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

Karin Becker
Instituto de Informática - Federal University of Rio Grande do Sul, Brazil

Shiva Jahangiri, Craig A. Knoblock
Information Sciences Institute, University of Southern California, USA

Introduction

- Statistical data is used as the foundation for policy prediction, planning and adjustments
- Growing consensus that Linked Open Data (LOD) cloud is the right platform for sharing and integrating open data
- The success of the LOD depends on basic principles
  - Common vocabulary reuse
  - Interlinking
  - Metadata provision
  - Otherwise, it is just another platform for making data available

Introduction

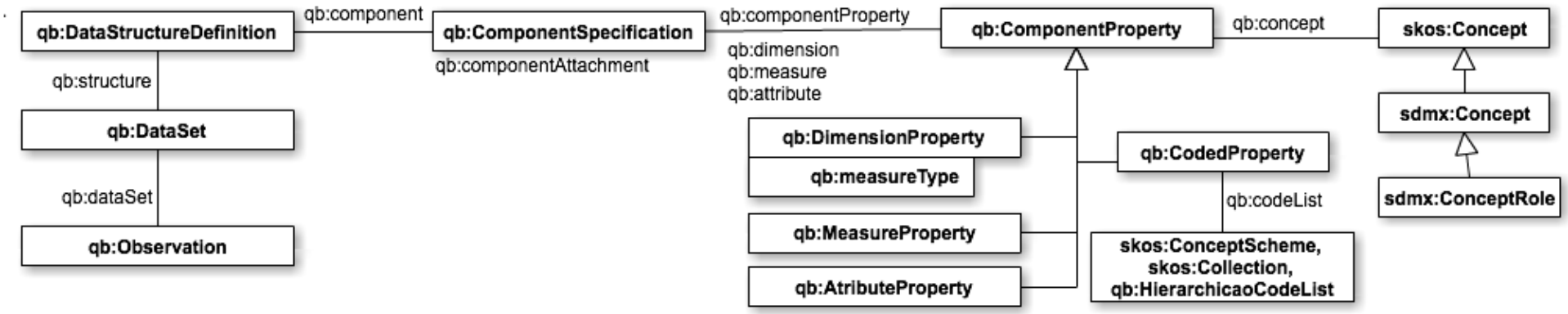
- Cube vocabulary
  - W3C recommendation
  - Multidimensional representation of data
    - But designed to be compatible with statistical ISO SDMX standard
  - Popular (62% of datasets in the LOD in the governmental domain)
  - Several projects address platforms for publishing data using the cube

- Is data being represented using the Cube in such a way that it can be easily found in the LOD cloud, consumed and integrated with other data ?

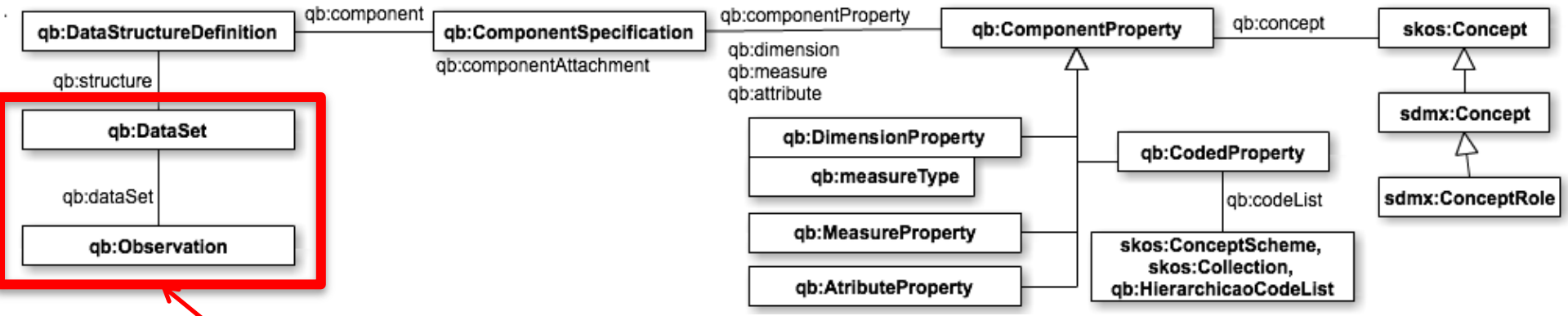
Goal

- Quantitative survey on the current usage of the Cube vocabulary
  - Governmental data identified in the last LOD census (2014)
- Focus: commonly used strategies for modeling multi-dimensional data
  - They affect how data can be found and consumed automatically
- Contributions
  - Analysis of various ways the Cube vocabulary is used in practice
  - Guidance on the most useful representations
  - Baseline for comparison with the evolution of Cube usage
  - Input for methodological support and platforms addressing Cube usage

