Applying FAIR principles to statistical classifications

Eurostat, Unit B1, Christine Laaboudi-Spoiden

COSMOS, Paris, 11/04/2024

Table of contents

1. Overview of the modernization of ESS Statistical Classifications
   - Data Transformation (SDMX, RDF)
   - Tools for the maintenance and dissemination of statistical classifications
   - Availability of ESS structural metadata as LOD
2. Compliance with the FAIR Principles
3. LOD Community of Practice

Modernisation of ESS classifications

UNTIL JUNE 2023:

RELATIONAL DATABASE

JANUARY 2021 – 2023:

TRANSFORMATION PHASES

FROM JULY 2023:

NEW DISSEMINATION PLATFORMS

ESS classifications (until June 2023)

Limited Findability	No Persistent identifiers Search only by Code No standardised knowledge representation
Limited Accessibility	Files in different formats or links to third party websites Different file structures
Limited Interoperability	Correspondences not standardised (different identifiers and file formats)
Limited Reusability	HTML description, not standardised

Major modernisation steps

Step 1. SDMX

- Converting all statistical classifications from RAMON into SDMX
- Converting the Standard Code Lists (Eurobase) into SDMX
- Exposing them in the Euro SDMX Registry

Step 2. Linked Open Data (LOD)

- Converting the statistical classifications used for the production of European Statistics into RDF (Eurostat is the custodian)
- Converting the correspondence tables, provided that targets are available in RDF
- Exposing them as Linked Open Data (LOD)

Step 1. Data transformation to SDMX

Conversion from RAMON to SDMX/XML

Conceptual mapping between RAMON elements and SDMX properties

Script based on the SDMX Information Model

- Basic structural elements (Identifier, code, name, parent)
- SDMX annotations (Explanatory notes, case law, levels, units of measure)

Storage in the Euro SDMX Registry

Additional filter for filtering the classifications

- Classifications are prefixed by CLS_ (CLS_NACE_REV2)
- Standard Code Lists are prefixed by SCL_ (SCL_GEO)

Dataset downloadable in SDMX-XML format or via a query to the SDXM Registry Rest API

Step 2. Data transformation as LOD

Transformation of the structure files (CSV, Excel) into RDF Triples

- Files delivered by Eurostat Business Units
- Forthcoming: **RDF to SDMX/XML Exporter** (SDMX 3.0)

Standardisation based on semantics standards

SKOS: Simple Knowledge Organization System (W3C)

Generic data model for representing RDF controlled vocabularies

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

XKOS Best practices (released in July 2013)

For the maintenance, storage and dissemination of classifications,
we use tools (open sources) offered by the EU Publications Office

LOD – Classifications tools

Maintenance & editing

Vocbench

- Transformation of structured data into RDF triples

Dissemination

ShowVoc

- Search, Download
- Advanced visualisation in XKOS
- Integrated **SPARQL Endpoint** (one per dataset)

Cellar (Triple Store)

- EU Vocabularies : EU Corporate website for the **dissemination** of vocabularies from Cellar
- SPARQL Endpoint (federated queries) + API query with a generated URL

Reuse **data.europa.eu**

- European Open Data portal
- Dataset descriptions (DCAT-AP)

Corporate Tools offered by the Publications Office of the EU

European Commission

LOD – ESS Linked Open structural metadata

Statistical classifications ✓

- Combined Nomenclature
- PRODCOM
- NACE (Rev. 2, Rev. 2.1)
- CPA (2.1, 2008)
- ECOICOP/HICP
- GEONOM
- NST 2007
- NUTS 2024, TERCET
- ACL HETUS
- CEPA, CREMA
- EWC, LoW (Wastes)
- CBF (Business functions)
- ESeG (Socio-economic groups)
- ICST-COM (Ships by type)

Metadata catalogues

- Business Statistics Manuals
- ESS approved Standards
- Glossaries: Prices, Quality

Code lists ✓

- Supplementary units
- Production units
- Waste categories

ESS Classifications used for the production of European statistics

❑ F
❑ A
❑ I
❑ R

FAIR principles: SDMX vs. XKOS

☑ Compliant
📋 Partially compliant
✕ Not compliant

FAIR	Principles	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	☑	✕
F4	(Meta)data are registered or indexed in a searchable resource	📋	☑
A1	Metadata are retrievable by their identifier using a standardised communication protocol	📋	☑
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	✕	☑
I3	(Meta)data include qualified references to other (Meta)data	☑	☑
R1	(Meta)data are richly described with a plurality of accurate and relevant attributes	☑	☑

