Achieving optimal interoperability with statistical classifications, codesets and concordances

Stéphane Crête and Flavio Rizzolo

Enterprise Architecture, Strategy & Innovation

June 27th, 2019

Delivering insight through data for a better Canada

Canada's Digital Journey

Canadian Data Strategy Roadmap of the Federal Public Service

Greater usability and availability of data

Digital 9 (D9)

Collaborative network of the world's leading digital governments with a common goal of harnessing digital technology to improve citizens' lives.

International Open Data Charter

A set of principles and best practices for the release of governmental open data.

Statistics Canada – Data Analytics as a Service

A set of capabilities enabling data and analytics.

Opportunity Statement

- Statistical Classification Stewardship is well established.
- Statistical Classifications are well represented in GSIM (Generic Statistical Information Model).
- Statistical Classifications are well represented in RDF (Resource Description Framework) with the use of SKOS (Simple Knowledge Organization System) and XKOS (Extended Knowledge Organization System).
- Optimal interoperability requires both common *information objects* and common *methods* to manipulate them (CRUD).
  - Long tradition of *information object* standardization via common data/information models in the statistical domain, e.g. GSIM, SDMX, DDI, Neuchâtel, etc.
  - No standardization of *methods*.
- We propose to define a standard Statistical Classification *API* : set of agreed-upon specifications for protocol messages (methods) and responses (information objects).

Interoperability layers

- Interoperability is often understood as being just about communication protocols and infrastructure
- In fact, interoperability has multiple levels
  - The European Interoperability Framework identifies 4 levels, i.e. technical, semantic, organizational and legal.
  - Others split the first two into three or four, e.g. system, syntactic, structural and semantic.
- Core interoperability:
  - System (technical): infrastructure and physical communication protocols, e.g. HTTP, REST, etc.
  - Syntactic (structure): common representations (data formats) and exchange models, e.g. SDMX schemas, DDI 3.x, NIEM, etc.
  - Semantic (meaning): conceptual models, vocabularies and ontologies, e.g. SDMX concepts, DDI 4 conceptual model, SKOS/XKOS, etc.

Some common endpoint types

	Approach	Usability	Protocol	Return type
REST	Entity-centric	Easy	HTTP	JSON, XML, HTML, YAML, TEXT, etc.
SOAP	Function-centric	Easy	HTTP, SMTP, UDP, etc.	XML
SPARQL	Query-centric	Hard	HTTP	JSON-LD, RDF-XML, Turtle, n-triples, HTML
ODATA	Model-centric	Medium	HTTP	JSON, XML

API (entities and methods) standardization

- "*Entities* [and their *operations*] are not to be multiplied beyond necessity" – (A variant of the famous Ockham's razor)
- We have tackled the entity multiplication in the statistical domain with the development of common data/information models, e.g. GSIM, SDMX, DDI, Neuchâtel, etc.
- However, we haven't standardized the mechanisms to manipulate entities in those models. This has led to a proliferation of ways to
  - Access and interact with the same information object (CRUD)
  - Define a fragment of an information object
  - Deal with composite information objects (including some or all of their parts)
- We need to standardize operations (API methods) to tackle their unnecessary multiplication.

Example: Get statistical classification

- **SOAP**: potentially large number of operation names (e.g. getClassification, getSC, Classification, StatClass, getStatClass, etc.)
- **REST**: just one operation name (GET), but many possible interpretations/implementations (e.g. whole classification, a subset of attributes, related entities like levels and items, etc.)
- **Both protocols**: many different ways of defining the same parameters (e.g. *items*, *withItems*, *hierarchy*, *includeHierarchy*, *complete*, etc. are all potential names for a Boolean parameter indicating whether or not to return classification items)
- Without some sort of entity operation standardization, interoperability is at risk.

API design

- What constitutes an entity?
  - A class (information object) in an information model
  - A group of classes related by part-whole relationships (e.g. UML aggregation/composition)
  - A group of classes realizing a pattern (e.g. DDI collections pattern)
- Entity-centric API design principles
  - Methods should be coupled to entities
  - Each entity should have only one method per CRUD operation
  - Each method should manipulate:
    - An entire entity, including all its sub-entities (parts)
    - A sub-entity (or a set of them)
    - An entity fragment, i.e. a subset of its attributes
    - Links to associated entities (for navigation)

Rest vs. SOAP

- **REST**: entity(or resource)-centric
  - CRUD operations (GET, PUT, etc.) apply to a resource. A resource is an entity in a data model, which can be a composite of classes
  - Aligns well with the design principles, as long as resources are mapped to the proper information model classes
- **SOAP**: function(or method)-centric
  - SOAP methods can do CRUD with arbitrary sets of entities
  - Methods should be designed to align with the design principles

Current State – Web Service APIs

Statistical Classifications and Codelists

A statistical classification is “official”, and its elements are mutually exclusive and complete. Code lists are not "official" and may be fitted to one specific statistic. Code lists may be included in the search process by ticking off the box. Note that the amount of search results can be vast, including a lot of code lists adapted to specific Statistics Norway needs.

