Reflections on: Modeling Linked Open Statistical Data

Evangelos Kalampokis[1,2], Dimitris Zeginis[1,2], and Konstantinos Tarabanis[1,2]

[1] University of Macedonia, Information Systems Lab, Egnatia 156, Thessaloniki 54006, Greece ekal@uom.edu.gr, zeginis@uom.gr, kat@uom.edu.gr
[2] Centre for Research & Technology Hellas, Information Technologies Institute, 6th km Xarilaou - Thermi, Thessaloniki 57001, Greece

Abstract. A major part of Open Data concerns statistics such as economic and social indicators. Statistical data are structured in a multidimensional manner creating data cubes. Recently, National Statistical Institutes and public authorities adopted the Linked Data paradigm to publish their statistical data on the Web. Many vocabularies have been created to enable modeling data cubes as RDF graphs, and thus creating Linked Open Statistical Data (LOSD). However, the creation of LOSD remains a demanding task mainly because of modeling challenges related either to the conceptual definition of the cube, or to the way of modeling cubes as linked data. The aim of this paper is to identify and clarify (a) modeling challenges related to the creation of LOSD and (b) approaches to address them. Towards this end, LOSD experts were involved in an interactive feedback collection and consensus-building process that was based on Delphi method. We anticipate that the results of this paper will contribute towards the formulation of best practices for creating LOSD, and thus facilitate combining and analysing statistical data from diverse sources on the Web.

Keywords: Linked Open Statistical Data · Modeling Challenges · Delphi Method.

1 Introduction

International organizations, governments, and companies are increasingly opening up their data for others to reuse [1]. A major part of open data concerns statistics [2] such as demographics, economic, and social indicators. Statistical data are organized in a multidimensional manner, and thus they can be conceptualized as data cubes. These data can be an important primary material for added value services and products, which can increase government transparency, contribute to economic growth and provide social value to citizens [3].

Linked data has been introduced as a promising paradigm for opening up data because it facilitates data integration on the Web [4]. In statistics, linked data enable performing analytics on top of disparate and previously isolated datasets [5]. As a result, many National Statistical Institutes and public authorities have already used the linked data paradigm to publish statistical data on the Web.

Many vocabularies have been created to enable modeling data cubes as RDF graphs. However, the creation of Linked Open Statistical Data (LOSD) remains a demanding task mainly because of modeling challenges related either to the conceptual definition of a cube, or to the way of modeling cubes as linked data. The former regards challenges such as the number of measures or the number of units to include in a cube, while the latter is related to the lack of clarity on the way to apply the proposed vocabularies and the lack of specialized standards. All the above modeling challenges are currently addressed by data publishers in an ad hoc manner, and thus they hinder publishing LOSD in a uniform way that would facilitate their wide exploitation.

The aim of this paper is to identify modeling challenges related to the creation of LOSD and approaches to address them. Towards this end, nine LOSD experts were involved in an interactive feedback collection and consensus-building process. The experts indicated and evaluated modeling challenges and approaches to address them. The goal is to build a consensus on the approaches that can be adopted to address LOSD modeling challenges.

The rest of the paper is organized as follows: Section 2 presents the method that was followed, Section 3 presents the state of the art analysis regarding LOSD standards. Section 4 briefly presents the results of the Delphi method. Finally, Section 5 discusses open challenges and Section 6 summarizes the results.

2 Method

The method employed for the LOSD experts involvement is Delphi [6], which facilitates consensus-building by using a questionnaire with multiple iterations to collect feedback until a stability in the responses is attained. One of the characteristics of Delphi is that participants remain anonymous to each other. This prevents the domination of some participants (e.g., because of their reputation). Delphi can be continuously iterated until consensus is achieved. However, literature has pointed out that two iterations are often enough to reach sufficient consensus [7]. The two rounds of the presented study are the following:

Round 1: Usually the first round uses an open-ended questionnaire. However, we adopted a common modification that uses a structured (aka closed) questionnaire based upon a preparatory phase. The preparatory phase contained: (i) state of the art analysis on the data cube model to identify the main LOSD modeling constructs, (ii) involvement of experts to identify LOSD modeling challenges, and (iii) analysis of LOSD standards to identify approaches related to the modeling challenges. The structured questionnaire asked experts to review, select or rank the initially identified approaches related to the modeling challenges. As a result, areas of disagreement/agreement were identified. The results included advantages/disadvantages of the publishing approaches as well as other publishing approaches not identified at the preparatory phase.

