A Quantitative Survey on the Use of the Cube Vocabulary in the Linked Open Data Cloud

Karin Becker[1], Shiva Jahangiri[2], and Craig A. Knoblock[2]

[1] Instituto de Informática - Universidade Federal do Rio Grande do Sul, Brazil
karin.becker@inf.ufrgs.br
[2] Information Sciences Institute, University of Southern California, USA
{shivajah,knoblock}@usc.edu

Abstract There is a striking increase in the availability of statistical data in the Linked Open Data (LOD) cloud, and the Cube vocabulary has become the *de facto* standard for the description of multi-dimensional data. However, the reuse of a standard vocabulary needs to pair with modeling strategies that make it easy to locate, consume and integrate information. In this paper, we developed a quantitative study on how the main concepts of the Cube vocabulary are applied in practice, using the governmental datasets identified in the 2014 LOD cloud census. Our focus was on the commonly used strategies for multi-dimensional modeling using Cube because they have an impact on the automatic location and consumption of data. The results provide feedback to work that addresses Cube usage and establish a baseline for evolution comparison.

Keywords: LOD, Cube, multidimensional data, quantitative analysis

1 Introduction

Statistical data is used as the foundation for policy prediction, planning and adjustments. There is a growing consensus that the Linked Open Data (LOD) cloud is the right platform for creating, locating and integrating heterogeneous and distributed open datasets for a myriad of analysis purposes [3,4,7,9,10]. However, the LOD cloud as a primary source for integrated data can only scale if there is a strong commitment to its basic principles in terms of vocabulary reuse, linking and metadata provision. A study that compared several vocabulary reuse strategies [12] concluded that practitioners favor the reuse of a single, popular vocabulary.

The Cube vocabulary [11] is a W3C recommendation for publishing multidimensional data in the web. It establishes that *datasets* contain *observations* about *measures*, according to one or more *dimensions*. A *data definition structure* (DSD) explicitly describes the structure and semantics of the observations in a dataset. Although not restricted to statistical data, it was designed to be compatible with statistical ISO SDMX standard. The last LOD census [13] revealed that the Cube vocabulary was adopted in 61.75% of the datasets in the government government domain, and several projects focus on infrastructure for using, publishing, validating and visualizing cube datasets [5,7,9,8].

However, the reuse of a standard vocabulary needs to pair with modeling strategies that make it easy to locate, consume and integrate information in the

LOD cloud. The use of the Cube vocabulary is complex, as it involves mastering both the multidimensional modeling paradigm, and the vocabulary itself. As a result, cube datasets available in the LOD cloud are very diverse [3,4,9,10]. Cube publishing platforms [5,7,9] are targeted at the correct and complete usage of the vocabulary, but until a thorough discussion on the best patterns is developed, publishers lack methodological support for making modeling decisions that influence how easily cube datasets can be located in the LOD cloud and consumed.

In this paper, we present a quantitative survey on the usage of the Cube vocabulary considering governmental datasets identified in the last LOD census. The survey focuses on the commonly used strategies for modeling multi-dimensional data according to the Cube, because they affect how data can be found and consumed automatically. Among the results, we found that: a) many cube datasets are not well-formed (e.g. 35% of DSDs do not define measures correctly); b) despite many dimensions represent the exact same concept and values (e.g. Year), they are seldomly reused across different organizations, are defined using the same domain or are inter-linked; c) there is a strong influence of SDMX modeling, but which is not adequately captured using Cube constructs; etc. The analysis is reproducible, and datasets/procedures are available in a public repository.

The main contributions of this survey are: a) it provides a detailed analysis of various ways Cube vocabulary is used in practice and guidance on the most useful representations, and b) it serves as a baseline for comparison with the evolution of Cube usage. Cube is a recent W3C recommendation, and its usage will evolve as more examples are available, the trade-offs of each modeling choice are better understood from different perspectives (i.e. publisher, consumer), and comprehensive and innovative supporting platforms become available. The results are also relevant to all works addressing Cube usage (e.g. methodological support, supporting platforms [5,7,9], cube discovery [2])

The rest of this paper is organized as follows. Section 2 describes the key terms of the Cube vocabulary. Section 3 discusses the trade-offs of the most common modeling strategies for modeling measures as Cube multi-dimensional data. Section 4 describes the GQM analysis framework, and discusses the results. Section 5 discusses related work, and Section 6 draws conclusions.

2 Cube Vocabulary

The key terms of the Cube vocabulary are depicted in Fig. 1, adapted from the specification [11]. A DSD (qb:DataStructureDefinition) defines the structure of a dataset. A dataset (qb:DataSet) must be linked to one (and only one) DSD, and each observation (qb:Observation) to one (and only one) dataset. Observations provide values to the dimensions, measures and attributes, according the corresponding DSD. Integrity constraints define rules for well-formed cubes.

