# D1.2 - Design of data structure definition for public budget data

**Authors**

- Jakub Klímek
- Jan Kučera
- Lucie Sedmihradská
- Jindřich Mynarz
- Jaroslav Zbranek

## Executive summary

In this demonstrator deliverable, a set of OpenBudgets.eu core RDF component properties for description of budget data is presented. The deliverable is based on results achieved in a survey and knowledge elicitation performed with domain experts and reported in Deliverable D1.1 (Klímek et al., 2015). In addition to the component properties definitions, a method of creation of additional component properties and the data structure definitions themselves is presented and illustrated by an example of the budget of the European Union.

## Table of contents

# Introduction

In this demonstrator deliverable, we focus on the methodology of creating data structure definitions for budget data using the RDF Data Cube Vocabulary (DCV) together with the RDF Schema (RDFS)[1] vocabulary as the formalism for describing component properties. We also provide a list of identified reusable component properties for dimensions, attributes and measures that appear frequently in budget datasets and will serve as a tool for comparability of various budget datasets. We decided to propose component properties (i.e. instances of `qb:ComponentProperty`) as reusable components out of which concrete data structure definitions can be built. Component properties are generic enough to be reusable across multiple datasets. Unlike component specifications, they do not describe the physical structure of the dataset (e.g., component order or attachment level). The full data model documentation will be presented in deliverable D1.4.

# The RDF Data Cube Vocabulary

For representation of data cubes such as budget data in RDF the RDF Data Cube Vocabulary[2] is the most appropriate option. It is a widely used vocabulary for representing multidimensional statistical data compatible with the well-known SDMX (Statistical Data and Metadata eXchange)

[1] http://www.w3.org/TR/rdf-schema/
[2] http://www.w3.org/TR/vocab-data-cube/

ISO standard. The key terms of DCV and their relationships are depicted in Figure 1.