Round 2: The collected feedback of the first round was organized and a second questionnaire was created. This questionnaire was re-structured to be more comprehensive and incorporated the advantages/disadvantages identified at the first round to provide additional insights to the experts. It also contained approaches of the first round in which consensus was achieved so that experts can review them. In every question, the experts were asked to state the rationale behind their choice. The result of Round 2 included all the LOSD modeling challenges, an analysis of the approaches related to these challenges, and all the approaches where consensus was achieved.

The selection of appropriate experts is very important in a Delphi study since it affects the quality of the produced results. Usually, around ten experts are sufficient. In our study, we included 9 experts in the area of LOSD:

- An expert involved in the creation of the LOSD portals of the Scottish Government (`http://statistics.gov.scot`) and the UK Department for Communities and Local Government (`http://opendatacommunities.org`).
- An expert involved in publishing of LOSD for the Flemish Government (`https://id.milieuinfo.be`).
- An expert involved in the creation of the LOSD portal for the European Commission's Digital Agenda (`http://digital-agenda-data.eu/data`).
- An expert involved in the creation of the portal of the Italian National Institute of Statistics (`http://datiopen.istat.it`).
- An expert involved in the creation of the QB vocabulary.
- An expert who created LOSD using data from international organizations such as Eurostat, OECD, IMF and World Bank.
- An expert working at National Institute of Statistics and Economic Studies.
- An expert working in academia.
- An expert working in industry.

The study took place in 2017 and comprised two rounds that lasted two months each. In order to facilitate the process we exploited Mesydel[3] an online service that supports Delphi enabling the participation of multiple experts.

3 Preparatory phase: State of the art analysis

Statistical data usually concern aggregated data monitoring social and economic indicators [8]. They can be described in a multidimensional way, where a measure is described based on a number of dimensions. Thus, statistical data can be conceptualized as a data cube. The data cube model has already been defined in the literature [9, 10], and comprises a set measures which represent numerical values, and dimensions, which provide contextual information. Each dimension comprises a set of values (e.g., "Greece", "France") that can be hierarchically organized into levels (e.g., country, region). The location of each cube's cell is specified by the dimension values, while the value of a cell specifies the measure (e.g., the unemployment rate of "Greece" in "2016" is "23.1%").

A number of linked data standard vocabularies have been proposed to enable the publishing of data cubes. The QB vocabulary [11] is a *W3C* standard for

[3] `https://mesydel.com`

publishing data cubes. The core class of the vocabulary is the *qb:DataSet* that represents a cube, which comprises a set of dimensions (*qb:DimensionProperty*), measures (*qb:MeasureProperty*), and attributes (*qb: AttributeProperty*). Each *qb:DataSet* has multiple *qb:Observation* that describe the cells of the cube.

At LOSD it is a common practice to re-use predefined code lists to populate the dimension values. For example, the values of the time dimension can be obtained from the code list defined by `reference.data.gov.uk` or the values of the unit of measure can be obtained from the QUDT units vocabulary[4]. However, predefined code lists does not always exist, so new should be specified using the QB vocabulary or the Simple Knowledge Organization System (SKOS) [12] or the Extended Knowledge Organization System (XKOS) [13].

Finally, a UK Government Linked Data Working Group[5] has developed a set of common dimensions (*timePeriod*, *refArea*, *sex*, *age*), measures (*obsValue*) and attributes (*unitMeasure*) that are intended to be reusable across data sets. The definition of these concepts is based on the SDMX guidelines.