☑ F
□ A
□ I
□ R

FAIR – Findability

(Meta)data are assigned globally unique identifiers

LOD (URI)

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

One namespace per classification serie

- **ux2** for NACE
- http://data.europa.eu/ux2/nace2.1/nace2.1

One URI by ressource

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

Data are described with rich metadata (XKOS)

SDMX (URN)

One URN per artefact:

- urn="urn:sdmx:org.sdmx.infomodel.codelist.**Codelist**=ESTAT:NACE21(1.0) **agencyID="ESTAT"** id="NACE21
- urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=ESTAT:NACE21(1.0).3600" id="3600"
- Codelist, Code, Hierarchies, RepresentationMap

Data are described with rich metadata (basic structural elements + annotations)

European Commission

□ F
☑ A
□ I
□ R

FAIR – Accessibility

Metadata are retrievable by their identifier using a standardised communication protocol

LOD

Resources are machine-readable, accessible via a SPARQL End-point or API

Resources are dereferencable (URI returns the elements about a resource)

SDMX

Some resources are machine-readable, accessible via the SDMX API

In SDMX, only the code list is retrievable (via download), not the individual code items.

□ F
□ A
☑ I
□ R

FAIR – Interoperability (Knowledge representations)

GSIM 2.0 Concepts	XKOS classes	SDMX objects
Statistical Classification Code List, Concept System	<skos:ConceptScheme>	<str:Codelist>
Classification item Code Item, Concept	<skos:Concept>	<str:Code>
Classification level	<xkos:ClassificationLevel>	SDMX Annotation (Type: HIER_LEVEL)
Node	<skos:Collection>	<str:Hierarchies>
Correspondence Table	<xkos:Correspondence>	<str:RepresentationMap>
Map	<xkos:ConceptAssociation>	<str:RepresentationMapping>

(Meta)data use of a formal, accessible, shared and broadly applicable language of for Knowledge representation

❑ F
❑ A
☑ I
❑ R

FAIR – Interoperability

(Meta)data use of a formal, accessible, shared and broadly applicable language of for Knowledge representation

R package for automatically generating candidate correspondence tables between classifications

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

- Facilitated data ingestion by a function directly **accessing** classifications & correspondence tables data via a SPARQL endpoint
  - Eurostat Classifications (OP Triple Store Cellar – EU Vocabularies)
  - International or national classifications available remotely (ISIC, CPC from FAO Caliper Triple Store)
- Interoperability enables by the XKOS common knowledge representation

□ F
□ A
□ I
☑ R

FAIR - Reuse

Meta(data) are richly described with a plurality of accurate relevant attributes

Eurostat data catalogue in data.europa.eu (European Data Portal)

- 8 000 datasets distributed in different formats (SDMX, TSV, CSV)
- Descriptions compliant with DCAT-AP (extension of DCAT)
- DOI registered to DataCite (enhanced data discovery and data citation)

Description of statistical datasets within statistical domains

- StatDCAT-AP (StatDCAT Application Profile, extension of DCAT-AP)
- Statistical dataset structure : dimensions, attributes, units of measurements, quality annotations, number of time series

Opportunity

- Linking statistical datasets with classifications (dimensions)
- Finding statistical datasets sharing the same dimensions

ESS & UNECE LOD Community of Practice

Objective

- Sharing experience and best practices as well as providing visibility to initiatives for querying and linking statistical classifications

Benefits

- Demonstrate the usefulness of Linked Open Data
- Better aligned to the need of the LOD community
- Discuss the challenges and added-value based on real use-cases

Participants

- 12 NSOs (ESS members, Statistics Canada), 1 international organisation (FAO)
- 4 workshops in 2023, UNECE LOD Community of Practice

4 Task Teams:

- Linking datasets and their structural metadata
- Linking statistical classifications
- API for querying statistical classifications
- Linking statistical datasets in data catalogue

Links

Access

- ShowVoc
- Cellar API
- European Open Data Portal

Contact email:

- ESTAT DATA METADATA SERVICES

ShowVoc training material

User guides

- Modeling of Eurostat's statistical classifications in ShowVoc
- SPARQL Queries User Guide

ESTAT Website > Metadata

List of ESTAT classifications used for the production of European statistics

Thank you

European Commission