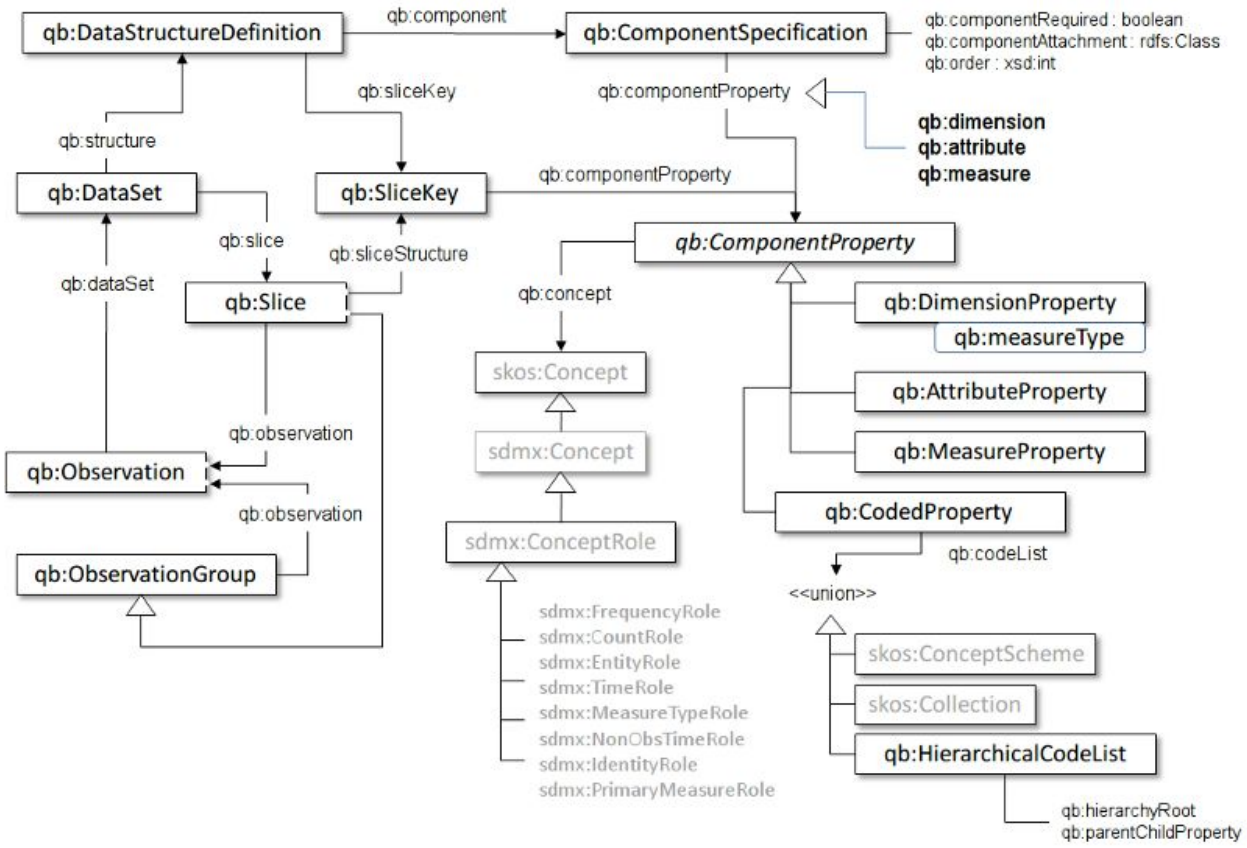


*Figure 1: Key terms and relationships in The RDF Data Cube Vocabulary, source: (Cyganiak & Reynolds, 2014)*

A data cube consists of dimensions, which describe properties of individual observations such as time period or geographical region. In the context of budget data, these are typically the fiscal year, organization and budget item category. Then there are measures representing the observed values such as height, width, amount, etc. Again, in the context of budget data, this is typically the budgeted amount of money. Finally, there are attributes, which specify additional properties of the measures, such as unit of measurement or multiplicator. In budget data, this is typically the currency of the budget. Data cube can be sliced by grouping observations with the same values on selected dimensions, e.g., budget items for a selected fiscal year and a specific organization. Using the RDF Data Cube Vocabulary we model the data structure definition using components (dimensions, attributes, measures) and then the defined components are used to classify individual observations.

## OpenBudgets.eu RDF prefixes

For OpenBudgets.eu we will use the following RDF prefixes based on a similar approach for SDMX[3]:

[3] https://github.com/UKGovLD/publishing-statistical-data/tree/master/specs/src/main/vocab

- `obeu: <http://data.openbudgets.eu/ontology/> .`
- `obeu-dsd: <http://data.openbudgets.eu/resource/dsd/> .`
- `obeu-dimension: <http://data.openbudgets.eu/ontology/dsd/dimension/> .`
- `obeu-measure: <http://data.openbudgets.eu/ontology/dsd/measure/> .`
- `obeu-attribute: <http://data.openbudgets.eu/ontology/dsd/attribute/> .`
- `obeu-codelist: <http://data.openbudgets.eu/resource/codelist/> .`
- `obeu-metadata: <http://data.openbudgets.eu/ontology/metadata/> .`

And then there are prefixes for defined code lists:

- `obeu-operation: <http://data.openbudgets.eu/resource/codelist/operation-character/> .`
- `obeu-budgetphase: <http://data.openbudgets.eu/resource/codelist/budget-phase/> .`

# Definition of core component properties for budget data

In this section, we describe identified core component properties. Component properties to consider have been identified in Deliverable D1.1 either as occurring in datasets, although using various identifiers, or they were considered important by domain experts based on knowledge elicitation. The core component properties were then selected as the most significant ones based on their frequent occurrence in datasets and the intended use of budget data based on the user requirements proceeding from the project's use cases.

## Core dimensions

Values of dimensions uniquely identify the measured value (observation). For budget data, the core dimensions are defined as follows:

### Budgetary unit

The budgetary unit dimension identifies the entity that plans the budget. The budgetary unit itself is an instance of `org:Organization`[4]. It is defined as:

```
obeu-dimension:budgetaryUnit a rdf:Property, qb:DimensionProperty,
qb:CodedProperty ;
      rdfs:label "budgetary unit"@en ;
      rdfs:range org:Organization .
```

**Technical note:** Note that the dimension has three types, `rdf:Property`, `qb:DimensionProperty` and `qb:CodedProperty`. According to the DCV specification, for

[4] http://www.w3.org/TR/vocab-org/

example `qb:DimensionProperty` is a subproperty of `rdf:Property`, so the `rdf:Property` type could seem superfluous. However, it is here to support applications that do not do type inference themselves and could use the information that the dimension is in fact also a property.