All the above standard vocabularies facilitate the publishing of LOSD. However, in some cases there is lack of clarity on how to apply these standards because they allow the adoption of different valid publishing approaches.

4 Delphi Results: Challenges and Approaches

This section briefly presents the results of the Delphi method. Tables 1 and 2 contain all the identified LOSD modeling challenges and the approaches where consensus was achieved among the experts. The following paragraphs elaborate on some challenges and approaches that need further clarification.

At measure definition (Ch1), a common property is *sdmx-measure:obsValue*. However, experts indicated that it should not be used because defining a measure as sub-property of *sdmx-measure:obsValue* is redundant. It does not add any additional semantics than defining the measure as a *qb:MeasureProperty*.

Regarding the definition of unit (Ch2.3) the QB vocabulary enables different levels, i.e., the *qb:DataSet*, the *qb:MeasureProperty*, and the *qb:Observation*. The level *qb:DataSet* or the *qb:MeasureProperty* facilitates the retrieval of units directly from the structure of the cube. While the level *qb:MeasureProperty* or the *qb:Observation* enables the definition of multiple units at one cube. The *qb:Observation* level enables observation to be re-used at another context since they contain all relevant information. Expert proposed to use a hybrid approach and define the unit both at *qb:Observation* and *qb:DataSet* if needed.

The QB vocabulary proposes two practices for the definition of multiple measures per cube (Ch4): i) "multi-measure observations" that define multiple *qb:MeasureProperty* in the cube structure and use all measures in every observation and ii) "measure dimension" that defines multiple *qb:MeasureProperty* at the structure, but restrict observations to having a single measure. The first approach produces smaller in sizes cube but cannot represent multiple units and

[4] `http://qudt.org/`

[5] `https://github.com/UKGovLD/publishing-statistical-data`

Table 1. Challenges and approaches

	Challenges	**Approaches**
Measure	**Ch1:**What property should be used to model a measure of a cube?	**Ap1:** A new measure property should be defined that is not sub-property of sdmx-measure:obsValue. The new measure enables the annotation with additional properties (e.g., labels, comments).
Unit	**Ch2.1:**Should a cube include the unit of the measure?	**Ap2.1:** A unit of measure should always be included in the cube. The measure on its own is a plain numerical value and thus unit is required to correctly interpret this value.
	Ch2.2: What RDF property should be used to define the unit?	**Ap2.2:** sdmx-attribute:unitMeasure should always be re-used to define units. This property can be used directly to assign values that are not part of a code list (e.g., QUDT). However, when annotation with additional properties (e.g., labels, codelist, etc.) is required, then new units that are sub-properties of sdmx-attribute:unitMeasure should be defined.
	Ch2.3: Where should the unit be defined?	**Ap2.3:** The unit should be defined at the qb:Observation. The unit can be additionally defined at the qb:DataSet in order to facilitate the retrieval of the available units in a cube.
	Ch2.4: What values should be used for the units?	**Ap2.4:** URIs from QUDT should be re-used. If QUDT is not sufficient, then DBpedia or other code lists can be used.
Multiple units	**Ch3.1:** Should one cube include multiple units for the same measure? **Ch3.2:** Where to define multiple units?	**Ap3:** One cube with multiple units should be created and the unit should be defined at each *qb:Observation*. Conceptually, it is preferable to have all related units of the same measure in the same cube. The unit can be additionally defined at the qb:DataSet in order to facilitate the retrieval of the available units in a cube.
Multiple measures	**Ch4:** How to model multiple measures per cube?	**Ap4.1:** If the data have multiple measures, then it is common to publish cubes with multiple measures only when measures are closely related to a single observational event (e.g. sensor network measurements). However, the approach to be followed is up to the data cube publisher. In case of modeling multiple measures in multiple cubes with one measure each, then Ap2 (if the measures have one unit) and Ap3 (if the measures have multiple units) should be followed. **Ap4.2:** In case of modeling multiple measures in one cube then the measure dimension approach (i.e. observations with a single measure) should be followed and the unit should be defined in each observation (see Ap 3).
Dimension	**Ch5:** What *rdf:Properties* should be used for common dimensions?	**Ap5.1:** If a dimension refers to time, geography, or age, then a new qb:DimensionProperty should be defined. This new qb:DimensionProperty should be also defined as rdfs:subPropertyOf the corresponding SDMX dimension. For example, a geospatial dimension of a cube should be defined as sub-property of sdmx-dimension:refArea. **Ap5.2:** If a dimension refers to gender, then sdmx-dimension:sex should be reused provided that the associated code list addresses the modeling needs, e.g., more notions of sex such as hermaphroditism, transgender, and asexual are not needed. Otherwise, a new dimension should be defined along with a controlled vocabulary.