Cube Vocabulary

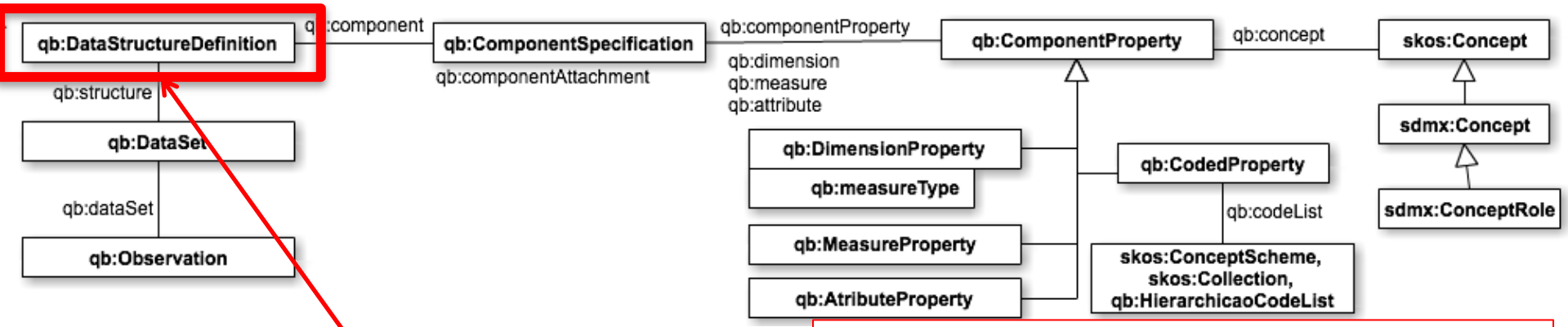


Cube Vocabulary



- The actual data
- The structure of the dataset is implicitly represented
- Possibly large volumes of data