Much of the information in a DSD is implicit within the observations, but its explicit declaration has several benefits [11]: a) verification that a dataset matches an expected structure; b) reuse in the publication process; c) simplification and confidence for data consumption; among others.

A DSD aggregates components (qb:ComponentSpecification), which in turn reference properties. A qb:ComponentProperty is an abstract class that encapsulates orthogonal pieces of information, namely: a) the nature of the component

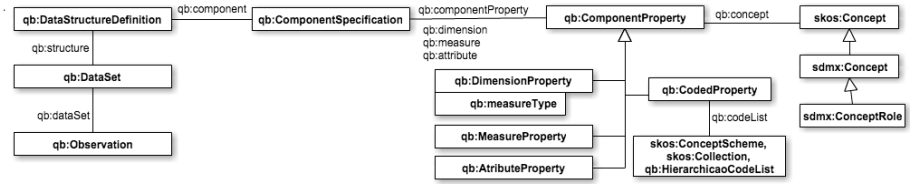

Figure 1: Cube key terms and relationships

(subclasses qb:DimensionProperty, qb:MeasureProperty and qb:AttributeProperty); b) the concept being represented (qb:concept); and c) the type of its values. The type can be defined through rdfs:range, mandatory in the case of dimensions, or through a qb:codeList connected to one of the classes depicted in Fig. 1. Common statistical concepts and associated code lists used across statistical datasets are defined by the SDMX standard as a set of content oriented guidelines (COG), for which corresponding RDF encodings were created. They are not part of the Cube specification, but in practice, they are widely used.

A DSD defines at least one measure. Two approaches are available for multiple measures [11], illustrated in Fig. 2: a) *multi-measure* (ST2) and b) *measure dimension* (ST3). In the former, typical in Business Intelligence and OLAP applications, each observation contains values for each measure defined in the DSD. The second approach is derived from the SDMX information model. To use this representation, the DSD also specifies different measures, and an implicit dimension (qb:measureType) plays the role of the "measure dimension". The dataset can contain different measures, but each observation is related to a single measured value, according to the value of the measure dimension.

3 Modeling Strategies

Cube provides a standard vocabulary, with degrees of freedom that result in different styles of multidimensional modeling, possibly influenced by diverse backgrounds. Let us consider two examples from the peace-building domain. FAO[1] collects and publishes open data on food security, which involves 21 indicators about health, transportation, economics, etc. FFP[2] defines a Fragile State Index (FSI), which ranks the vulnerability of countries according to 14 social, economical and political indicators. These are annual indicators that refer to a specific country. FSI indicators are consistently collected, but food security data is quite sparse.

Fig. 2 shows DSD excerpts according to 3 modeling options: single-measure **(ST1)**, multi-measure **(ST2)** and measure dimension **(ST3)**. For FSI, all strategies are valid. For food security, strategy ST2 is not valid because measures cannot be optional in well-formed cubes. According to ST1, there will be one DSD per indicator. The use of ST2 or ST3 will result in DSDs that contain several measures.

Fig. 2 also shows excerpts for measure and dimension properties. The dimension fao:refArea represents the location, as indicated by the related concept (sdmx-concept:refArea). Two reuse strategies are exemplified involving time dimension (fao:refPeriod). First, it is defined as a sub-property of sdmx-dimension:refPeriod,

[1] Food and Agriculture Organization: faostat3.fao.org
[2] Fund for Peace (FFP): global.fundforpeace.org

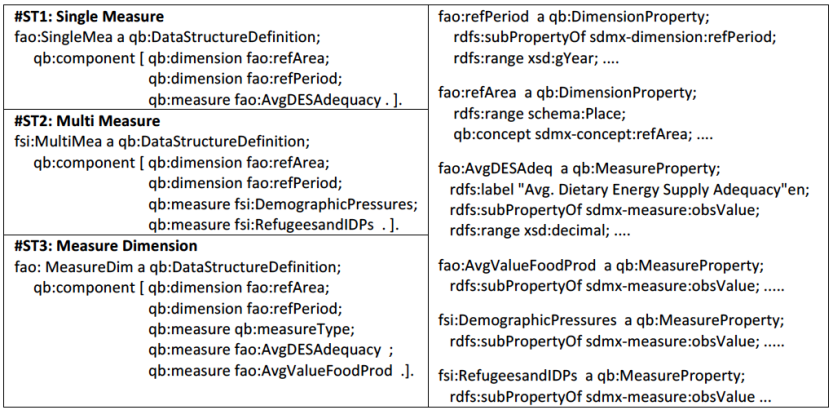

Figure 2: Modeling strategies according to measures.

