On Modeling Linked Open Statistical Data

Evangelos Kalampokis, Dimitris Zeginis, Konstantinos Tarabanis

http://kalampok.is  @kalampokis
Outline

- Motivation
- Objective
- Approach
- Results
- Future Work
Outline

- Motivation
- Objective
- Approach
- Results
- Future Work
Open Statistical Data

- A major part of open data concerns statistics such as demographic, economic, and social indicators.
- We refer to these aggregated data as **Open Statistical Data (OSD)**.
  - This type of data covers also, but is not limited to, official statistics produced and disseminated by National Statistics Institutes or Eurostat.
- OSD can be conceptualized as cubes.
Linked Open Statistical Data

- **Linked Open Statistical Data (LOSD)** refer to the application of Linked Data technologies in Open Statistical Data.
- Many linked data vocabularies enable modeling data cubes as RDF graphs.
  - The RDF Data Cube Vocabulary
  - SKOS – Simple Knowledge Organization System
  - XKOS – Extended Knowledge Organization System
- In addition, controlled vocabularies that define widely used statistical concepts have been created.
  - QUDT units vocabulary
  - Linked data transformation of the SDMX
Official LOSD portals

- Public Authorities and Governments
  - The environmental department of the Flemish government
  - Ministry of Housing, Communities & Local Government (MHCHG) – UK
  - Scottish Government
  - Japanese Government

- National Statistical Institutes
  - Italian National Institute of Statistics (ISTAT)
  - Irish CSO
Problem definition

- Although Web standards (e.g., QB and SKOS) and supporting vocabularies have been created, LOSD creators make a number of modeling decisions in an ad-hoc manner resulting in
  - Fragmented LOSD (silos) that is hard to combine
  - Software tools that cannot be reused across datasets
Outline

- Motivation

- Objective

- Methodology

- Results

- Future Work
Objective

▪ Identify all the modeling questions that a LOSD publisher has to answer in order to be able to create LOSD
  – We refer to these questions as **modeling challenges** related to the creation of LOSD
  
▪ Provide answers to these modeling questions
  – We refer to these answers as **modeling approaches** (or practices) that address the modeling challenges
Outline

▪ Motivation
▪ Objective
▪ Approach
▪ Results
▪ Future Work
Delphi method

- The Delphi method is a process of arriving at **group consensus** by providing experts with rounds of questionnaires, as well as the group response before each subsequent round.

![Diagram](image)

- Experts
- Facilitator

→
Delphi characteristics

- It enables experts to **re-assess** an initial judgment based on the feedback provided by other experts.

- Participants **remain anonymous** to each other even after the completion of the final report.
  - This prevents the domination of some participants (e.g., because of their personality or reputation), which is usually a concern when using group-based methods to collect information.

- Delphi can be continuously iterated until consensus is achieved
  - **Two or three iterations** are often enough to collect the needed information and reach sufficient consensus.
Our approach - Involvement of Experts

- Participation of practitioners that were involved in publishing LOSD
  - for the Scottish and the UK governments
  - for the Flemish government
  - for European Commission’s Digital Agenda
  - for the Italian National Institute of Statistics (ISTAT)
  - of Eurostat, OECD, IMF and World Bank

- Participation of experts that were involved
  - in the creation of the QB vocabulary
  - in the creation of software tools in academia and in industry

- In total 9 experts participated in our study
Our approach

- The study took place in 2017
- Two Delphi rounds that lasted two months each
- **Mesydel** an online service that supports Delphi

Preparatory phase

First questionnaire with initial set of challenges and practices

Delphi round 1

- Areas of disagreement/agreement
- Pros and cons of suggested practices
- More challenges and practices

Second questionnaire based on the collected feedback

Delphi round 2

- Modeling challenges
- Agreed (or best) practices
- Analysis of practices in challenges where consensus was not achieved
Outline

▪ Motivation
▪ Objective
▪ Approach
▪ Results
▪ Future Work
## Modeling Challenges

<table>
<thead>
<tr>
<th>Challenge</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>Defining a measure</td>
<td>What property should be used to model a measure of a cube?</td>
</tr>
<tr>
<td>Defining the unit</td>
<td>Should a cube include the unit of the measure?</td>
</tr>
<tr>
<td></td>
<td>What property should be used to define the unit?</td>
</tr>
<tr>
<td></td>
<td>Where should the unit be defined?</td>
</tr>
<tr>
<td></td>
<td>What values should be used?</td>
</tr>
<tr>
<td>Defining multiple units per measure</td>
<td>Should one cube include multiple units for the same measure?</td>
</tr>
<tr>
<td></td>
<td>Where to define multiple units?</td>
</tr>
<tr>
<td>Defining multiple measures</td>
<td>How to model multiple measures per cube?</td>
</tr>
<tr>
<td>Defining dimension properties</td>
<td>What <em>Properties</em> should be (re-)used for common dimensions?</td>
</tr>
<tr>
<td>Associating dimensions with their values</td>
<td>How to associate a dimension to its values?</td>
</tr>
<tr>
<td>Defining values of common dimensions</td>
<td>What values should be used in time/geo/age related dimensions?</td>
</tr>
<tr>
<td>Modeling single value dimensions</td>
<td>How to model single value dimensions?</td>
</tr>
<tr>
<td>Creating code lists</td>
<td>How to model a new code list?</td>
</tr>
<tr>
<td></td>
<td>How to model hierarchical structures in a code list?</td>
</tr>
</tbody>
</table>
Modeling approach: units