### Fiscal period

The fiscal period is the period of time reflected in financial statements. We use the more general term fiscal period instead of the more common term fiscal year because in our survey we found out that some budget datasets are published quarterly. We reuse the data.gov.uk representations of time intervals[5] to specify it:

```
obeu-dimension:fiscalPeriod a rdf:Property, qb:DimensionProperty,
qb:CodedProperty ;
    rdfs:label "fiscal period"@en ;
    rdfs:subPropertyOf sdmx-dimension:refPeriod ;
    rdfs:range time:Interval ;
    qb:concept sdmx-concept:refPeriod .
```

### Fiscal year

The fiscal year is a more common dimension among the surveyed datasets. It is defined as a subproperty of fiscal period:

```
obeu-dimension:fiscalYear a rdf:Property, qb:DimensionProperty,
qb:CodedProperty ;
    rdfs:label "fiscal year"@en ;
    rdfs:comment "The year reflected in financial statements."@en ;
    rdfs:subPropertyOf obeu-dimension:fiscalPeriod ;
    rdfs:range interval:Year ;
    qb:concept sdmx-concept:refPeriod .
```

### Classification

There is a vast amount of classifications used to classify budget lines. Although we cannot cover all of them, we can at least split them into categories. All dimensions that are classifications should be subproperties of:

```
obeu-dimension:classification a rdf:Property, qb:DimensionProperty,
qb:CodedProperty ;
     rdfs:subPropertyOf dcterms:subject ;
     rdfs:label "classification"@en .
```

In our survey in D1.1 we have identified 4 recurring types of classifications:

#### Functional classification

The functional classification categorizes expenditures according to the purposes and objectives for which they are intended. When there is only one functional classification used in the given dataset, this property should be used. All other functional classification dimensions should be subproperties of:

[5] http://datahub.io/dataset/data-gov-uk-time-intervals

```
obeu-dimension:functionalClassification a rdf:Property,
qb:DimensionProperty, qb:CodedProperty ;
     rdfs:label "functional classification"@en ;
     rdfs:comment "Categorizes expenditures according to the
purposes and objectives for which they are intended."@en ;
     rdfs:subPropertyOf obeu-dimension:classification .
```

### Programme classification

The programme classification is a grouping of expenditures by common objective for budgeting purposes. When there is only one programme classification used in the given dataset, this property should be used. All other programme classification dimensions should be subproperties of:

```
obeu-dimension:programmeClassification a rdf:Property,
qb:DimensionProperty, qb:CodedProperty ;
     rdfs:label "programme classification"@en ;
     rdfs:comment "Grouping of expenditure by common objective for
budgeting purposes."@en ;
     rdfs:subPropertyOf obeu-dimension:classification .
```

### Economic classification

The economic classification identifies the type of expenditure incurred. When there is only one economic classification used in the given dataset, this property should be used. All other economic classification dimensions should be subproperties of:

```
obeu-dimension:economicClassification a rdf:Property,
qb:DimensionProperty, qb:CodedProperty ;
     rdfs:label "economic classification"@en ;
     rdfs:comment "Identifies the type of expenditure incurred."@en
;
     rdfs:subPropertyOf obeu-dimension:classification .
```

### Administrative classification

The administrative classification identifies the entity responsible for managing the public funds concerned. This can be for example a specific department of a municipality, whereas the budgetary unit is the municipality itself. When there is only one administrative classification used in the given dataset, this property should be used. All other administrative classification dimensions should be subproperties of:

```
obeu-dimension:administrativeClassification a rdf:Property,
qb:DimensionProperty, qb:CodedProperty ;
     rdfs:label "administrative classification"@en ;
     rdfs:comment "Identifies the entity responsible for managing
the public funds concerned."@en ;
     rdfs:subPropertyOf obeu-dimension:classification .
```

### Operation character

There are budget datasets that contain revenues and expenditures. To distinguish among those, a dimension property and a corresponding code list are defined:

```
obeu-dimension:operationCharacter a rdf:Property,
qb:DimensionProperty, qb:CodedProperty ;
      rdfs:label "operation character"@en ;
      rdfs:comment "Distinguishes among expenditure and revenue."@en
;
      rdfs:range obeu:OperationCharacter ;
      qb:codeList obeu-codelist:operationCharacter .
```

The supplied code list has 2 items: `obeu-operation:Expenditure`, `obeu-operation:Revenue`.

