ModernStats 2022

Statistical Classifications and Metadata Catalogues exposed as Linked Open Data

*ModernStats Workshop 2022, 29 June 2022 (Belgrade)*
Table of contents

1. RAMON
   (Reference and Management of Nomenclature)
   • What does it contain?
   Migration
   • To SDMX
   • As LOD (RDF)

2. Migration as LOD
   • Transformation in RDF
   • EC Corporate tools
   • Dissemination (State of play)
     • Statistical classifications
     • Organisations
     • Metadata catalogues

Challenges and opportunities
• Statistical classifications
• Metadata catalogues
• Use Case: the Statistical Compendium
Link to ModernStat Standards
Eurostat Reference And Management Of Nomenclatures platform
Hosting, since the early 2000s:

• Statistical classifications
• Correspondence tables
• Standard code lists used in dissemination (Eurobase)
• Metadata catalogues
  • ESS standards, statistical legislation, statistical methodologies
Statistical classifications in RAMON

RAMON offers

• All the classifications managed by Eurostat (EU Level)
• Classifications developed by other custodians (World Level)

Semantic compliance

• « basic principles for standard statistical classifications » (Hancock, 2013)

Technical compliance

• Structure but not standardised: properties depending on the classification
• Human readable: browsable, searchable by classification code (HTML)
• Classifications are downloadable in various formats (additional files)
  • Open (CSV, XML) or proprietary (MDB, Excel) formats
Correspondence tables in RAMON

Correspondence tables between Eurostat or International classifications

Downloadable but **not** Findable

- Different structures and formats (Excel, CSV, ZIP, URL)

Classifications and correspondences are **not** Interoperable

- Mapping based on the classification code, sometimes **different** from the classification item ID
- **No access** to the (semantics) context about the source or target(s) codes
- **No access** to the Correspondence(s) of a code item (in the classification)
RAMON migration

RAMON runs on ColdFusion, phased out at the end of 2022.

Therefore necessary to move the Ramon objects within it to other systems

Migration strategy in two phases:

- SDMX Registry
- Expose as Linked Open Data (LOD)

Why?

- Render the resources **FAIR** (Findable, Accessible, Interoperable, Reusable)
- Open Standards and formats (SDMX and Linked Open Data)

Also an occasion to rethink about what we store and expose in RAMON (and why, how)
Migration to SDMX

Euro SDMX REGISTRY

- Script for mapping RAMON elements to SDMX Properties
  - The SDMX Information Model
  - The Guidelines for SDMX Annotations
- Structural Interoperability
- Availability in the Registry: 2nd Semester 2022

Make this information to be more easily Findable, Accessible and Reusable for data collection

What?
- Classifications managed by Eurostat
- Classification managed by third parties (for reuse purpose)
- Standard Code Lists

Where?
- Euro SDMX REGISTRY

Output
SDMX/XML
Migration as LOD

Exposed as **Linked Open Data (LOD)**

- Requires more modelling flexibility
- Less integration with our production systems

**EC corporate tools**

Offered by the Publications Office of the EU

For the maintenance and dissemination of EU controlled vocabularies

- Taxonomies, Nomenclature, Thesaurus
- Authority Tables
- Ontologies

**What?**

- All the classifications managed by Eurostat
- Correspondence tables (Eurostat, + Remote classification available as LOD)
- Metadata catalogues

**Where?**

- EU Vocabularies (Dissemination)

**Output**

RDF
LOD – EC corporate tools

**Vocbench**
- **Maintenance**
  - Ontologies
  - RDF datasets

**ShowVoc**
- **Visualisation** component of Vocbench (similar GUI)
  - Advanced visualisation

**EU Vocabularies**
- EU Corporate website for the dissemination of EU controlled vocabularies
- Business collection: “Eurostat Statistical Classification”

**data.europa.eu**
- European Open Data portal
  - Description of the classifications in the data catalogue
LOD – Transformation

Integrated tool in Vocbench

• Transformation of structured content (XLS, CSV) into RDF triples

Semantic Interoperability

• Conceptual mapping between SDMX and SKOS/XKOS properties

Statistical classifications are developed by the business units in Eurostat and uploaded in Vocbench for dissemination purpose only.

