UNECE High-level Group for the Modernisation of Official Statistics

Business Case for linking and querying statistical classifications through the semantic web standards

This business case was prepared by Christine LAABOUDI (Eurostat), and is submitted to the HLG-MOS for their approval.

Type of Activity

☒	New project	□	Extension of existing project

Purpose

Relations between statistical classifications is given by the international system of economic classifications maintained by the United Nations Statistical Division. The economic classifications forming this system are linked either by a common structure, which is more and more detailed from international to European and then national levels, or by semantic correspondences between the economic fields covered. The current practice for establishing the correspondence tables between the reference and derived classifications is not standardised, for example, based on different identifiers or distributed in formats not always machine-readable, preventing their interoperability and reuse across organisations.

The harmonisation of statistical reference and national classifications in XKOS, a data model for representing statistical classifications in the semantic web, and their dissemination according to the LOD (Linked Open Data) principles in RDF format provides an opportunity for increasing a further collaboration between the statistical organisations and implementing a federated search interface for querying our statistical classifications.

This project shall demonstrate the usefulness of linking statistical classifications to the statistical community but also, with the usage of a common language and understanding, being better aligned to the need of the Linked Data communities. In addition to sharing good practices for linking and dissemination as LOD, the federated search interface shall facilitate the extension of national versions or the annotation of statistical datasets.

Description of the project and the Work Packages/sub-activities

The project consists in implementing a federated search engine that should enable to query Linked Data classifications stored in different Triple Store and federate the search results in one interface, and thus demonstrating the usefulness of linking and querying statistical classifications at the different levels (UN, EU, national).

Project composition

As there are already a few statistical organisations with some experience with LOD, the project could be built in a way that members with already some experience and new members, not having any experience at all can both add great value.

The implementation project divided in 2 phases and 3 Work Packages (WP)

Phase 1: Project set-up

WP1 – Selecting the statistical classifications and correspondence tables

In the first activity, the group will identify the relevant classifications and correspondence tables (datasets) that are included in the scope of the project, such as ISIC/NACE, CPC/CPA or geographical/regional classifications that are all have national variants.

LOD correspondence tables exist between UNSD and EU classifications (NACE – ISIC, CPA – CPC), and are accessible in EU Vocabularies [i] or in the Caliper platform [ii] run by FAO (Food and Agriculture Organization of the UN (FAO).

The selected datasets will delivered in XKOS and stored in a Triple Store in RDF format by the group members. Technical assistance will be provided to the members with less experience for transforming and storing their data and making it compliant to LOD principles.

Phase 2: Implementation phase

WP2 – Development of the federated search interface

The interface will query multiple classifications and correspondences and propose a set of predefined queries to the End-User. Once the End-User have selected a query and entered the parameter(s), the System runs the query across the different datasets simultaneously via SPARQL endpoints and then federates the search results on the screen of the interface.

Examples of queries : Get the correspondences of a code in the targeted classifications, display a correspondence in a target classification with its code hierarchy (narrower levels), search a text in the source classification and get the correspondences from the target

The first benefit of exposing datasets in RDF is machine-readability, enabling a machine (the Interface) to query to data via a SPARQL Endpoint. The second benefit is human readability, returning hyperlinked referring to the webpage of a classification item.

The interface must enable the modification and storage of the SPARQL queries in order to allow further extension of the queries once new datasets are available in RDF.

WP3: Project management. The project should be managed by a project manager and frequent reporting should be done according to all expectations. (Phase 1 and Phase 2)

Deliverables and timeline

The proposal has the potential for a project managed in 1 years. The scoping of the project (WPs included and the expected outputs) and their timing are to be adjusted accordingly to needs and resources (mainly number of countries/experts) committed to the project.

2023 January – 2023 June

Phase 1 – Project set-up

- WP 1.1 List of selected classifications and correspondence tables (handbook)
- WP 1.2 Classifications and Correspondence tables selected in WP 1.1 available in RDF (Datasets)

2023 April – December 2023

Phase 2 – Implementation phase

- WP2.1: Specifications of the SPARQL queries (handbook)
- WP2.2: Federated search interface (tool)
- WP2.3: Federated search interface documentation (handbook)

Phase 1 – Phase 2

- WP 3: Project Management reporting

The Phase 2 (implementation phase) can start as soon as two families of classifications at UN – EU and National level and their correspondences are available in RDF.

The development of the Interface will be taken over by one member of the group.

Offices/Countries committed

Eurostat, Hungarian Central Statistical Office, INSEE (France), Statistics Canada, ILO (International Labour Organization), Statistics Netherlands	
Alternatives considered	
Without cooperation, statistical organisations will likely continue implementing solutions on their own, for maintaining or consuming their classifications and correspondence tables, which are not interoperable or reusable in terms of data model or format.	
How does it relate to the HLG-MOS vision and other activities under the HLG-MOS?	
The project will provide a continuity to the efforts that the HLG-MOS have been developing in the areas of Implementing ModernStats Standards, specifically in relation with the deliverables and output of the Linked Statistical Metadata Project. In the framework of the "HLG Project proposal on Linked Statistical Metadata" (2015), a Demo tool (classification explorer) was developed for browsing RDF classifications uploaded in a common repository. The code is available in the UNECE Github.	
Proposed start and end dates	
Start: *January 2023*	**End:** *December 2023*

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

[ii] Caliper Statistical Classifications in a Linked Open World https://www.fao.org/statistics/caliper/web/