This dimension is paired with the supplied codelist and from knowledge elicitation it seems that these two operation characters are widely adopted. However, some datasets distinguish transactions in financial assets and liabilities (such as loans) as a separate operation character. In such case, a dataset-specific version of this dimension property with an extended code list can be devised.

### Budget phase

Many budget datasets capture the state of the budget at multiple points in time. We call those the budget phases. To distinguish among budget phases, a dimension property and a corresponding code list are defined. Note that various authorities publish budgets in various phases. It is expected that each dataset may have its own budget phase dimension as a subproperty and a code list mapped to the provided one.

```
obeu-dimension:budgetPhase a rdf:Property, qb:DimensionProperty,
qb:CodedProperty ;
      rdfs:label "budget phase"@en ;
      rdfs:comment "Distinguishes among phases of the budget."@en ;
      rdfs:range obeu:BudgetPhase ;
      qb:codeList obeu-codelist:budgetPhase .
```

The supplied code list has four items: `obeu-budgetphase:Draft`, `obeu-budgetphase:Revised`, `obeu-budgetphase:Approved` and `obeu-budgetphase:Executed`.

## Core measures

### Amount

In the surveyed approaches there were various types of financial amounts regarding budgets, typically the amounts in budget lines in various stages of the budget. In OpenBudgets.eu we model budget phases as a dimension. With this in mind, we have identified a single measure -

the budgeted amount. Therefore, all budget datasets should use this measure to specify the amount in the budget line:

```
obeu-measure:amount a rdf:Property, qb:MeasureProperty ;
      rdfs:label "amount"@en ;
      rdfs:comment "The amount budgeted."@en ;
      rdfs:subPropertyOf sdmx-measure:obsValue ;
      rdfs:range xsd:decimal ;
      qb:concept sdmx-concept:obsValue .
```

Note that we specify `xsd:decimal` as the data type to be used with the amount measure. This is because we want to avoid loss of precision that could happen when `xsd:float` or `xsd:double` are used.

## Core attributes

### Currency

Currency of the budgeted amount is a very important fact that is often omitted in budget datasets. To improve comparability and machine readability, in OpenBudgets.eu all datasets must have their currency specified. The currency is modelled here as an attribute because it specifies a property of the measure (amount) rather than identify it[6]. Note that according to the RDF Data Cube Vocabulary, the attribute components can be attached at observation, slice or dataset level. This allows the datasets to have one or even multiple currencies.

```
obeu-attribute:currency a rdf:Property, qb:AttributeProperty,
qb:CodedProperty ;
    rdfs:label "currency"@en ;
    rdfs:comment "The currency of the financial amount"@en ;
    rdfs:subPropertyOf sdmx-attribute:currency ;
    rdfs:range obeu:Currency ;
    qb:concept sdmx-concept:currency .
```

### Taxes included

Taxes included attribute indicates whether the reported monetary amount includes taxes. It is a boolean property that can be either true or false. This mirrors the typical granularity of budget data, in which either all applicable taxes or no taxes are included. Examples of common taxes included in expenditure budget lines are value-added tax or excise duty (e.g., for fuel). Taxes typically apply only to expenditure budget lines and are not included in revenue budget lines. In some cases, payments of taxes are reported in dedicated budget lines aggregated by the tax type.

```
obeu-attribute:taxesIncluded a rdf:Property, qb:AttributeProperty,
qb:CodedProperty ;
  rdfs:label "tax included"@en ;
```

[6] Currency could be modelled as a dimension when we would model a dataset like currency exchange rates where the currency would be part of the measure identifier.

```
  rdfs:comment "Indicates whether the reported amount includes
taxes."@en ;
  rdfs:range xsd:boolean .
```

# Metadata

Data Cube Vocabulary recommends describing datasets with basic metadata[7], which we endorse. The purpose of providing metadata include helping users to discover the dataset via subject classifications, learning about conditions for reuse, guiding interpretation of its content by additional explanations, assessing dataset's dynamics based on modification timestamps, or establishing trustworthiness from provenance metadata. Based on the findings from previous interviews with domain experts we decided to extend the metadata elements recommended by DCV with others specifically geared towards the budgetary domain.