Table 2. Challenges and approaches

	Challenges	**Approaches**
Dim. values	**Ch6:** How to associate a dimension to its values?	**Ap6.1:** The rdfs:range of a qb:DimensionProperty should always be defined. **Ap6.2:** If a code list is modelled 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.
Common dimension values	**Ch7.1:** What values should be used in time related dimensions?	**Ap7.1a:** In case of a specific point in time a new dimension should be defined. This dimension should be rdfs:subPropertyOf sdmx-dimension:refPeriod and have rdfs:range xsd:dateTime. **Ap7.1b:** In case of a period of time, a new dimension should be defined. This dimension should be rdfs:subPropertyOf sdmx-dimension:refPeriod and have rdfs:range the interval:Interval class of the `http://reference.data.gov.uk`, which uses this class to define years. However, if the approach of `http://reference.data.gov.uk` is not sufficient, then new code lists can be also created and used .
	Ch7.2: What values should be used in geospatial dimensions? **Ch7.3:** What values should be used in age related dimensions?	**Ap7.2:**In case of a geography or age a new dimension should be defined. This dimension should be rdfs:subPropertyOf the sdmx-dimension:refArea or sdmx-dimension:age respectively. The rdfs:range and/or qb:codeList of this dimension should be defined as described in Ap6.2. If a code list or reference dataset that addresses the modeling needs exists, then it should be re-used. Otherwise, a new code list should be created.
Single	**Ch8:** How to model single value dimensions?	**Ap8:** A single value dimension should be always included in all observations of the cube
Code list	**Ch9.1:** How to model a new code list?	**Ap9.1:** A code list should be modelled using SKOS. This is also suggested by the QB vocabulary. Specifically, individual code values should be modelled using skos:Concept and the overall set of values should be modelled using skos:ConceptScheme or skos:Collection. Always define a separate code list for each distinct set of values (e.g., age groups and geographical areas).
	Ch9.2: How to model hierarchical structures in a code list?	**Ap9.2:** In case of hierarchical data, hierarchical code lists should be always used to describe them. SKOS should be preferred when the hierarchies are simple. In case where the hierarchical levels are fully separated and depth is a meaningful concept then XKOS is appropriate. Finally, when there is a need to express more relations that are not covered by SKOS or XKOS (e.g., administeredBy in contrast to within) then the QB vocabulary should be preferred.
	Ch9.3: Should aggregate values be included as dimension values?	**Ap9.3:** Aggregate values (e.g., Total) should be included in a dimension if the measured variable in this dimension can be aggregated. The aggregate value should be modelled on the top a hierarchy.

measures while the second approach enables the definition of multiple units and multiple measures. Experts proposed the use of the "measure dimension".

The association of a dimension to its potential values (Ch6) can be achieved using two complementary approaches: i) use the property *rdfs:range* to define the class of the values of a *qb:DimensionProperty* and ii) use the property *qb:codeList* to associate a *qb:DimensionProperty* with a code list. Experts proposed to use always the *rdfs:range* and the *qb:codeList* when a code list is available.

