Publishing Statistical Data as Linked Data «The RDF Data Cube Vocabulary»

George Papastefanatos
Irene Petrou

Institute for the Management of Information Systems
RC "Athena"

Open Data Day 2013@Athens

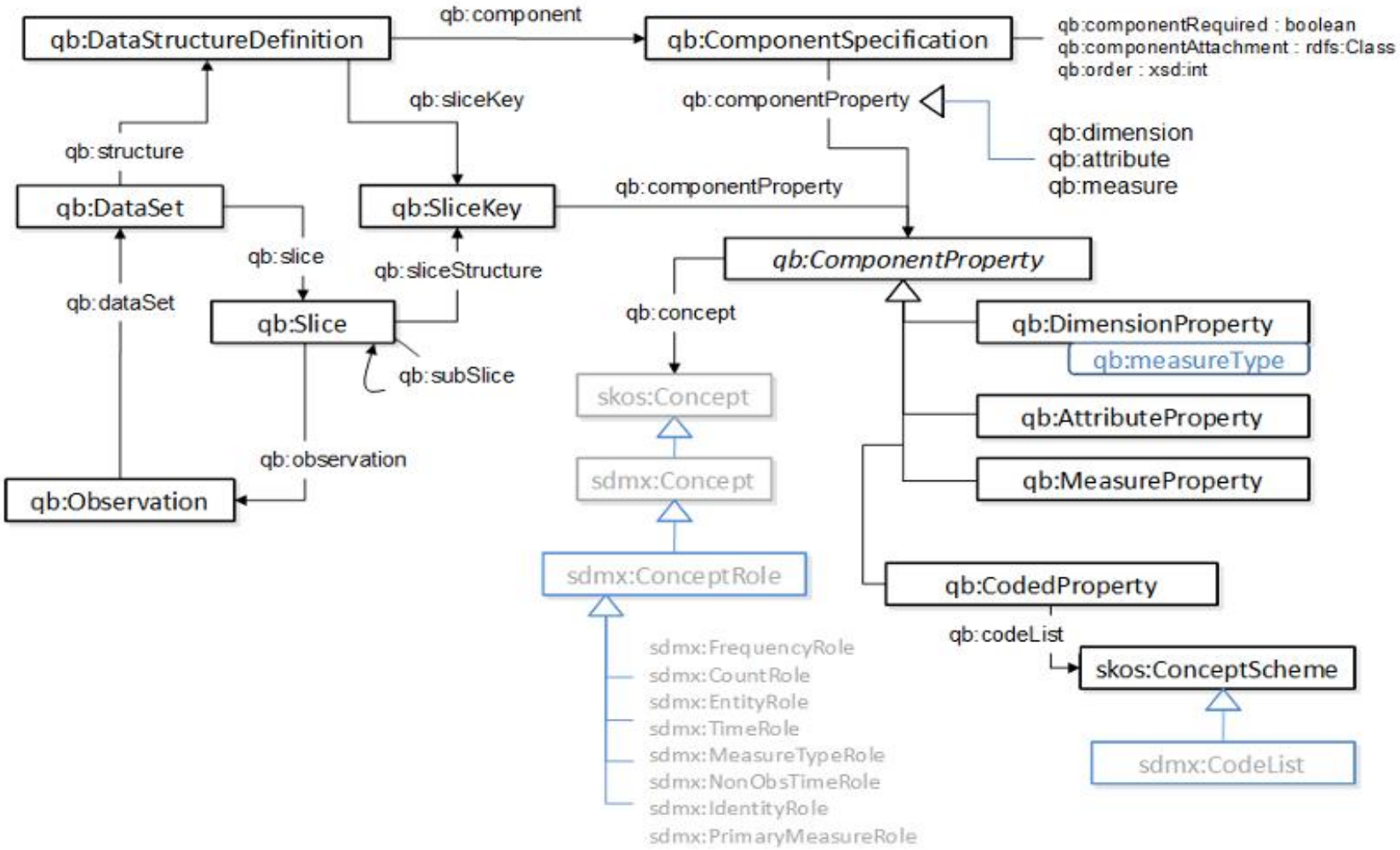
Data Cube Vocabulary

- Source:
  http://www.w3.org/TR/2012/WD-vocab-data-cube-20120405/

- W3C Working Draft 05 April 2012

- The Data Cube Vocabulary is based on the following existing RDF vocabularies:
  - SKOS for concept schemes
  - SCOVO for core statistical structures
  - VoiD for data access
  - FOAF for organisations
  - Dublin Core Terms for metadata