## DSD-level metadata

### Used methodology

In order to improve the chance to interpret a budget correctly, we should allow to link to the methodology that was used when the DSD was created. The need for explicit links to methodologies was emphasized by several interviewed domain experts, who expect it to help prevent misinterpretation. In order to be express links to documents describing the methodologies used to define DSDs we provide the `obeu-metadata:methodologyUsed` property.

# Additional component properties for specific datasets

As budget datasets are very variable in terms of the number and the nature of component properties (dimensions, measures and attributes), in OpenBudgets.eu we only specify a core set of components which should be used whenever appropriate. However, often it will be the case that additional data needs to be represented. In this case, additional component properties are to be defined using DCV similarly to how the core components were defined. For instance, in the case of classifications, new classification dimensions may be subproperties of the core classification dimension types.

# Creating DSDs using component properties

Using the provided core component properties and optional additional component properties a data structure definition (DSD) is to be created according to DCV to describe the structure of a given dataset. A data structure definition comprises components that specify which component properties are used. For attribute components, the `qb:componentAttachment` property can be useful. It defines whether a component will be specified for each observation or for the dataset as a whole. For example, the currency attribute will be typically attached to the dataset, meaning that each observation (budget line) from the dataset uses the same currency. While missing values of dimensions are invalid in DCV, values of attributes are optional. However,

[7] http://www.w3.org/TR/vocab-data-cube/#metadata

there may be attributes that are necessary for correct interpretation of data, such as currency. DCV allows to define such attributes as required using the `qb:componentRequired` property. This approach can be further extended to other attributes when designing DSDs for concrete datasets. Moreover, DSDs may optionally specify the order of components via the `qb:order` property. Doing so can be useful for presentation purposes and thus, similarly to ordering columns in a table, components may be highlighted by bringing them to the front or ordered in a logical way.

The mandatory components are (can be refined using a subproperty such as `fiscalYear`):

- `obeu-dimension:budgetaryUnit`
- `obeu-dimension:fiscalPeriod`
- `obeu-measure:amount`
- `obeu-attribute:currency`

Typically, there will be also some classifications, but those are not mandatory.

An example of an OpenBudgets.eu compliant budget DSD is:

```
obeu-dsd:Budget1 a qb:DataStructureDefinition ;
  rdfs:label "Sample budget Data Structure Definition"@en ;
  rdfs:comment "Made for D1.2 of OpenBudgets.eu"@en ;
  qb:component [
     rdfs:label "Budgetary unit"@en ;
     qb:dimension obeu-dimension:budgetaryUnit
  ], [
     rdfs:label "Fiscal period"@en ;
     qb:dimension obeu-dimension:fiscalPeriod
  ], [
     rdfs:label "Programme classification"@en ;
     qb:dimension obeu-dimension:programmeClassification
  ], [
     rdfs:label "Currency"@en ;
     qb:attribute obeu-attribute:currency ;
     qb:componentAttachment qb:DataSet ;
     qb:componentRequired true
  ], [
     rdfs:label "The budgeted amount"@en ;
     qb:measure obeu-measure:amount
  ].
```

There can be a situation where we have a dataset where for some budget lines the dimension value is missing. In that case, either the data is incomplete and needs to be fixed or the component is in fact an attribute, not a dimension. The values on dimensions should together

uniquely identify the measure. With a missing value on a dimension, one should be unable to uniquely identify the measure (incomplete data) or it is an attribute - additional information related to the measure, but not necessary to identify it.

## Example: EU budget

Budget of the European Union is one of the key datasets the OpenBudgets.eu project will work with. In particular, it is a fundamental dataset for the transparency use case in WP6. This led us to select it to demonstrate how the proposed component properties can be used to assemble a DSD for a concrete dataset.

