Modernization of statistical classifications in the EU

Christine Laaboudi-Spoiden, Eurostat unit B1

ModernStats Community of Practice webinar on Linked Open Data
4 December 2023

Table of contents

European
Commission

RAMON

Eurostat Reference And Management Of Nomenclatures platform

- Previous dissemination platform, **phased out in June 2023**

What did RAMON contain?

- Statistical classifications + Correspondence Tables

   Maintained by Eurostat and Third parties (datasets were replicated)

- Standard Code Lists, disseminated in Eurostat Reference Database (Eurobase)

- Specialised glossaries « Coded » (concepts and definitions)

- Statistical metadata catalogues (ESS officials standards, statistical methodologies)

FAIR principles in RAMON

Limited **A**ccessibility

- Access URL, HTML pages (hierarchical view)

Limited **F**indability

- Search by code

- Conceptual Knowledge representation, not technically standardised

Limited **R**eusability

- Additional files for downloading attached in different formats

    - Open (CSV), Proprietary (XLS, MDB, PDF)

- Links to third party websites

European Commission

FAIR principles in RAMON

Limited Interoperablity

Classifications and Correspondences were not interoperable

- Mapping based on the classification code or Key

- Stored in different formats (Excel, CSV) and different file structures (headings)

- Issues:
  - missing leading zero;
  - different codes or IDs across classification series (Code 12.4 or 124)

- The semantics of the source/target was not Accessible (Name, Explanatory Texts)

European Commission

Modernisation of structural metadata

- Innovate the dissemination of structural metadata
    - necessary to move the content of Ramon to other systems

- New dissemination channels

- Compliant to the **FAIR** Principles

    Findable, Accessible, Interoperable, Reusable

- Open standards and formats

   SDMX/XML, RDF (SKOS, XKOS)

New Dissemination channels

WHAT ?	EU Vocabularies LINKED OPEN DATA (LOD)	Euro SDMX Registry	CIRCABC RAMON ARCHIVES
Statistical classifications	ESS classifications	All classifications (from RAMON)	Additional files (from RAMON)
Correspondence tables	Correspondences ESS classifications Third parties if LOD	Correspondences (from RAMON)	Correspondences XLS format
Standard Code Lists		OFFICIAL REPOSITORY	
Glossaries (Coded)	Selection Quality glossary	Full Coded dataset (from RAMON)	
Metadata catalogues	Selection EBS Manuals, ESS Standards		
FORMATS	RDF	SDMX/XML	VARIOUS

European Commission

LOD – Dissemination (state of play)

Statistical classifications

- Combined Nomenclature
- PRODCOM
- NACE (Rev. 2, Rev. 2.1)
- CPA (2.1, 2008)eco
- ECOICOP/HICP
- GEONOM (Geonomenclature)
- NST 2007 (Goods for transport)
- NUTS (Terrritorial units)
- ACL HETUS
- CEPA, CREMA (Environment)
- EWC, LoW (Wastes)
- CBF (Business functions)
- ESeG (Socio-economic groups)
- ICST-COM (Ships by type)
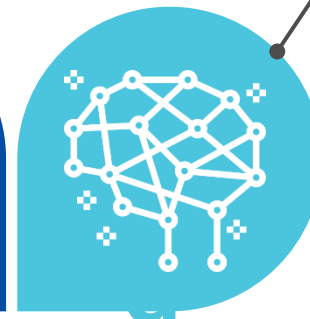
Metadata catalogues

- EBS (Business Statistics Manuals)
- ESS-Standards
- Quality glossary

Code lists

- Supplementary units
- Production units
- Waste categories

Classifications used for the production of European statistics

European Commission

LOD – EC corporate tools

Vocbench

Maintenance
- Transformation of structured data into RDF
- Input: Eurostat Business Units

ShowVoc

Visualisation component of Vocbench
- **LOD Dissemination**
- Advanced visualisation in XKOS
- Integrated **SPARQL Endpoint**

EU Vocabularies
- EU Corporate website for the **dissemination** of EU controlled vocabularies
- TripleStore (Cellar)
- SPARQL Endpoint (federated + remote queries)

data.europa.eu

European Open Data portal
- **Data catalogue**

Tools offered by the Publications Office of the EU

European Commission

LOD dissemination platforms

ShowVoc

Classifications exposed by Eurostat are **immediately** available

- as soon as validated in Vocbench
- All their properties are displayed

Limited SPARQL Endpoint enabling

Federated search, only across datasets hosted **in the ShowVoc Knowledge Graph**

EU Vocabularies (Cellar)