Outline of the Data Cube Vocabulary



Basic Steps

1. Define the prefixes to be used
2. Publish your schema
   - Define the dimension(s) – used to identify the observations (ex. time, region), what the observation applies to
   - Define the measure(s) – the phenomenon being observed
   - Define the attribute(s) - unit of measure
   - Define the DSD (attach components)
3. Publish your data
   - Define the Dataset (attach DSD)
   - Define Observations – the actual data (ex. the values in the cells of a table)
4. Including Slices (views) on your data
   - Define SliceKey(s) - the fixed dimensions
   - Define the DSD (attach SliceKey(s))
   - Define the Dataset (attach Slices to be defined)
   - Define Slices and Observations
5. Selecting appropriate URIs

Example of Data Cube Vocabulary

- Source: StatsWales Report number 003311
- Topic of the Data Set: Life expectancy broken down by region (unitary authority), sex and time.

	2004-2006		2005-2007		2006-2008	
	Male	Female	Male	Female	Male	Female
Newport	76.7	80.7	77.1	80.9	77.0	81.5
Cardiff	78.7	83.3	78.6	83.7	78.7	83.4
Monmouthshire	76.6	81.3	76.5	81.5	76.6	81.7
Merthyr Tydfil	75.5	79.1	75.5	79.4	74.9	79.6

- Three **dimensions**:
  1. Time period
  2. Region
  3. Sex
- **Measure**: Each observation represents the life expectancy for that population – a single measure which corresponds to all data set
- **Attribute**: Years (the units of the measured values)

Prefixes

@prefix eg: <http://example.org/life-expectancy#>.

@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .

@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@prefix sdmx: <http://purl.org/linked-data/sdmx#> .

@prefix sdmx-concept: <http://purl.org/linked-data/sdmx/2009/concept#> .

@prefix sdmx-dimension: <http://purl.org/linked-data/sdmx/2009/dimension#> .

@prefix sdmx-attribute: <http://purl.org/linked-data/sdmx/2009/attribute#>

@prefix sdmx-measure: <http://purl.org/linked-data/sdmx/2009/measure#> .

@prefix qb: <http://purl.org/linked-data/cube#> .

Components definition

- All dimensions, attributes and measures are components

Time:

- Reuse of Predefined concept in the SDMX-COG, *refPeriod*
- To represent the time period itself it would be convenient to use the data.gov.uk reference time service. Example and more information on time intervals can be found at:

http://www.epimorphics.com/web/wiki/using-interval-set-uris-statistical-data

```
eg:refPeriod  a rdf:Property, qb:DimensionProperty;
      rdfs:label "reference period"@en;
      rdfs:subPropertyOf sdmx-dimension:refPeriod;
      rdfs:range interval:Interval;
      qb:concept sdmx-concept:refPeriod .
```

Note:

rdfs:range is an instance of rdf:Property that is used to state that the values of a property are instances of one or more classes.

The triple *P rdfs:range C* states that P is an instance of the class rdf:Property, that C is an instance of the class rdfs:Class and that the resources denoted by the objects of triples whose predicate is P are instances of the class C.

Components definition

Region (unitary authority):

- Reuse of Predefined concept in the SMDX-COG, *refArea*
- For this example, the Ordanance Survey Administrative geography ontology is used. More details can be found at:

http://data.ordnancesurvey.co.uk/ontology/admingeo/UnitaryAuthority

```
eg:refArea  a rdf:Property, qb:DimensionProperty;
    rdfs:label "reference area"@en;
    rdfs:subPropertyOf sdmx-dimension:refArea;
    rdfs:range admingeo:UnitaryAuthority;
    qb:concept sdmx-concept:refArea .
```

Sex:

- Direct use of SMDX-COG component, *sdmx-dimension:sex*

Components definition

Measure:

- Use of the default *sdmx-measure:obsValue*
- It is recommended to use a specific measure corresponding to the phenomenon being observed.
- *sdmx-measure:obsValue is defined by:*
  http://sdmx.org/wp-content/uploads/2009/01/01_sdmx_cog_annex_1_cdc_2009.pdf

```
eg:lifeExpectancy  a rdf:Property,
qb:MeasureProperty;
      rdfs:label "life expectancy"@en;
      rdfs:subPropertyOf sdmx-measure:obsValue;
      rdfs:range xsd:decimal .
```

Components definition

Unit measure attribute:

- Plain decimal value
- Define what units it is measured in (years)
- For this example, we can use the predefined *sdmx-attribute:unitMeasure*
- Use of a common thesaurus of units of measures. For this example, a Dbpedia resource is used, *http://dbpedia.org/resource/Year*

We can now define the structure of the data set. However, before we define the structure of the data set we should note that:

- The dimensions can be usefully ordered
- There is only one attribute, the unit measure, and this is required
- The attribute will be attached at the level of the data set, since it is only one

Data Structure Definition (DSD)

```
eg:dsd-le a qb:DataStructureDefinition;
# The dimensions
qb:component [qb:dimension eg:refArea;        qb:order 1];
qb:component [qb:dimension eg:refPeriod;      qb:order 2];
qb:component [qb:dimension sdmx-dimension:sex; qb:order 3];
# The measure(s)
qb:component [qb:measure eg:lifeExpectancy];
# The attributes
qb:component [qb:attribute sdmx-attribute:unitMeasure;
qb:componentAttachment qb:DataSet;] .
```

Notes to remember:

- The URI of the DSD can be reused across different datasets with the same structure
- The component properties can also be reused, but across different DSDs
- The component specifications are only useful within the scope of a particular DSD
- Multiple observed values are allowed to be attached to an individual observation, by declaring multiple *qb:MeasureProperty* components in the DSD and attach an instance of each property to the observation within the data set. However if one measure is applied to an observation, we can distinguish which measure applies to the observation using the *qb:measureType* dimension component.

Dataset and Observations

- To represent the entire dataset (the collection of observations) a resource is defined, the *qb:DataSet*
- A link between the DSD and the DataSet is created via the *qb:structure* property

```
eg:dataset-le1 a qb:DataSet;
    rdfs:label "Life expectancy"@en;
    rdfs:comment "Life expectancy within Welsh Unitary authorities - extracted from Stats Wales"@en;
    qb:structure eg:dsd-le ;
.
```

- Each observation is represented as an instance of a *qb:Observation*
- The values for each of the attributes, dimensions and measurements are attached directly to the observation
- The observation is linked to the containing data set via the *qb:dataSet* property.

Observations

```
eg:o1 a qb:Observation;
    qb:dataSet  eg:dataset-le1 ;
    eg:refArea              admingeo:newport_00pr ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex   sdmx-code:sex-M ;
    sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
    eg:lifeExpectancy       76.7 ;     .

eg:o2 a qb:Observation;
   qb:dataSet  eg:dataset-le1 ;
   eg:refArea              admingeo:cardiff_00pt ;
   eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
   sdmx-dimension:sex   sdmx-code:sex-M ;
   sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
   eg:lifeExpectancy       78.7 ;     .

eg:o3 a qb:Observation;
   qb:dataSet  eg:dataset-le1 ;
   eg:refArea              admingeo:monmouthshire_00pp ;
   eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
   sdmx-dimension:sex      sdmx-code:sex-M ;
   sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
   eg:lifeExpectancy       76.6 ;     .
...
```

Removing repeated component

- We notice that the measure is attached to each observation and its repeated.
- The unit of measure is not changed between the different observations.
- To remove redundancy we attached the attribute at the level of the dataset.
- Our example will now turned to:

Before

```
eg:dataset-le1 a qb:DataSet;
    rdfs:label "Life expectancy"@en;
    rdfs:comment "Life expectancy within Welsh Unitary authorities - extracted from Stats Wales"@en;
    qb:structure eg:dsd-le ;
.
```

After

```
eg:dataset-le1 a qb:DataSet;
    rdfs:label "Life expectancy"@en;
    rdfs:comment "Life expectancy within Welsh Unitary authorities - extracted from Stats Wales"@en;
    qb:structure eg:dsd-le ;
    sdmx-attribute:unitMeasure <http:/dbpedia.org/resource/Year>;
.
```

- Therefore it can be skipped by the Observations.