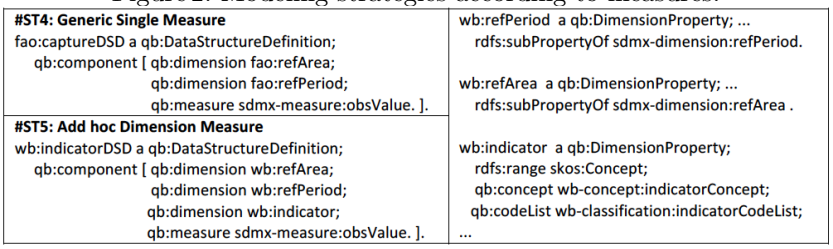

Figure 3: Modeling strategies for generic single measures.

which in turn is related to the SDMX time period concept. It also adds the property range. Second, FSI DSDs reuse fao:refPeriod dimension "as-is".

Fig. 3 illustrates two additional modeling patterns[3], frequently adopted when SDMX descriptions are automatically converted (e.g. [4]). Compared to ST1 in Fig. 2, there are some subtle differences. First, the measure is generic in both cases (sdmx:obsValue), corresponding to a statistical concept valid in any domain. In the case of fao:CaptureDSD **(ST4)**, the conceptual context is provided by the DSD title ("Capture Fisheries"). In the case of wb:indicatorsDSD **(ST5)**, the dimension wb:indicator is meant to represent the specific measure in an observation (a "measure dimension"), where an associated codelist describes the possible measures. In that case, the conceptual context is partially provided by the codelist, but only the presence of a code within a dataset provides information for the available measures.

Although strategies ST4 and ST5 result in correct and well-formed cubes, they do not benefit from constructs designed to represent a measure dimension (qb:measureType). Standard vocabulary and standard usage of vocabulary increase the ability to find, understand and consume data automatically in the LOD.

In terms of semantic expressiveness, ST1-ST3 are the most expressive ones in that the most relevant information needed to identify, understand, and consume datasets is explicit in the DSD. Additionally, measures and dimensions can be associated to both domain and statistical concepts, such that interlinking and inferences can be done automatically. The other two modeling strategies imply

[3] These excerpts are simplifications of real DSDs describing FAO and Worldbank data, available at fao.270a.info/Sparql and worldbank.270a.info/Sparql.

discovering in an *ad hoc* manner which constructs hide the intended semantics (e.g. a DSD/dataset label or description). fao:captureDSD can virtually represent any measure indexed by time and location. wb:indicatorDSD aims at representing the ST3 strategy, but without the appropriate constructs. If one is searching for datasets that contain demographic pressure measures, the answer is straightforward by querying the DSDs of Fig. 2. For the other ones, the DSD provides limited explicit information: it would require identifying that a dimension has a particular role (the measure dimension), finding in the respective codelist if one of the codes represents the desired measure, and locating the relevant datasets by searching the observations that contain that code. Building generic applications to consume data would have to deal with all these issues.

Strategies ST1-ST3 also make it easy to verify if datasets are conformant with regard to the respective DSD. ST4 provides no means for verifying such conformance. In ST5, that capacity is limited: the codelist can be used, but a change on the codelist itself, without concerns for retro-compatibility, can affect the meaning of a DSD, and possibly impact existing consumers. For instance, the removal of a code from the codelist would change the semantics of the DSD, but it will be harder to notice such a change compared to the removal of a measure.

4 Quantitative Analysis

This survey aims at quantifying the usage of the Cube vocabulary in terms of practices concerning the Data Structure Definition (DSD), reuse of dimensions and measures, and conceptual annotation. The evaluation of the current usage provides feedback to other works (e.g. discussion on the best practices, methodological support, Cube supporting platforms, cube discovery), as well as establish a baseline for comparing with usage evolution. The quantitative analysis is reproducible, since all datasets used, extraction procedures, and queries used to compute the metrics are available in a public repository[4].

4.1 Scope

The analysis framework was defined according to the Goal-Question-Metric (GQM) approach [1], which defines a measurement model at three levels. The *conceptual level* is represented by the Goal of the measurement, stated in terms of five components: entity, purpose, focus, point of view and context. The second level is *operational*, where a set of Questions define models of the object of study, in order to characterize the assessment or achievement of a specific goal. Finally, the *quantitative level*, defines a set of Measures that enable to answer the questions in a measurable way. We defined three goals for this survey:

– **Goal 1: Analyze DSD and Datasets for the purpose of *understanding* with respect to *DSD relevance and reuse* from the point of view of the *publisher* in the context of the *LOD cloud*.** Defining a DSD for each dataset is a basic rule for well-formed cubes, and benefits are claimed with regard to DSD explicit information [11], such as discoverability, conformance checking, and reuse. We aim at verifying whether publishers do agree with these benefits, by declaring and reusing DSDs or DSD properties.