A unit of measure should always be included

sdmx-attribute:unitMeasure should always be re-used

- The unit should be defined at the qb:Observation.
- Use also qb:DataSet to facilitate indexing

URIs from QUDT should be re-used.

sdmx-attribute:unitMeasure

qb:Observation

qb:DataSet

qb:structure

qb:component/qb:measure

qb:component/qb:attribute

sdmx-attribute:unitMeasure

*Re-use URIs from QUDT

qudt:Percent

qudt:EUro

qudt:Kilometer

qudt:Person

:dsd

:cube

:obs1

16.8
### Modeling approach: multiple measures

- QB vocabulary proposes 2 practices for the definition of multiple measures per cube
  - **Multi-measure observations**: define multiple measures in the structure and use all measures in every observation
  - **Measure dimension**: define multiple measures in the structure, restrict observations to having one measure

<table>
<thead>
<tr>
<th>Size of produced cube</th>
<th>Multi-measure observations</th>
<th>Measure dimension</th>
</tr>
</thead>
<tbody>
<tr>
<td>Support of multiple units and measures</td>
<td>Small</td>
<td>Large</td>
</tr>
<tr>
<td>Support of multiple measures with the same unit</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Modeling approach: values of dimensions

The range of a dimension property should always be defined.

If a code list is modeled as skos:ConceptScheme, qb:HierarchicalCodeList, or skos:Collection, then it should be associated with the qb:DimensionProperty using the qb:codeList property.

In addition, the object that is related to the rdfs:range property should be set to skos:Concept.
Modeling approach: single value dimensions

- Some datasets describe a measure using only a single value of a dimension.
  - For example, census data describe multiple measures for a specific year.
- The QB vocabulary enables defining this single value at different levels:
  - Including the single value at the *qb:Dataset*
  - Creating a *qb:Slice* for the single value
  - Including the single value at every *qb:Observation*
Modeling approaches

- In total we have described 19 approaches for modeling LOSD
- Definitions along with a detailed working example can be found in the JWS paper

Journal of Web Semantics

On modeling linked open statistical data

Evangelos Kalampokis \(^a\), \(^b\), Dimitris Zegini\(^a\), \(^b\), Konstantinos Tarabanis \(^a\), \(^b\)

Show more

https://doi.org/10.1016/j.websem.2018.11.002

Get rights and content

Under a Creative Commons license

open access

https://doi.org/10.1016/j.websem.2018.11.002
Outline

- Motivation
- Objective
- Approach
- Results

- Future Work
Open challenges

- Code lists for measures that could be re-used by LOSD publishers.
- The method (e.g. formula) a measure is calculated.
  - For example, unemployment can be calculated based on different methods or different base periods
- Composite measures may be derived from other measures
  - “Unemployment Rate” as ratio of the number of unemployed people in the labor force to the total labor force
- Having explicitly defined which aggregation functions are applicable to a measure is useful for presentation and further processing purposes
  - QB4OLAP provides a solution
- When should we define multiple measures per cube
Interoperability conflicts in LOSD portals

- In (Kalampokis et al., 2019) we have identified 20 interoperability conflicts in LOSD portals
  - 11 of them can be resolved based on the results of this paper
  - 9 of them require further analysis

<table>
<thead>
<tr>
<th>Conflict Name or Type</th>
<th>Status</th>
<th>Resolvable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homonym conflicts</td>
<td>Solved</td>
<td>-</td>
</tr>
<tr>
<td>C1.1 Naming the measure property</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C1.2 Naming the unit of measure property</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.3a Naming the geospatial dimension properties</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.3b Naming the gender dimension properties</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C1.3c Naming the age dimension properties</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.3d Naming the temporal dimension properties</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.4 Naming hierarchical relations</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.5 Populating the unit of measure</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C1.6 Populating the temporal dimension</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C1.7 Populating the gender dimension</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C2.1 Modelling the measure and its parameters</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C2.2 Selecting the number of measures</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C2.3 Modelling the unit of measure and its parameters</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C2.4 Selecting the number of units per measure</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C2.5 Modelling hierarchical levels</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C3.1 Modelling multiple measures</td>
<td>Open</td>
<td>No</td>
</tr>
<tr>
<td>C3.2 Defining the unit of measure in the structure of the cube</td>
<td>Solved</td>
<td>-</td>
</tr>
<tr>
<td>C3.3 Defining dimensions with a single value in the structure of the data cube</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C3.4 Modelling aggregated values</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C3.5 Modelling the values of the temporal dimension</td>
<td>Open</td>
<td>Yes</td>
</tr>
<tr>
<td>C3.6 Associating dimensions to code lists</td>
<td>Open</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Thank you for your attention!!

http://kalampok.is

@kalampokis