Removing repeated component

```
eg:o1 a qb:Observation;
    qb:dataSet  eg:dataset-le1 ;
    eg:refArea              admingeo:newport_00pr ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex   sdmx-code:sex-M ;
    sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
    eg:lifeExpectancy        76.7 ;      .

eg:o2 a qb:Observation;
    qb:dataSet  eg:dataset-le1 ;
    eg:refArea              admingeo:cardiff_00pt ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex   sdmx-code:sex-M ;
    eg:lifeExpectancy        78.7 ;      .

eg:o3 a qb:Observation;
    qb:dataSet  eg:dataset-le1 ;
    eg:refArea              admingeo:monmouthshire_00pp ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex        sdmx-code:sex-M ;
    eg:lifeExpectancy        76.6 ;      .
...
```

Slices

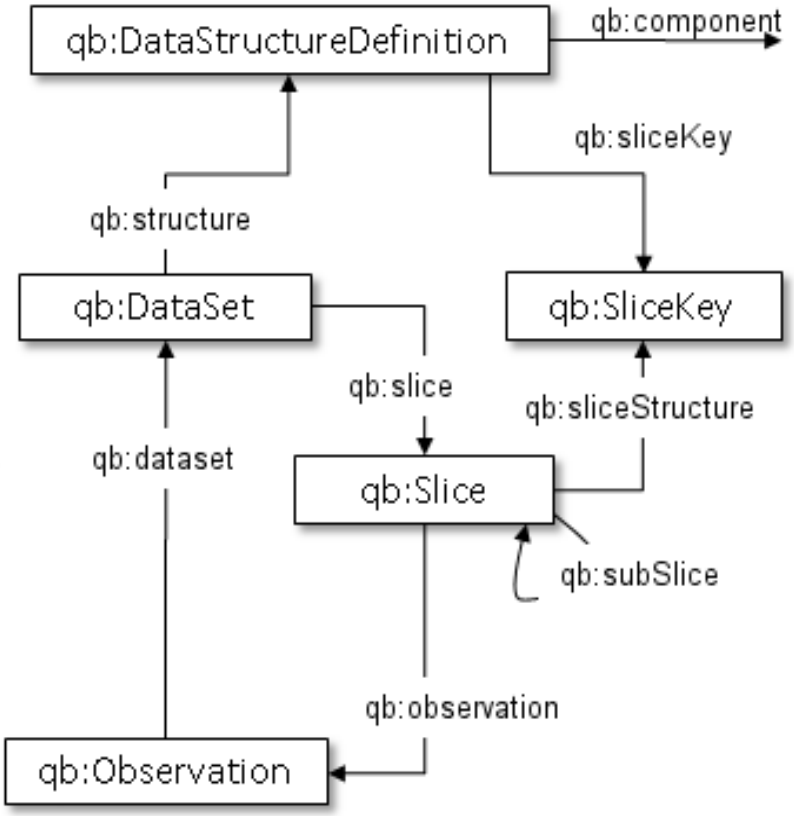
- Slices allow us to group subsets of observations together
- We first define the structure of the slices we want by associating a "slice key" with the DSD, by creating a *qb:SliceKey* which lists the component properties (which **must be** dimensions) which will be **fixed** in the slice.
- The key is attached to the DSD using qb:sliceKey.
- For example "male life expectancy observations for 2004-6" regions. In this case, we keep the dimensions sex and time fixed.

```
eg:sliceByRegion a qb:SliceKey;
   rdfs:label "slice by region"@en;
   rdfs:comment "Slice by grouping regions together, fixing sex and time values"@en;
   qb:componentProperty eg:refPeriod, sdmx-dimension:sex .

eg:dsd-le-slice1 a qb:DataStructureDefinition;
   qb:component [qb:dimension eg:refArea;        qb:order 1];
                [qb:dimension eg:refPeriod;      qb:order 2];
                [qb:dimension sdmx-dimension:sex; qb:order 3];
                [qb:measure eg:lifeExpectancy];
                [qb:attribute sdmx-attribute:unitMeasure;
                 qb:componentAttachment qb:DataSet;] ;
    qb:sliceKey eg:sliceByRegion .
```

Slices

- In the instance data then slices are represented by instances of *qb:Slice* which link to the observations in the slice via *qb:observation* and to the key via the *qb:sliceStructure*. Data sets indicate the slices they contain by means of *qb:slice*.