Semantic Web representation

SKOS: Simple Knowledge Organization System

• W3C recommendation
• Ontology used for representing controlled vocabularies

XKOS: An SKOS extension for representing statistical classifications

• DDI specification
## Crosswalk between SDMX and SKOS/XKOS

<table>
<thead>
<tr>
<th>SDMX</th>
<th>SKOS/XKOS</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classification</strong></td>
<td><strong>Code List</strong></td>
</tr>
<tr>
<td><strong>Codes</strong></td>
<td><strong>Code Item</strong></td>
</tr>
<tr>
<td><strong>Correspondence table</strong></td>
<td><strong>Structure Set</strong></td>
</tr>
</tbody>
</table>
| **Code Mapping** | **Map** | **xkos:ConceptAssociation**  
| | | **skos:mapping**  
| | | **skos:Collection**  
| | | **xkos:classificationLevel** |
| **Collection** | **N/A** | **skos:Collection** |

*Crosswalk between SDMX and SKOS/XKOS*
LOD – Dissemination (state of play)

**Statistical classifications**
- Combined Nomenclature CN2022, CN2021, CN2020, CN2019
- PRODCOM 2019, 2021
- NACE Rev. 2
- CPA 2.1
- ECOICOP/HICP
- LUCAS 2022
- GEONOM (Geonomenclature)
- NST 2007
- NUTS 2021 + Statistical Regions

**Metadata catalogues**
- ESTAT-Legis
- STAT-Manuals
- ESS-Standards

**Code lists**
- Statistical organisations
- Units of measurements

Availability in EU Vocabulary: 15/07/2022
LOD – Identifiers (URI)

- SDMX
  - URN (based on the classification Codes or CNKey for CN)
- LOD
  - Persistent URI (PUlRI) based on SDMX URN
  - One namespace by Classification family
  - A Permanent URI by Resource

Make this information to be more **F**indable, **A**ccessible and **R**eusable

Ressources are human-readable (HTML) through Content Negotiation

Ressources are machine-readable (RDF)

- dereferencable via content-negotiation
- Accessible via a SPARQL-end Point
Dataset of organisations transmitting data to Eurostat

- Used by EDAMIS and ESS-MH

Dataset maintained in Vocbench

- Multiple codes (SDMX, EDAMIS)
- One conceptScheme by Country
- Different <skos:collection>
  - Eurostat Organisation type
    ONA, EU-NSI, MED-NSI

Customised extraction for the SDMX Registry

- SDMX Organisation Unit Scheme used by ESS-MH

Make this information to be more **FAIR**

**S**tandardise the information based on

- SKOS
- ORG (Organisation) ontology
- COOS: Organisation type

Dataset available in **ShowVoc**
Metadata catalogues

ESTAT-LEGIS

- Catalogue of statistical legislative acts
  - Being maintained in Ramon without involvement with the legal unit (ESTAT A2).
  - ESTAT A2 maintain their own list of legislative acts, which are available in the Legislation in Force web page
    - subdivided by domain and year in PDF files

Make this information to be more FAIR

S
Standardise the information based on

- ELI Ontology (European Legislation Identifier)
- Information linked and Interoperable with the EU legislation disseminated in EUR-LEX

POC in Vocbench for two domains

- Transport, Energy
Metadata catalogues

ESS Standards

- Catalogue of statistical standards officially adopted by the ESS Standard Working Group (11 standards)
- Reference to GSBPM (phases)

STATMANUALS

- Catalogue of methodological manuals relating to statistics (around 1,000)
- Organised by statistical domains
- Transformed in RDF in Vocbench

Make this information to be more FAIR

Standardise the information based on

- SKOS, Dublin Core (dct:Type)
- Core Ontology for Official Statistics (COOS) + Business processes (GSBPM phases)

Missing properties

- Property for linking a methodology/standard to a process
- Property for linking a methodology to a statistical domain
Challenges and opportunities (classifications)

Technical interoperability

• Nice to have: Automated conversion of XKOS to SDMX (in Vocbench)

• Transformation of a statistical classification is not straightforward
  • RDF Transformation rules applicable to a classification family only
  • Analysis of the structure of a classification before transformation (case by case)

Semantic interoperability (opportunities)

• Reuse Eurostat classifications as a basis for developing derived classifications (national level)

• Develop and disseminate correspondence tables between reference and derived classifications
Challenges and opportunities (Metadata catalogues)

Semantic interoperability

• Between different types of data:
  legislation, methodologies, organisations

• Represented with different data model

The « Statistical compendium » use case
Statistical compendium

Statistical requirements compendium

Eurostat yearly publication: relevant documents for the compilation of statistics

Organised by domains of the CSA (Classification of Statistical Activities)

Each entry is structured as follows:

- Eurostat responsible unit (Organisation)
- Statistical legislative acts (ESTAT-LEGIS)
- Statistical methodologies (STAT-MANUALS)
- Statistical datasets
- International statistical Organisations (international agreement)

Make this information to be more Interoperable

Linked the datasets (Organisations, Legislation, methodologies) to the Classification of Statistical Activities (CSA)

Reuse the COOS classes/properties

Issue #84

Or opportunity to acquire new properties in COOS
Thank you

Christine Laaboudi-Spoiden, Eurostat, christine.laaboudi@ec.europa.eu

© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the CC BY 4.0 license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Slide xx: element concerned, source: e.g. Fotolia.com; Slide xx: element concerned, source: e.g. iStock.com