[4] https://github.com/KarinBecker/LODCubeSurvey/wiki

- **Goal 2: Analyze *DSD* for the purpose of *understanding* with respect to *modeling strategy* from the point of view of the *publisher* in the context of the *LOD cloud*.** The way measures and dimensions are explicitly and implicitly declared in the DSD influences the discoverability, understanding and processing of cube datasets, as discussed in Section 3. We aim at discovering how frequent is each modeling strategy, and how easy it is to identify hidden semantics about measures and dimensions.
- **Goal 3: Analyze *DSD* for the purpose of *understanding* with respect to *DSD conceptual enrichment* from the point of view of the *publisher* in the context of the *LOD cloud*.** Semantics are added to DSD properties through relationships to concepts. We aim at verifying whether publishers practice semantic annotation.

4.2 Operations

Context Selection. We selected the LOD cloud as reported by the Mannheim Catalogue[5], because it contains the most recent compilation of Linked Datasets (Aug. 2014), and it encompasses all entries from Datahub.io Catalogue[6]. Relevant entries were identified using tags *format-cube*.

Questions and Metrics. Questions and respective metrics are summarized in Table 1. Names of the metrics are self-describing, where direct metrics are represented by absolute numbers, and indirect metrics, by percentages. For example, M1 is the number of datasets (nbDatasets), M2 is the number of datasets with a related DSD (nbDatasetsWithDSD), and M2% refers to the proportion M2/M1.

With regard to Goal 1, our expectation is that all datasets are related to a DSD, as it is a basic rule in well-formed cubes. Two types of reuse are considered. NbReusedDSDs corresponds to DSDs related to more than one dataset (M6). Similarly, reused dimensions/measures in DSD (M7 and M8) concern dimension/measure properties used in more than one DSD description. We also examine the reuse through the definition of new dimensions/measures as sub-properties of existing ones (M9 and M10). Metrics M11 and M12 aim to reveal the most reused dimensions/measure properties, regardless the method. Both types of reuse demonstrate the interest in defining standards for widely used properties present in statistical datasets (e.g. temporal and geographical dimensions).

The metrics for Goal 2 aim at quantifying each modeling strategy described in Section 3. Metrics M13-15 are focused on strategies ST2, ST3 and ST1, respectively. M16 evaluates DSDs with a single generic measure (ST4 or ST5). M17 measures the use of strategy ST5, which is hard to identify with precision, and we searched for different patterns to identify this situation (M18).

The metrics concerning Goal 3 focus on determining whether component properties (dimensions and/or measures) are qualified by concepts (M19-22), and to which extent these are standard SDMX concepts (M23-25). Metrics M26 and M27 study DSDs that include at least one component property qualified by a concept, and M28 reveals the most frequent concepts used.

[5] http://linkeddatacatalog.dws.informatik.uni-mannheim.de/
[6] datahub.io catalog

Table 1: Questions and Metrics

Goal 1: Datasets and DSDs with respect to relevance and reuse	
Q1: Do all datasets have a corresponding DSD?	M1: NbDatasets M2 : NbDatasetWithDSD (M2%=M2/M1)
Q2: Are DSDs, dimensions and measures reused?	M3: NbDSDs M4 : NbReusedDSDs (M4% = M4/M3) M5: NbDimensionProp M6: NbMeasureProp M7 : NbReusedDimensionPropInDSD (M7%=M7/M5) M8 : NbReusedMeasurePropInDSD (M8%=M8/M6) M9 : NbReusedDimensionSubProperty (M9 %=M9/M5) M10 : NbReusedMeasureSubProperty (M10%=M10/M6) M11: TopReusedDimensionProp M12:TopReusedMeasureProp
Goal 2 : DSD with respect to modeling strategy	
Q3: How many DSDs apply the multi-measure strategy (ST2)?	M13 : nbDSDsWithMultipleMeasures (M13%=M13/M3)
Q4: How many DSDs adopt the measure dimension strategy (ST3)?	M14 : nbDSDsWithMeasureDimensionApproach (M14%=M14/M3)
Q5: How many DSDs define a single measure (ST1 and ST4/ST5)?	M15 : nbDSDsWithSingleDomainMeasure (M15%=M15/M3) M16 : nbDSDsWithSingleGenericMeasure (M16%=M16/M3)
Q6: How many DSDs with a single measure contain a dimension representing measures (ST5)?	M17 : nbDSDsWithDimensionReprMeasure (M17%=M17/M3) M18: TopStrategiesDimensionRepresentingMeasure
Goal 3 : DSD with respect to conceptual enrichment	
Q7: Do publishers relate component properties to concepts for conceptual annotation?	M19 NbComponentProp M20: NbCompPropRelatedToConcept (M20%=M20/M19) M21: NbDimPropRelatedToConcept (M21%=M21/M5) M22: NbMeasurePropRelatedToConcept (M22%=M22/M6) M23: NbCompPropRelatedToSDMXConcept (M23%=M23/M19) M24: NbDimPropRelatedToSDMXConcept (M24%=M24/M5) M25: NbMeasurePropRelatedToSDMXConcept (M25%=M25/M6) M26: NbDSDsComPropRelatedToConcept (M26%=M26/M3) M27: NbDSDsComPropRelatedToSDMXConcept (M27%=M27/M3) M28: TopPopularConcepts