Slices

```
eg:dataset-le2 a qb:DataSet;
    rdfs:label "Life expectancy"@en;
    rdfs:comment "Life expectancy within Welsh Unitary authorities - extracted from Stats Wales"@en;
    qb:structure eg:dsd-le-slice2 ;
    sdmx-attribute:unitMeasure <http://dbpedia.org/resource/Year> ;
    qb:slice eg:slice2;      .

eg:slice2 a qb:Slice;
    qb:sliceStructure  eg:sliceByRegion ;
    eg:refPeriod              <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex        sdmx-code:sex-M ;
    qb:observation eg:o1b, eg:o2b; eg:o3b, ... .

eg:o1b a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:newport_00pr ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex      sdmx-code:sex-M ;
    eg:lifeExpectancy       76.7 ;      .
```

Slices

```
eg:o2b a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:cardiff_00pt ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex   sdmx-code:sex-M ;
    eg:lifeExpectancy        78.7 ;      .

eg:o3b a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:monmouthshire_00pp ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex   sdmx-code:sex-M ;
    eg:lifeExpectancy       76.6 ;     .

...
```

- However, looking at the observations we can see that we have redundancy, since we are repeating the dimension values on the individual observations: the *eg:refPeriod* and *sdmx-dimension:sex*.
- Redundancy can be reduced by declaring different attachment levels for the dimensions at the DSD.

Slices

- Before

```
eg:dsd-le-slice1 a qb:DataStructureDefinition;
  qb:component [qb:dimension eg:refArea;        qb:order 1];
               [qb:dimension eg:refPeriod;      qb:order 2];
               [qb:dimension sdmx-dimension:sex; qb:order 3];
               [qb:measure eg:lifeExpectancy];
               [qb:attribute sdmx-attribute:unitMeasure; qb:componentAttachment qb:DataSet;] ;
   qb:sliceKey eg:sliceByRegion .
```

- After

```
eg:dsd-le-slice3 a qb:DataStructureDefinition;
  qb:component
            [qb:dimension eg:refArea; qb:order 1];
            [qb:dimension eg:refPeriod; qb:order 2; qb:componentAttachment qb:Slice];
            [qb:dimension sdmx-dimension:sex; qb:order 3; qb:componentAttachment qb:Slice];
            [qb:measure eg:lifeExpectancy];
            [qb:attribute sdmx-attribute:unitMeasure; qb:componentAttachment qb:DataSet;] ;
   qb:sliceKey eg:sliceByRegion .
```

Slices

- Therefore, the observations that belong to slices now can be written in the following way:

```
eg:o1c a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:newport_00pr ;
    eg:refPeriod            <http://reference.data.gov.uk/id/gregorian-interval/2004-01-01T00:00:00/P3Y> ;
    sdmx-dimension:sex      sdmx-code:sex-M ;
    eg:lifeExpectancy       76.7 ;      .

eg:o2c a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:cardiff_00pt ;
    eg:lifeExpectancy       78.7 ;      .

eg:o3c a qb:Observation;
    qb:dataSet  eg:dataset-le2 ;
    eg:refArea              admingeo:monmouthshire_00pp ;
    eg:lifeExpectancy       76.6 ;      .

...
```

- The Data Cube vocabulary allows slices to be nested. We can declare multiple slice keys in a DSD and it is possible for one slice key to be a narrower version of another, represented using *qb:subSlice*.

Creating appropriate URIs

- Schema – Structural components

  e.g., http://linkedstatistics.gr/schema/

http://linkedstatistics.gr/schema/PopulationCensus2011/

- Vocabulary

  e.g. http://linkedstatistics.gr/dic/

http://linkedstatistics.gr/dic/geocode/

http://linkedstatistics.gr/dic/geocode#010101

  - Dataset - Observation Values

    e.g. http://linkedstatistics.gr/data/

http://linkedstatistics.gr/data/PopulationCensus2011#010101