Some datasets describe a measure using only a single value of a dimension (Ch8) e.g. census data describe measures for a specific year. The QB vocabulary enables defining this single value at different levels: i) *qb:Dataset*, ii) *qb:Slice* and iii) *qb:Observation.* The first does not enable future addition of observations with a different value for that particular dimension while the second imposes an extra burden of defining *qb:Slices.* The last approach is proposed by the experts since it enables the addition of observations with different dimension values in the same dataset and the easy re-use of *qb:Observations* at another context. This approach has, however, the cost of an increased number of triples.

5 Open challenges

During the Delphi process experts indicated a number of open challenges. These open challenges regard, limitations of existing standards, lack of standards and modeling decisions. An important challenge is the definition of code lists for measures that could be re-used by LOSD publishers. Currently each publisher defines its own *qb:MeasureProperty* for the same measure (e.g. *p1:unemployment* and *p2:unemployment*). The definition of a "standard" code list would enable the publishing of LOSD in a uniform way, thus facilitating the integration and combination of related statistical data from different sources. Another challenge is related with the method a measure is calculated. For example, unemployment can be calculated based on different methods or different base periods. In this case there is a discussion whether to use the same *qb:MeasureProperty* or not.

Composite measures may be derived from other measures [14, 15] e.g. "Unemployment Rate" as ratio of the number of unemployed people to the total labour force. This relation should somehow be expressed by linking the two measures. In this case the computation of aggregated "total" values for composite measures would also be possible. For example the computation of the "total unemployment rate" based on "male" and "female" unemployment rate. Currently, there is no LOSD standard to express these relations. Additionally, having explicitly defined which aggregation functions (e.g. sum, average) are applicable to a measure is useful for further processing purposes. QB4OLAP [16] proposes an extension of QB vocabulary to express aggregation function. The applicability of aggregate functions to measures depends on various factors [17] (e.g. cube dimensions, units) and needs to be further explored.

A modeling challenge is related to the definition of multiple measures at a cube. Our study has shown that cubes with multiple measures should be published only when measures are closely related to a single observational event (e.g. sensor measurements). If the measures are independent then they should be modelled at separate cubes. However, there's a large grey area between the two since the "observational event" is not clearly defined.

Finally, there are some challenges related with the performance of applications that consume LOSD. For instance, the use of the *qb:codeList* indicates all the potential values of a *qb:DimensionProperty*. However, it is common to not use all the values at the cube e.g. a code list may contain values for the geography of Europe, but the cube uses only values for Greece. In this case there is no way to retrieve only the used values from the cube structure. They can be only retrieved by demanding SPARQL queries that iterate over all the cube observations. The same case also applies to units of measure.

6 Conclusion

A major part of open data concern statistics. Recently many National Statistical Institutes and public authorities have adopted the linked data paradigm to publish LOSD since it facilitates data integration on the Web. Towards this direction many standard vocabularies have been proposed (i.e., QB, SKOS, XKOS).

The publication of high quality LOSD can be an important primary material for added value services, which can increase government transparency, contribute to economic growth and provide social value. However, the creation of LOSD remains a demanding task because of modelling challenges. These challenges are usually addressed by data publishers in an ad hoc manner thus hindering the publishing of LOSD in a uniform way and lead to the creation of LOSD silos. As a result LOSD from different sources cannot be easily integrated and generic software tools cannot be developed.

Towards this direction, experts that directly participate at the publishing of LOSD, are involved through an iterative approach in order to comprehend the modeling challenges, identify relevant publishing approaches and propose ways to address these challenges. The result is a set of proposed approaches that support LOSD publishers to model their data and to apply common standards. However a set of open challenges related with the limitations of existing standards, lack of standards and modeling decisions still remain to be explored.

We anticipate that the analysis of the modelling challenges as well as the proposed approaches presented at this paper will trigger and contribute towards a discussion on the development of best practices for publishing LOSD, facilitating the combining and analysing of linked statistical data from diverse sources.