Cube Vocabulary

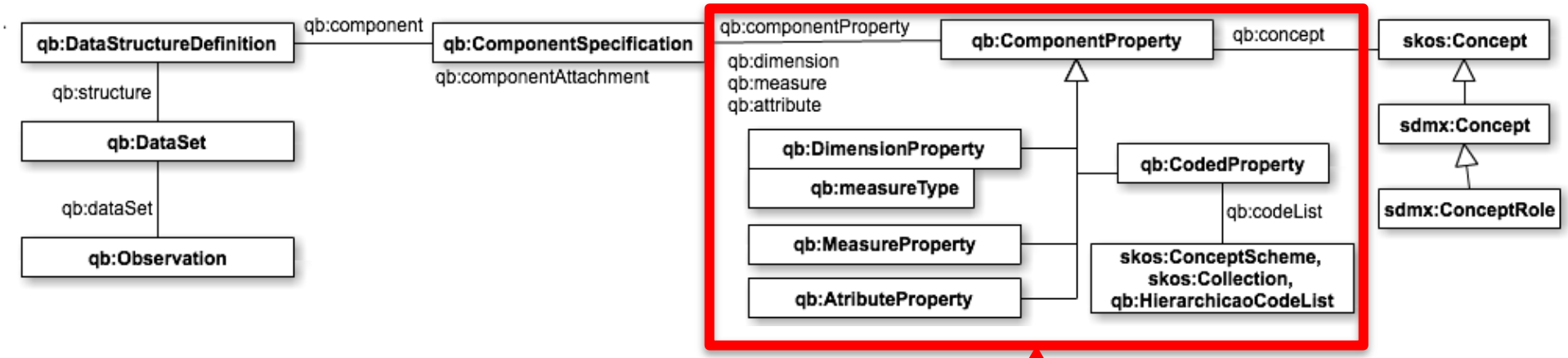


- The description of the data
- Explicit representation
- Concise description

Advantages

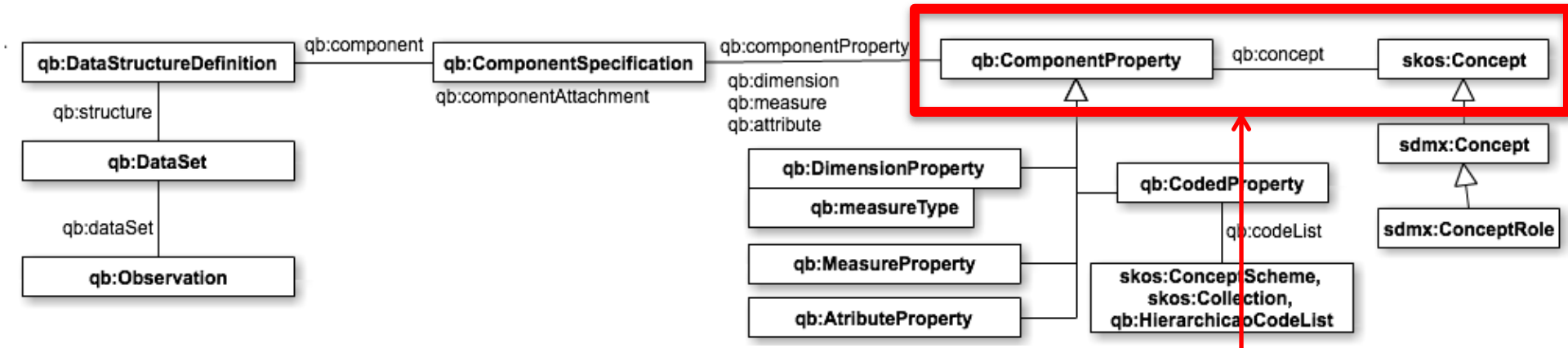
- Checking conformance of actual data with regard to expected structure
- Simplification of data consumption, due to explicit properties
- Reuse in the publication process
- Build trust and normatization for consumption

Cube Vocabulary



- Measures and dimensions
- "measure dimension" (qb:measureType)
- Possible values for dimensions

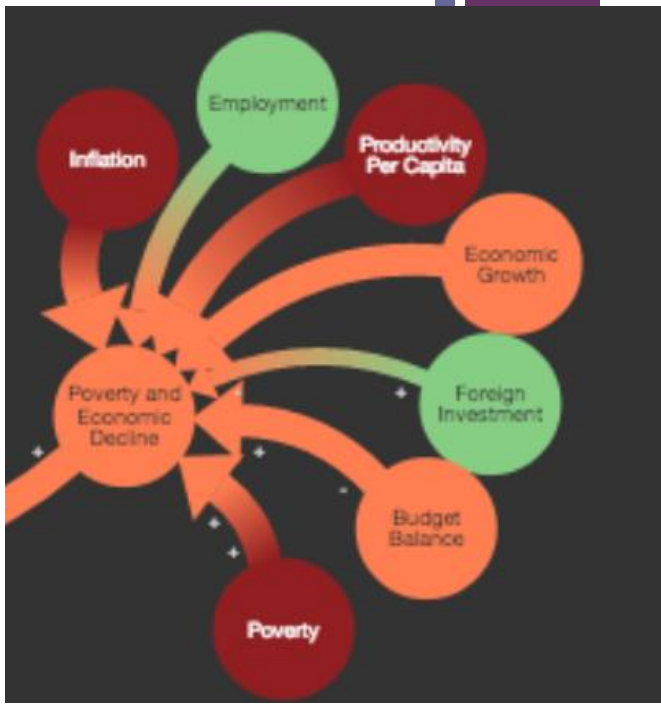
Cube Vocabulary



- Concepts represented by measures and dimensions
- Possibly SDMX concepts

Motivating Example

- Prediction of public indicators: Fragile State Index (FSI)
  - 14 social, economic and political indicators
  - Methodology
    - software that collects millions of documents, select relevant ones, and values indicators (CAST)
    - human analysis
- Can we predict FSI indicators using other indicators and data available in the LOD Cloud?
  - Automatic location and consumption
  - Otherwise, it is just another media where data is available ...