Data Collection Procedures. Data was collected through the months of March and April 2015. We developed crawlers to extract the entries from the Mannheim Catalogue, resulting in 114 entries. Then, we created a local repository to integrate DSD descriptive data extracted from all the entries that were operational [7]. We queried for resources of all the classes depicted in Fig. 1, except for qb:Observation. We were also concerned by resources related to the cube properties, as well as properties from other vocabularies (e.g. rdfs:subPropertyOf, rdfs:range). For this purpose, we developed a program that iterates through a set of Sparql endpoints, issues a set of queries that select and constructs the resources and properties of interest as triples, and save the results in a repository (Open RDF). As a result, we extracted data about 16,563 datasets and 6,847 DSDs.

Measuring Operations. We developed Sparql queries to collect each one of the direct metrics in Table 1, and the indirect measures were derived using a spreadsheet, according to the corresponding formulae. Considering the large dominance of a single publisher (Linked Eurostat, with 6,539 DSDs), we decided to run our measuring queries considering three scenarios: a) all non-Eurostat data; Linked Eurostat data only; and combined data. In this way, in the analysis of the

[7] The accompanying material in the Github wiki describes all issues faced during data extraction, and how we handled them.

Table 2: Metrics results

Goal	Question	Metric	Measure	Non-Eurostat		Eurostat		Both	
				Count	%	Count	%	Count	%
G1	Q1	M1	nbDatasets	9,997		6,539		16,536	
		M2	nbDataSetsWithDSD	9,724	97.3%	6,539	100%	16,263	98.3%
	Q2	M3	nbDSDs	309		6,538		6,847	
		M4	NbReusedDSD	11	3.6%	1	0%	12	0.2%
		M5	NbDimensionProp	538		506		1,044	
		M6	NbMeasureProp	163		1		163	
		M7	NbReusedDimensionPropInDSD	191	35.5%	447	88.3%	638	61.1%
		M8	NbReusedMeasurePropInDSD	31	19%	1	100%	32	19.6%
		M9	NbReusedDimensionSubProperty	4	0.7%	0	0%	4	0.4%
		M10	NbReusedMeasureSubProperty	1	0.6%	0	0%	1	0.6%
G2	Q3	M13	nbDSDsWithMultipleMeasures	54	17.5%	0	0%	54	0.8%
	Q4	M14	nbDSDsWithMeasureDimensionApproach	0	0%	0	0%	0	0%
	Q5	M15	nbDSDsWithSingleDomainMeasure	0	0%	0	0%	0	0%
		M16	nbDSDsWithSingleGenericMeasure	244	79%	6,538	100%	6782	99.1%
	Q6	M17	nbDSDsWithDimensionReprMeasure	33	10.7%	2,233	34,2%	2266	33.1%
G3	Q7	M19	NbComponentProp	701		509		1209	
		M20	NbCompPropRelatedToConcept	411	58.6%	507	99,6%	916	75.8%
		M21	NbDimPropRelatedToConcept	395	73.4%	506	100%	901	86.3%
		M22	NbMeasurePropRelatedToConcept	16	9.8%	1	100%	16	9.8%
		M23	NbCompPropRelatedToSDMXConcept	44	6.3%	2	99.6%	45	75.8%
		M24	NbDimPropRelatedToSDMXConcep	36	6.7%	1	0.2%	37	3.5%
		M25	NbMeasurePropRelatedToSDMXConcept	8	4.9%	1	100%	8	4.9%
		M26	NbDSDsComPropRelatedToConcept	266	86.1%	6,538	100%	6804	99.4%
		M27	NbDSDsComPropRelatedToSDMXConcept	215	69.6%	6,538	100%	6,754	98.6%

results and practices per publisher[8] could be identified based on the URIs. We run additional queries to investigate the accuracy of the numbers, the hidden patterns, and the usual practices.