Acknowledgement

This paper is an extended abstract of [18] published at the Journal of Web Semantics. Part of this work was funded by the European Commission within the H2020 Programme in the context of the project OpenGovIntelligence under grant agreement no. 693849. The authors would like to cordially thank all the experts who participated in the study.

References

1. E. Kalampokis, E. Tambouris, K. Tarabanis, A classification scheme for open government data: towards linking decentralised data, Int. J. Web Eng. Technol 6 (3) (2011) 266—285.
2. S. Capadisli, S. Auer, A.-C. Ngonga Ngomo, Linked sdmx data, Semantic Web 6 (2).
3. E. Kalampokis, E. Tambouris, A. Karamanou, K. Tarabanis, Open statistics: The rise of a new era for open data?, in: H. J. Scholl, O. Glassey, M. Janssen, B. Klievink, I. Lindgren, P. Parycek, E. Tambouris, M. A. Wimmer, T. Janowski, D. Sá Soares (Eds.), Electronic Government, Springer International Publishing, Cham, 2016, pp. 31–43.
4. C. Bizer, T. Heath, T. Berners-Lee, Linked data-the story so far, Semantic Services, Interoperability and Web Applications: Emerging Concepts (2009) 205–227.
5. E. Kalampokis, E. Tambouris, K. Tarabanis, Linked open cube analytics systems: Potential and challenges, IEEE Intelligent Systems 31 (5) (2016) 89–92.
6. C.-C. Hsu, B. A. Sandford, The delphi technique: making sense of consensus, Practical assessment, research & evaluation 12 (10) (2007) 1–8.
7. R. L. Custer, J. A. Scarcella, B. R. Stewart, The modified delphi technique: A rotational modification., Journal of Vocational and Technical Education 15 (2) (1999) 1 – 10.
8. R. Cyganiak, M. Hausenblas, E. McCuirc, Official statistics and the practice of data fidelity (2011) 135–151.
9. F. S. Tseng, C.-W. Chen, Integrating heterogeneous data warehouses using xml technologies, Journal of Information Science 31 (3) (2005) 209–229. doi:10.1177/0165551505052467.
10. S. Berger, M. Schrefl, From federated databases to a federated data warehouse system, in: Proceedings of the 41st Annual Hawaii International Conference on System Sciences, IEEE, 2008, pp. 394–394.
11. R. Cyganiak, D. Reynolds, The rdf data cube vocabulary:w3c recommendation, 2014.
12. A. Miles, S. Bechhofer, Skos simple knowledge organization system reference: W3c recommendation, 2009.
13. R. Cyganiak, D. Gillman, R. Grim, Y. Jaques, W. Thomas, Xkos: An skos extension for representing statistical classifications, 2017.
14. S. F. Pileggi, J. Hunter, An ontological approach to dynamic fine-grained urban indicators, Procedia Computer Science 108 (2017) 2059 – 2068, international Conference on Computational Science, ICCS 2017, 12-14 June 2017, Zurich, Switzerland.
15. M. Denk, W. Grossmann, Towards a best practice of modeling unit of measure and related statistical metadata, IMF working paper.
16. J. Varga, A. A. Vaisman, O. Romero, L. Etcheverry, T. B. Pedersen, C. Thomsen, Dimensional enrichment of statistical linked open data, Journal of Web Semantics 40 (2016) 22 – 51. doi:http://dx.doi.org/10.1016/j.websem.2016.07.003.
17. E. Kalampokis, E. Tambouris, K. Tarabanis, ICT tools for creating, expanding, and exploiting statistical linked open data, Statistical Journal of the IAOS 33 (2) (2017) 503–514.
18. E. Kalampokis, D. Zeginis, K. Tarabanis, On modeling linked open statistical data, Journal of Web Semantics 55 (2019) 56 – 68. doi:https://doi.org/10.1016/j.websem.2018.11.002.