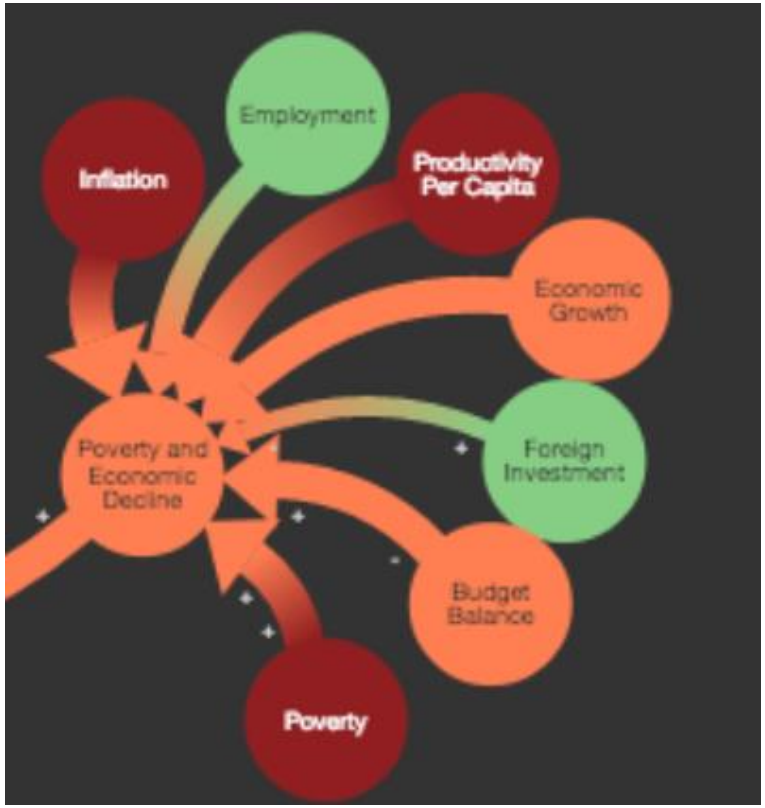
FAO FIAT PANIS

The World Bank

UNdata
A world of information

http://ffp.statesindex.org/methodology

Motivating Example



- Find datasets that
  - Measures
    - Have the label "poverty"
    - Are described by using the term “poverty”
    - Are related to the concept poverty
    - etc
  - Dimensions
    - year time series
    - countries

Modeling Strategies

#ST1: Single Measure

```
fao:SingleMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fao:AvgDESAdequacy . ].
```

#ST2: Multi Measure

```
fsi:MultiMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fsi:DemographicPressures;
                   qb:measure fsi:RefugeesandIDPs  . ].
```

#ST3: Measure Dimension

```
fao: MeasureDim a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure qb:measureType;
                   qb:measure fao:AvgDESAdequacy  ;
                   qb:measure fao:AvgValueFoodProd  .].
```

```
fao:refPeriod  a qb:DimensionProperty;
   rdfs:subPropertyOf sdmx-dimension:refPeriod;
   rdfs:range xsd:gYear; ....

fao:refArea  a qb:DimensionProperty;
   rdfs:range schema:Place;
   qb:concept sdmx-concept:refArea; ....

fao:AvgDESAdeq  a qb:MeasureProperty;
  rdfs:label "Avg. Dietary Energy Supply Adequacy"en;
  rdfs:subPropertyOf sdmx-measure:obsValue;
  rdfs:range xsd:decimal; ....

fao:AvgValueFoodProd  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:DemographicPressures  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:RefugeesandIDPs  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue ...
```

Modeling Strategies

#ST1: Single Measure

```
fao:SingleMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fao:AvgDESAdequacy . ].
```

#ST2: Multi Measure

```
fsi:MultiMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fsi:DemographicPressures;
                   qb:measure fsi:RefugeesandIDPs  . ].
```

