://www.go-fair.org/fair-principles/

[10] SDMX Web Service enables the retrieval of all the objects that reuse a particular Code List. This is currently not possible in SKOS as Statistical Datasets are not implemented in RDF. (Aracri, et al., 2015)

[11] In SDMX, only the ID, Name and Description are searchable, not the Annotations.

[12] In SDMX, only the Code List is retrievable, not the individual Code Items.

[13] SDMX Artefacts includes references to other data, whose content is not dereferencable.

5. Challenges and opportunities

All statistical classifications hosted in Ramon have been converted in SDMX XML and will be uploaded in the Euro SDMX Registry by mid-2022, enabling the SDMX implementers to select classification codes representing the allowed dimensions in Data or Metadata Structure Definitions (DSD and MSD). The main European Statistical classifications (NACE Rev. 2, CPA 2.1, Combined Nomenclature 2022, 2021, 2020, 2019, PRODCOM List 2022, 2019-2020) developed and maintained by Eurostat have been transformed in RDF in Vocbench, and are already accessible in ShowVoc. Their dissemination in the EU Vocabularies website will take place during the summer of 2022.

A recommended optimisation would be to consider the direct conversion from XKOS to SDMX, transforming the XKOS specific properties into SDMX annotation types and properties (URL, if the RDF value is a persistent identifier or description, if the RDF value is a literal), rendering the classifications automatically findable, accessible and reusable by the SDMX Community in the Euro SDMX Registry – and making them technically interoperable with the other SDMX artefacts. The XKOS format would thus be the source version (the single point of classification data entry), with the SDMX registry containing a representation (view) thereof.

The usage of XKOS is increasing opportunities for further collaboration between Eurostat and the ESS Members for developing, reusing and linking reference and derived classifications although the main challenge remains the availability of these statistical classifications in RDF for enabling their interoperability. While RDF representation of statistical classifications gives a greater compliance with the FAIR principles, European statistics datasets are generally not represented in RDF and thus cannot reuse RDF vocabularies for annotating their variables, rendering the implementation of principle F3 difficult to achieve. (Aracri, et al., 2015).

References

Aracri, et al. (2015), Official Statistics meets the Semantic Web: How SDMX and RDF can live together. Paper presented to New Techniques and Technologies for Statistics 2015. Available at: https://europa.eu/!cWJnhN

Cox et al. (2021), Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6): e1009041. https://doi.org/10.1371/journal.pcbi.1009041

DDI (2019), XKOS – Extended Knowledge Organization System. Available at: https://ddialliance.org/Specification/RDF/XKOS

Eurostat (2017), NACE Rev. 2: statistical classification of economic activities in the European Community. Available at: https://op.europa.eu/s/v6cb

Eurostat (2019), Formalisation of the Structure and Content of Statistical Classifications (Version 1.0). Available at: https://europa.eu/!Fhhk7c

Eurostat (2020), Euro SDMX Registry Web Services Guide. Available at: https://ec.europa.eu/eurostat/cros/content/euro-sdmx-registry-web-services-guide_en

GO FAIR, FAIR Principles. Available at: https://www.go-fair.org/fair-principles/

Hancock A. (2013), Best Practice Guidelines for Developing International Statistical Classifications. Available at: https://unstats.un.org/unsd/classifications/bestpractices/Best_practice_Nov_2013.pdf

Hoffman, E. and Chamie M. (1999), M. Standard Statistical Classifications: Basic Principles. Paper presented to the 30th session of the UN Statistical Commission. Available at: https://unstats.un.org/unsd/classifications/bestpractices/basicprinciples_1999.pdf

SDMX (2018), Guidelines for the Creation and Management of SDMX Code Lists (Version 3.0). Available at: https://sdmx.org/?page_id=4345#CodeListGuideline

SDMX (2021), Guidelines on using SDMX Annotations (Version 1.0). Available at: https://sdmx.org/?page_id=4345#Annotations

SDMX (2021), SDMX Technical Specifications. Available at: https://sdmx.org/?page_id=5008

W3C (2019), SKOS Simple Knowledge Organization System Reference. Available at: https://www.w3.org/TR/2009/REC-skos-reference-20090818/