4.3 Analysis and Interpretation

The raw values for each one of the indirect and direct metrics in Table 1 are presented in Table 2, and discussed in the remaining of this section.

Goal 1. Most datasets are defined by a DSD (M2%=98.3%), with 273 exceptions. Two publishers, responsible for 263 and 10 datasets, respectively, do not adopt this practice at all. DSDs are not often reused (M4=12), a practice identified only for three publishers. One of the DSDs, which follows the modeling strategy ST5, is reused 9,416 times, probably according to the different values of the dimension representing a measure. Otherwise, reused DSDs are related to 2-4 datasets.

Regarding the reuse of dimensions and measures, two patterns can be found: reuse inside the scope of a same publisher, as part of the definitions that are consistently adopted, or reuse of SDMX standards. Although most publishers have dimensions representing the same real world of entity (e.g. countries, years), they all define their in-house representation for them. Linkage was found just at instance level (e.g. geographical entities). Both dimensions (M7%=61.1%) and measures (M8%=19.6%) were reused in DSDs. Linked Eurostat adopts this practice more consistently: it adopts generic sdmx-measure:obsValue as measure (strategies ST4

[8] The term *publisher* refers to the entity to which the data refers to, such as FAO, FMI, Eurostat, etc., and not necessarily the organization, possibly a third party, which actually published the data in the LOD cloud.

or ST5), and reuses consistently the same in-house dimension definitions across all datasets, with the exception of sdmx-dimension:freq and dcterms:time, used in 100% of the DSDs. SDMX dimensions and measures are reused as super property of other dimensions (M9=4) and measures (M10=1), particularly obsValue (193), refArea (20) and refPeriod (15). Among the top 5 reused dimensions (M11), two are generic (sdmx-dimension:freq and dcterms:time), used 6,476 times each only in Eurostat. The other three ones are in-house representations for location, measure unit, and sex. Outside Eurostat, the popular dimensions are publisher dependent representations for Time, Time period, and location, where just two of them are defined as sub-property of the respective SDMX dimension. The top 5 reused measures (M12) are sdmx:obsValue (6,539) and its variations according to the publishers (all defined as sub-property of sdmx:obsValue).

Goal 2. Strategies ST1 (M15) and ST3 (M14) are not adopted at all, and ST2 is rarely adopted ((M13%=17.5% referring two non-Eurostat publishers). Every time a DSD presented a single measure, it was a generic one (sdmx-measure:obsValue or publisher-dependent variations) (M16$=99.1%). To distinguish between ST1 and ST4/ST5 in non-Eurostat DSDs, we selected all DSDs defining a single measure (244). From these, we filtered the ones involving combinations of the strings 'obs' and 'value' in the URI, which resulted in the exact same set. To confirm, we manually inspected 50 of these measure properties and confirmed that labels, descriptions or concepts conveyed no specific domain information. The numbers referring to DSDs following strategy ST5 (M17%=33.1%) may not be accurate due to the difficulty in identifying dimensions representing indicators in a reliable manner. We also found 11 DSDs related to a same publisher without any measure.

To identify indicator dimensions (M18), we searched for patterns involving concepts, codelists and URI rules. Labels involving the string 'indicator' sometimes would reveal such type of dimension, but the most encompassing strategy involved merely URI patterns. The best substrings were 'indic', 'variab' and 'measure'. For Eurostat, 36 dimensions were identified in this way, and for other 6 publishers, 13. We obtained 100% of precision with the first two substrings, and 58% with the latter. In all cases, the error referred to a dimension representing a unit of measure (an attribute in Cube), rather than a measure. We also examined randomly 70 other dimensions that did not follow this pattern, and none of them was an indicator dimension, which could be considered as 100% of recall.

Goal 3. Most dimensions are related to concepts (M21%=86.3%), but measures are not (M22%=9.8%). However, most concepts are also internal to the publisher, normally paired with the codelists defined for the dimensions. This could be easily verified by pairing the URIs, and verifying their formation patterns. When measuring the number of dimensions/measures related specifically to SDMX concepts, we noticed that the figures were exactly the same. A manual inspection revealed the common practice of defining a concept as an instance of sdmx:Concept, which is not adequate considering SDMX is a standard to be shared across datasets of various domains. To measure linkage with SDMX concepts, we adopted the following definitions: a concept that belongs to the standard SDMX COG, an SDMX dimension/measure (which is always associated to a SDMX concept), or a dimension/measure defined as sub-property of an SDMX dimension/measure. According to this more strict interpretation, the number of component properties

related to the SDMX standard was limited (M24%=3.5% and M25%4.9%). DSDs were considered related to (SDMX) concepts if at least one dimension/measure was, which explains the high figures for M26 and M27.