Official Knowlege Graph of OP

Planning for Eurostat dissemination in Cellar (every 3 Month)

- Once disseminated, the classification are automatically listed under EU Vocabularies > taxonomies
- but the visualisation is limited to the basic properties (skos)

SPARQL Endpoint enables:

Remote federated search between Triple Store

Cellar End-Point and API query with a generated URL

European Commission

LOD dissemination platforms

ShowVoc

Hierarchy and full representation

EU Vocabularies (Cellar)

Hierarchy and of basic elements

LOD – Data modelling and standardisation

SKOS: Simple Knowledge Organization System (W3C)

- A data model for representing controlled vocabularies (generic)

XKOS: An SKOS extension for representing statistical classifications (DDI)

- XKOS Best practices web document (July 2023) = **cookbook**

DCAT, DCAT-AP:

- Description of the classification (dataset) in open data catalogue

COOS (Core Ontology for Official Statistics, Modernstats Standards)

- Description of information resources (standards, methodologies)

LOD – Persistent Identifiers

Resources are defined in the domain **data.europa.eu**

- One namespace per classification serie

  - **ux2** for NACE, **ehl** for CPA

- A Permanent URI by ressource

  Classification
  http://data.europa.eu/ux2/nace2.1/nace2.1

  Item http://data.europa.eu/ux2/nace2.1/3600

  Level http://data.europa.eu/ux2/nace2.1/sections

Resources are human and machine-readable

Dereferencable URIs
(return the information about the thing identified by the URI)

Accessible via a Web API
SPARQL Endpoint

European Commission

LOD – RDF triples

Transformation of the structure content into RDF Triples

- Integrated tool in Vocbench

What is a RDF triple ?

- Subject (a resource) → Predicate (a property) → Object (a resource, a Literal)

- Examples:

NACE 36.00 **has a parent** NACE 36.0

http://data.europa.eu/ux2/nace2.1/3600   skos:broader   http://data.europa.eu/ux2/nace2.1/360

has a name (@en, @fr)

http://data.europa.eu/ux2/nace2.1/3600> skos:prefLabel 36.00 Water collection, treatment and supply@en

LOD – Linking via Persistent Identifiers

- Association between URIs

- Linking resources (Correspondence tables)

  - Accessing the semantics of the Source and Target(s) code items

    Example: NACE Rev. 2.1 - NACE Rev 2.

- Linking classification properties to resources

  - *Territorial Typologies (Tercet)* categorising NUTS (Nomenclature of Territorial Units)

  - *Supplementary Units* qualifying goods in Combined Nomenclature or products in Prodcom

LOD – Added-value of XKOS

XKOS enables different types of representations

- Hierarchical (broader to narrower levels)

- By classification level: NACE Sections, Divisions, Groups, Classes

- Breakdown of codes in categories

  Main Industrial Groupings (MIGS 2007) of NACE codes

- One dataset can contain more than one classifications (variants)

  A code item (URI) can be **reused** in different classifications

    Example

    ECOICOP and ECOICOP-HICP sharing the same codes

    HICP being limited to ECOICOP codes used in the index of consumer prices

Benefits to expose classifications as LOD

Eurostat Reuse use case: R package for automatically generating candidate correspondence tables between classifications

https://github.com/eurostat/correspondenceTables/

Data ingestion facilited by a function directly **accessing** classifications & correspondence tables data via a SPARQL endpoint

- Eurostat Classifications (OP Triple Store Cellar – EU Vocabularies)

- Classifications available remotely (ISIC, CPC from FAO Caliper Triple Store)

Interoperability enables by XKOS (common terminology)

Benefits to expose classifications as LOD

Develop a global ecosystem of FAIR statistical classifications

- Link European, national dimensions to the global counterparts

    - Classifications used for the production of European statistics (Eurostat)

    - UNSD and UN Agencies are custodians of reference classifications, also used for the production of European statistics (ISCED, ISCO, COICOP)

    - Some NSOs offers national versions (NACE) as LOD

- Offering reference classifications

    Example: The Classification of Statistical Activities, widely used:

    - for aggregating the classifications by classification families

    - for categorising the statistical activities (Eurostat statistical compendium)

Links

Tools

ShowVoc desc

Training courses

ShowVoc (link to the training material)

Access to

ShowVoc

Cellar API

User guides

Modeling of Eurostat's statistical classifications in ShowVoc

SPARQL Queries User Guide

ESTAT Website > Metadata

Classifications used for the production of European statistics

Thank you

ESTAT Datat Metadata Services , ESTAT-DATA-METADATA-SERVICES@ec.europa.eu

European Commission