#ST3: Measure Dimension

```
fao: MeasureDim a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure qb:measureType;
                   qb:measure fao:AvgDESAdequacy  ;
                   qb:measure fao:AvgValueFoodProd  .].
```

Single Measure

- Each observation contains a value for the measure

Several Dimensions

Measures and dimensions can be related to both

- generic (statistical) concepts
- domain concepts

```
rdfs:subPropertyOf sdmx-measure:obsValue;
  rdfs:range xsd:decimal; ....

fao:AvgValueFoodProd  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:DemographicPressures  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:RefugeesandIDPs  a qb:MeasureProperty;
  rdfs:subPropertyOf sdmx-measure:obsValue ...
```

Modeling Strategies

#ST1: Single Measure

```
fao:SingleMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fao:AvgDESAdequacy . ].
```

#ST2: Multi Measure

```
fsi:MultiMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fsi:DemographicPressure
                   qb:measure fsi:RefugeesandIDPs  . ].
```

#ST3: Measure Dimension

```
fao: MeasureDim a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure qb:measureType;
                   qb:measure fao:AvgDESAdequacy  ;
                   qb:measure fao:AvgValueFoodProd  .].
```

```
fao:refPeriod  a qb:DimensionProperty;
   rdfs:subPropertyOf sdmx-dimension:refPeriod;
   rdfs:range xsd:gYear; ....
```

Multiple Measures

- Each observation must contain values for **all** measures

Several Dimensions

Measures and dimensions can be related to both generic and domain concepts

```
fao:AvgValueFoodProd  a qb:MeasureProperty;
   rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:DemographicPressures  a qb:MeasureProperty;
   rdfs:subPropertyOf sdmx-measure:obsValue; .....

fsi:RefugeesandIDPs  a qb:MeasureProperty;
   rdfs:subPropertyOf sdmx-measure:obsValue ...
```

Modeling Strategies

#ST1: Single Measure

```
fao:SingleMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fao:AvgDESAdequacy . ].
```

#ST2: Multi Measure

```
fsi:MultiMea a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure fsi:DemographicPressure
                   qb:measure fsi:RefugeesandIDPs  . ].
```

#ST3: Measure Dimension

```
fao: MeasureDim a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure qb:measureType;
                   qb:measure fao:AvgDESAdequacy  ;
                   qb:measure fao:AvgValueFoodProd  .]
```

```
fao:refPeriod  a qb:DimensionProperty;
   rdfs:subPropertyOf sdmx-dimension:refPeriod;
   rdfs:range xsd:gYear; ....

fao:refArea  a qb:DimensionProperty;
   rdfs:range schema:Place;
   qb:concept sdmx-concept:refArea; ....

fao:AvgDESAdeq  a qb:MeasureProperty;
  rdfs:label "Avg. Dietary Energy Supply Adequacy"en;
```

Measure Dimension

- Each observation contains one value for one of the measures
- The specific measure is the value of the "measure dimension"

Several Dimensions

Measures and dimensions can be related to both generic and domain concepts

Modeling Strategies

```
#ST4: Generic Single Measure
fao:captureDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure sdmx-measure:obsValue. ].
```

```
#ST5: Ad| hoc Dimension Measure
wb:indicatorDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension wb:refArea;
                   qb:dimension wb:refPeriod;
                   qb:dimension wb:indicator;
                   qb:measure sdmx-measure:obsValue. ].
```

```
wb:refPeriod  a qb:DimensionProperty; ...
```

Single Generic Measure

- each observation contains a value for the measure
- **a generic statistical measure**
- **cannot be related to domain concepts**

Several Dimensions

DSD is limited in the explicit information it provides

Modeling Strategies

```
#ST4: Generic Single Measure
fao:captureDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure sdmx-measure:obsValue
#ST5: Ad hoc Dimension Measure
wb:indicatorDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension wb:refArea;
                   qb:dimension wb:refPeriod;
                   qb:dimension wb:indicator;
                   qb:measure sdmx-measure:obsValue
```

Ad hoc Dimension Measure

- each observation contains a value for a measure
- **a generic statistical measure**
- **cannot be related to domain concepts**

Several Dimensions

- **one dimension is implicitly a measure dimension**
- a codelist might describe the measure, but only the actual dataset defines the measure