The top 5 concepts used in DSDs were sdmx-concept:obsValue, sdmx-concept:freq, and three concepts within the domain of Eurostat representing location, measuring unit and sex. Among non-Eurostat publishers, the top concepts were sdmx-concept:obsValue, and distinct representations for time period, location and frequency. We also examined the range of dimensions in the search for conceptual clues, but mostly they are defined merely as generic skos:Concept (100% in case of Eurostat), or publisher-dependent concepts. Using properties owl:sameAs and skos:exactMatch, we found that geographic locations and currencies were frequently linked at instance level. Otherwise, interlinked resources are quite rare.

4.4 Threats of Validity

We discuss the threats that may influence the validity of our experiments according to the categories proposed in [14]. *Construct validity* refers to the extent to which the experiment setting actually reflects the construct under study. A major threat is the inability of correctly identifying the Cube datasets in the LOD cloud. We used the Mannheim Catalogue, the most update collection of linked data. The extraction procedures also constitute a risk, as we limited the scope of the properties and resources to be investigated, compared to the original sources. We mitigated this risk by defining a well-defined set of vocabulary terms and by exploring the different paths that connect resources, as well as by focusing the analysis solely on the type of resources extracted. *Conclusion validity* threats the ability to draw the correct conclusion about relations between the treatment and the outcome of an experiment. We assumed the existence of specific modeling strategies, but eventually others exist. We mitigated this risk by basing our study on the cube specification and real datasets found in the LOD cloud. The representativeness of the measures is another concern. As one publisher represented about half of the datasets and DSDs studied, we mitigated this risk by developing a separate analysis to detect dominance on the practices regarding this publisher. *External validity* is concerned with generalization. We noticed that the vast majority of cube datasets available were conversions of SDMX data, so the results might be circumstantial. It can also be argued that some datasets are prior the recommendation of W3, but we noticed that most people responsible for the actual publication were also involved in the Cube working group.

5 Related Work

The 2014 LOD status report [13] revealed the tendency of growth in the number of statistical datasets and Cube prevalent usage over other vocabularies, and resulted in the Mannheim Linked Data Catalogue, used in this work. A study [12] examined different strategies of vocabulary reuse in the LOD cloud, and by far participants preferred a single, popular vocabulary. However, both standard vocabulary and standard usages are important.

Different works have addressed the modeling of data using the Cube in case studies, either from scratch [10,9] or converting existing SDMX data using rather

straightforward conversion rules (e.g. [4]). Our work contributes with feedback that can make such conversion rules evolve, such that the purposes and benefits of both LOD and SDMX can be achieved.

LOD2 Statistical Workbench [7], OpenCube [8], Vital [5] and OLAP4LD [9] are platforms that support using, publishing, validating and visualizing Cube datasets. The results of this survey can be leveraged to integrate components that also provide methodological guidance to support modelings choices. An approach for identifying potentially relevant datasets in the LOD cloud related to seed concepts was presented in [2], where the diversity of modeling approaches and conceptual enrichment needs to be taken into account.

6 Conclusions

In this paper, we discussed common Cube modeling strategies,and their impact on automatically finding and consuming data. We then developed a quantitative study to understand current usage of the Cube in the LOD cloud.The analysis framework uses GQM to translate concerns about acknowledgement of DSD advantages and reuse, modeling strategies, and conceptual enrichment. The results establish a baseline for future comparison, and provides input for works on Cube methodological support, supporting platforms, and cube datasets discovery.

Although outside of the scope of this study, we were surprised by the elevated number of cubes that are not well-formed. We found cube datasets without a corresponding DSD, DSDs without measures, dimensions defined without ranges, among other trivial issues. Existing Cube validators [7] can easily support the syntactic usage of the vocabulary, but we lack methodological and infrastructure support for modeling decisions, and translating them in the appropriate Cube constructs.

The results reveal that most Cube datasets are straightforward conversions of SDMX data. This type of conversion explores mainly the properties of SDMX as a standard for exchanging statistical data. It is effective from an interoperability point of view, but limited with regard to the LOD purposes, in which the focus is the ability of automatically processing of data. The adoption of underused cube constructs would introduce more normative ways of modeling multidimensional data, and explicitly defining in the structure and semantics of DSDs, particularly with regard to measures and dimensions, as discussed in Section 3.