Klass API guide. REST API with the formats JSON, XML and CSV

Search for classifications (and codelists)

Search | Search

Filter: Include codelists | Responsible division: All divisions

Navigate by subject area | Open hierarchy

- Labour market and earnings (1)
- Banking and financial markets (1)
- Population (6)

Klass API Guide

1. Overview

1.1. Rest Client

Examples in this documentation uses curl. If this tool is unknown an alternative is to use a Rest client to explore Klass Rest interface, see instructions Rest client guide.

1.2. URL Encoding

The current version of this API requires you to use Percent-encoding for symbols and characters that are not part of the standard unreserved URI characters.

For more information on Percent-encoding see this wikipedia article

You can also see it in use in the request example for presentationNamePattern

1.3. HTTP status codes

Current State – Downloadable DDI

EBOPS 2010 - SERVICE CODES V1.0

NAME: EBOPS 2010 - Service codes v1.0

DESCRIPTION:
Denmark's foreign trade in services is calculated on the basis of the United Nations Service Companion Nomenclature EBOPS 2010 (Extended Balance of Payments Services Classification), as evidenced by the International Trade Service Manual (Manual on Statistics of International Trade in Services, MSITS), issued by the United Nations and others by 2010. EBOPS is incorporated as a nomenclature in the EU Balance of Payments Regulation, etc., cf. above.

VALID FROM: 2010-01-01T00:00:00 **OFFICE**: External Economy **CONTACT**:

Codes and categories

OPEN HIERARCHY | DOWNLOAD

- CSV
- DDI

0: SERVICES
1: MANUFACTURING S… INPUTS OWNED BY OTHERS
2: MAINTENANCE AND REPAIR SERVICES NOT INCLUDED ELSEWHERE (N.I.E.)
+ 3: TRANSPORT
+ 4: TRAVEL
+ 5: CONSTRUCTION

Sex - New Zealand Standard Classification V1.0.0

Overview | Browse | Advanced | Usage | Discussion

1 Male
2 Female

Comparison
EXCEL
PDF
CSV
HTML
SAS
SPSS
STATA
DDI-L
MS-WORD

The Copenhagen Mapping

Current State – SDMX

Reference area code list (2.0)

Description

Represents a combination of reference area codes in M49 and ISO-3611 classifications.

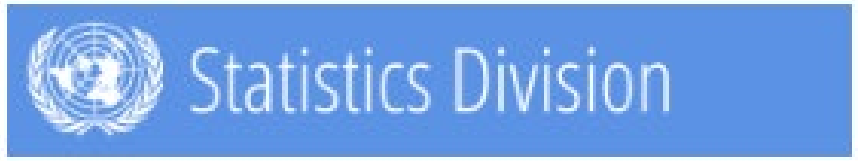
Values

Id	Name
M49	M49 Group +
ISO3166_2	ISO Group −

AF	Afghanistan
AL	Albania
AQ	Antarctica
DZ	Algeria
AS	American Samoa

```
- <mes:Structures>
  - <str:Codelists>
    - <str:Codelist id="CL_AREA" version="2.0" isFinal="true" agencyID="SDMX"
      isExternalReference="false" urn="urn:sdmx:org.sdmx.infomodel.codelist.Codelist=SDMX:CL_AREA
      (2.0)">
        <com:Name xml:lang="en">Reference area code list</com:Name>
        <com:Description xml:lang="en">Represents a combination of reference area codes in M49
          and ISO-3611 classifications.</com:Description>
      - <str:Code id="M49" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).M49">
          <com:Name xml:lang="en ">M49 Group</com:Name>
          <com:Description xml:lang="en">M49 Group</com:Description>
        </str:Code>
      + <str:Code id="000" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).000">
      + <str:Code id="001" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).001">
      + <str:Code id="002" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).002">
      + <str:Code id="003" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).003">
      - <str:Code id="004" urn="urn:sdmx:org.sdmx.infomodel.codelist.Code=SDMX:CL_AREA
        (2.0).004">
          <com:Name xml:lang="en">Afghanistan</com:Name>
          <com:Description xml:lang="en">ISO 3166-2 code: AF</com:Description>
        - <str:Parent>
            <Ref id="034"/>
          </str:Parent>
        </str:Code>
```

Current State – Custom Response Messages in XML/JSON



Moving Forward - Governance (Who & How)

- United Nations – Statistics Division
- United Nations – Economic Commission for Europe
- DDI
- SDMX