- DSD is limited in the explicit information it provides

Modeling Strategies

#ST4: Generic Single Measure

```
fao:captureDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension fao:refArea;
                   qb:dimension fao:refPeriod;
                   qb:measure sdmx-measure:obsValue. ].
```

#ST5: Ad hoc Dimension Measure

```
wb:indicatorDSD a qb:DataStructureDefinition;
    qb:component [ qb:dimension wb:refArea;
                   qb:dimension wb:refPeriod;
                   qb:dimension wb:indicator;
                   qb:measure sdmx-measure:obsValue. ].
```

```
wb:refPeriod  a qb:DimensionProperty; ...
   rdfs:subPropertyOf sdmx-dimension:refPeriod.

wb:refArea  a qb:DimensionProperty; ...
   rdfs:subPropertyOf sdmx-dimension:refArea .

wb:indicator  a qb:DimensionProperty;
   qb:concept wb-concept:indicatorConcept; ...
...qb:codeList wb-classification:indicatorCodeList.
```

- Correct with regard to the Cube, but …
  - DSD fulfills its role partially
  - Conformance of the actual data with regard to structure is limited to structural properties
  - Semantics is poor
- Harder to automatically locate useful datasets in the LOD cloud and consume

Goal-Question-Metric (GQM)

- Proposed by Basili et al. in experimental SW engineering
- Measurement model at three levels
  - Conceptual: **Goal** of the measurement
    - entity, purpose, focus, point of view and context
  - Operational: **Questions** define models of the object of study
    - characterize the assessment or achievement of a specific goal
  - Quantitative: a set of **Metrics**
    - defines a set of Measures that enable to answer the questions in a measurable way.

Survey: Goals

- 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
  - Do publishers agree that DSDs have several benefits?
  - Do publishers reuse DSDs and its underlying definitions?
- Goal 2: Analyze DSD for the purpose of understanding with respect to modeling strategy from the point of view of the publisher
  - how frequent is each modeling strategy?
  - 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
  - Do publishers practice semantic annotation on DSDs?

Survey: Method

- Context
  - Data from the LOD cloud census (Aug. 2014)
  - Manheim Catalogue
- Data Collection
  - 114 catalogue entries
  - March-Apr. 2015
  - Tag cube-format
- Operations
  - Sparql queries to all entries
  - All triples involving Cube constructs (except qb:Observation)
  - Results integrated in a local repository
  - Several issues for data extraction
  - Data about 16,563 cube datasets and 6,847 DSDs
  - Half of the data referred to a single publisher (Linked Eurostat)

https://github.com/KarinBecker/LODCubeSurvey/wiki

Goal 1: DSD and Reuse

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 1: DSD and Reuse

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)

- We found 273 datasets without DSDs, referring to 2 publishers
- Non-conformant cubes

Goal 1: DSD and Reuse

Goal 1: Datasets and DSDs with respect to relevance and reuse	
Q2: Are DSDs, dimensions and measures reused?	M3: NbDSDs M4 : NbReusedDSDs (M4% = M4/M3)

Metric	Measure	Non-Eurostat		Eurostat		Both	
		Count	%	Count	%	Count	%
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%

- DSD reuse is not a practice (3 publishers)
- Reuse is limited within a same publisher despite they all share similar dimensions (e.g. time, location)
  - No interlinking of concepts
- Reuse of SDMX concepts
- Popular dimensions: in-house variations of Time, Location and Sex
- Popular measures: sdmx:obs-value and its in-house variations

Goal 2: DSD Modeling Strategy

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 dimension representing measures (ST5)	M17 : nbDSDsWithDimensionReprMeasure (M17%=M17/M3) M18: TopStrategiesDimensionRepresentingMeasure

Goal 2: DSD Modeling Strategy

Goal 2 : DSD with respect to modeling strategy	
Q3: How many DSDs apply the multi-measure strategy (ST2)?	M13 : nbDSDsWithMultipleMeasures (M13%=M13/M3)