Publishers are also very concerned with establishing a proper, standard vocabulary which is uniformly applied within the scope of a specific organization. There is a serious concern with linkage, but currently, it is mainly focused on instances of specific types (e.g. geographical). Despite dimensions often represent the same concepts (age, year, frequency), they are hardly integrated across datasets of different domains, nor interlinked. The study points for an opportunity of integrating commonly used dimensions, either by reuse, adoption of standard concepts, or concept-based linkage.

Overall, the use of Cube is new, and its usage will reveal the importance of certain constructs (e.g. non-normalized cubes). Features for modeling aggregated measures, typical of BI environments, are certainly missing [6].

We are currently using the investigated patterns of Cube usage in an approach to automatically identify and integrate cube datasets for developing knowledge

discovery applications [2]. The main challenges have been to explore how the semantics about dimensions and measures are represented in order to match with seed concepts. In addition, the compatibility of dimension values is key for integrating indicators distributed in different datasets. Future work includes the expansion of the survey to explore other aspects such as additional constructs, linkage and provenance; development of Cube modelling guidelines and methodological support; development of automatic converters for the identified modeling strategies; formalization of dimension compatibility verification and conversion algorithms; among others.

Acknowledgments This research is sponsored in part by CAPES (Brazil) and in part by the U.S. National Science Foundation under Grant No. 1117913.

References

1. Basili, V.R., Caldiera, G., Rombach, H.D.: The goal question metric approach. In: Encyclopedia of Software Engineering. Wiley (1994)
2. Becker, K., Jahangiri, S., Knoblock, C.A.: Finding, assessing, and integrating statistical sources for data mining. In: Proc. of the 4rth Workshop on Knowledge Discovery and Data Mining - co-located with ESWC2015, Portoroz, 2015. (2015), http://ceur-ws.org/Vol-1365/paper5.pdf
3. Bouza, M., Elliot, B., Etcheverry, L., Vaisman, A.: Publishing and querying government multidimensional data using QB4OLAP. In: Web Congress (LA-WEB), 2014 9th Latin American. pp. 82–90 (Oct 2014)
4. Capadisli, S., Auer, S., Ngomo, A.C.N.: Linked sdmx data: Path to high fidelity statistical linked data. Semantic Web 6(2) (2015)
5. Daga, E., d'Aquin, M., Gangemi, A., Motta, E.: Early analysis and debugging of linked open data cubes. In: Proc. of the Second International Workshop on Semantic Statistics, 2014
6. Etcheverry, L., Vaisman, A.A., Zimányi, E.: Modeling and querying data warehouses on the semantic web using QB4OLAP. In: Proc. of the 16th International Conference on Data Warehousing and Knowledge Discovery (DAWAK), 2014. pp. 45–56 (2014)
7. Janev, V., Mijovic, V., MiloSevic, U., Vranes, S.: Supporting the linked data publication process with the LOD2 statistical workbench. Semantic Web Journal (2014)
8. Kalampokis E. et al: Exploiting linked data cubes with opencube toolkit. In: Proceedings of the ISWC 2014 Posters & Demonstrations Track, Riva del Garda, Italy, Oct. 2014. pp. 137–140 (2014)
9. Kämpgen, B., Harth, A.: OLAP4LD–A Framework for Building Analysis Applications Over Governmental Statistics. In: The Semantic Web: ESWC 2014 Satellite Events, pp. 389–394. Springer International Publishing (2014)
10. Omitola, T. et al., Integrating public datasets using linked data: Challenges and design principles. In: Proc. of the Workshop on Linked Data in the Future Internet, Ghent, 2010 (2010), http://ceur-ws.org/Vol-700/Paper8.pdf
11. Reynolds, D., Cyganiak, R.: The RDF Data Cube vocabulary. Tech. rep., W3C Recommendation (Jan 2014), http://www.w3.org/TR/2014/REC-vocabdata-cube-20140116
12. Schaible, J., Gottron, T., Scherp, A.: Survey on common strategies of vocabulary reuse in linked open data modeling. In: The Semantic Web: Trends and Challenges. Lecture Notes in Computer Science, vol. 8465, pp. 457–472 (2014)
13. Schmachtenberg, M., Bizer, C., Paulheim, H.: Adoption of the linked data best practices in different topical domains. In: 13th International Semantic Web Conference, Riva del Garda, Italy, Oct. 2014. 245–260 (2014)
14. Wohlin, C. et al: Experimentation in Software Engineering. Springer-Verlag (2012)