Metric	Measure	Non-Eurostat		Eurostat		Both	
		Count	%	Count	%	Count	%
M13	nbDSDsWithMultipleMeasures	54	17.5%	0	0%	54	0.8%
M14	nbDSDsWithMeasureDimensionApproach	0	0%	0	0%	0	0%
M15	nbDSDsWithSingleDomainMeasure	0	0%	0	0%	0	0%
M16	nbDSDsWithSingleGenericMeasure	244	79%	6,538	100%	6782	99.1%
M17	nbDSDsWithDimensionReprMeasure	33	10.7%	2,233	34,2%	2266	33.1%

- 1st strategy: a single generic measure (ST4)
- 2nd strategy: a dimension implicitly representing a measure dimension (ST5)
- Strategies to find dimensions representing measures (ST5):
  - Patterns involving the URI (e.g. included indic, variab, measur)
  - Concepts and codelists were not useful at all
- Strategies to find generic measures also involved URI patterns

Goal 3: DSD Conceptual Enrichment

Goal 3 : DSD with respect to conceptual enrichment	
Q7: Do publishers relate component properties to concepts for conceptual enrichment?	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

Goal 3: DSD Conceptual Enrichment

Goal 3 : DSD with respect to conceptual enrichment	
Q7: Do publishers relate component properties to concepts for	M19 NbComponentProp M20: NbCompPropRelatedToConcept (M20%=M20/M19)

Goal	Question	Metric	Measure	Non-Eurostat		Eurostat		Both	
				Count	%	Count	%	Count	%
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%

- Dimensions are often related to concepts, however …
  - in-house concepts, not interlinked with external concepts (e.g. owl:same-as, skos:exactMatch)
  - frequently concepts are paired with codes from codelists (uri patterns)
- Top concepts:
  - sdmx-concept:obsValue, sdmx-concept:freq
  - Different in-house representations for location, time, measuring unit and sex

Goal 3: DSD Conceptual Enrichment

Goal 3 : DSD with respect to conceptual enrichment	
Q7: Do publishers relate component properties to concepts for	M19 NbComponentProp M20: NbCompPropRelatedToConcept (M20%=M20/M19)

Goal	Question	Metric	Measure	Non-Eurostat		Eurostat		Both	
				Count	%	Count	%	Count	%
G3	Q7	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%

- Common practice of defining a concept as an instance of sdmx:Concept
  - not adequate considering SDMX is a standard to be shared across datasets of various domains, with well-defined concepts (COG)
- For the survey, we adopted a more strict interpretation
  - concept that belongs to the standard SDMX COG
  - (subproperty of) SDMX dimension/measure (which is always linked to a sdmx-concept)
- Top concepts: sdmx-concept:obsValue, sdmx-concept:freq

Related Work

- Surveys
  - LOD Census : growing importance of the Cube and governmental topical domain (Schmachtenberg et al. 2014)
  - Preferred reuse strategy: a single, popular vocabulary (Schaible et al.2014)
- platforms that support using, publishing, validating and visualizing Cube datasets
  - LOD2 Statistical Workbench, OpenCube, Vital, OLAP4LD
  - Our results can be leveraged to integrate components that also provide methodological guidance to support modeling choices
- Automatic search of open data for data mining (Becker et al. 2015; Janpuangtong et al. 2015)

Conclusions

- Survey current practices of modeling datasets with the Cube vocabulary
  - Surprised by the number of non-conformant cube datasets
  - most Cube datasets are straightforward conversions of SDMX data
    - standard for exchanging statistical data: interoperability
    - LOD cloud: ability of automatically processing of data requires
    - Next step: more complex conversion rules
  - Cube constructs are underused
    - more normative ways of modeling multidimensional data, and explicitly defining in the structure and semantics of DSDs
    - the use of Cube is new, and its usage will reveal the importance of certain constructs/modeling strategies

Conclusions and Future Work

- Publishers are concerned with establishing a proper, standard vocabulary to uniformly apply within the scope of a specific organization
  - Opportunity integrate commonly used dimensions, either by reuse, adoption of standard concepts, or concept-based linkage
- Survey has a specific focus
  - Baseline for future comparison
  - Extended to other aspects
  - Results can be leveraged into supporting platforms
- currently we are using the investigated patterns of Cube usage to automatically identify and integrate cube datasets for data mining applications