The EU budget is organized hierarchically by the Activity-Based Budgeting nomenclature that divides it into sections, chapters, articles, items, and sub-items. For example, in the 2014's budget, there is an article *"Coordination and promotion of awareness on development issues"*. Each budget contains data about 3 years: the current fiscal year (represented as *n*), the previous fiscal year (*n-1*), and the year before that (*n-2*). While *n* and *n-1* contain approved budget lines, budget lines for *n-2* are already executed and the final aggregated spending is reported. Budget lines are classified using the Multiannual Financial Framework Political Category (MFF/CATPOL) headings. For example, the previously mentioned article is classified as *"Actions financed under the prerogatives of the Commission and specific competences conferred to the Commission"*.

For the purpose of the example, we will be using the EU budget from 2014, which is the latest one released on the European Union Open Data Portal[8]. The source dataset is available in XML with a simplified CSV view. More information (e.g., draft amending budgets) is available only in PDF documents[9].

Thanks to the flexibility of the proposed data model we can cherry-pick only the component properties that match the structure of the dataset in question and fill in specific properties by extending the core data model. To model the EU budget we will reuse several core component properties defined above and mint subproperties of the core component properties to capture more specific aspects of the dataset. We directly reuse `obeu-dimension:budgetaryUnit`, `obeu-dimension:budgetPhase`, `obeu-dimension:fiscalYear`, `obeu-attribute:currency`, and `obeu-measure:amount`. In case of EU budget there is a single budgetary unit, which is the European Union, so that it can be attached on the dataset level. Budget lines in the dataset are either approved (`obeu-budgetphase:Approved`), which is the case for the years *n* and *n-1*, or they are executed (`obeu-budgetphase:Executed`) in the case of the year *n-2*. Since EU budget is a dataset with a single currency (EUR), the `obeu-attribute:currency` attribute too can be attached at the dataset level. Monetary amounts associated with budget lines are described using the `obeu-measure:amount` property. To cover the remaining parts of the dataset we extend

[8] http://open-data.europa.eu/data/dataset/budget-of-the-european-union-2014
[9] http://eur-lex.europa.eu/budget/www/index-en.htm

`obeu-dimension:classification`. Since neither Activity-Based Budgeting nomenclature, nor MFF/CATPOL fit the predefined classification subproperties, we create specific subproperties for them. The dataset contains only expenditures, but these may be further classified as either commitments or payments. To be able to capture this distinction we define a subproperty of `obeu-dimension:operationCharacter` and extend its code list with concepts for commitment and payment that specialize `obeu-operation:Expenditure`. We define one new attribute property for reserves, which are particular for the EU budget and thus cannot be mapped to any of the core component properties. Reserves account for funding set aside for some budget lines. Currency of the reserves is the same as the measure's currency.

Complete DSD for the EU budget can be found in the Appendix C. Example budget line described in terms of the DSD is shown in the Appendix D.

## Conclusions

In this deliverable, we specified the core component properties of OpenBudgets.eu budget RDF data representation identified based on the survey and knowledge elicitation with domain experts presented in deliverable D1.1. We presented means of creating additional component properties where necessary as well as means of creating data structure definitions out of these component properties. We illustrated the process with an example of the European Union budget.

## References


- Klímek J., Kučera J., Mynarz J., Sedmihradská L., Zbranek J.: OpenBudgets.eu - Deliverable D1.1 - Survey of modelling public spending data & Knowledge elicitation report, 2015, https://openbudgets.atlassian.net/browse/OB-12


## Appendix A - Core component properties

This appendix (`budget-components.ttl`) contains the definition of core component properties in RDF (Turtle) using the RDFS and DCV vocabularies.

https://github.com/openbudgets/data-model/blob/master/budget/budget-components.ttl

## Appendix B - Core code lists

This appendix (`budget-codelists.ttl`) contains a preliminary definition of core code lists in RDF (Turtle) using the RDFS and DCV vocabularies.

https://github.com/openbudgets/data-model/blob/master/budget/budget-codelists.ttl

## Appendix C - Example DSD and component properties

This appendix (`eu-budget-dsd.ttl`) contains an example DSD and component properties for the budget of the European Union in RDF (Turtle). It shows the reuse of core component

properties, definition of new component properties and the creation of the data structure definition for a given dataset.

https://github.com/openbudgets/data-model/blob/master/budget/eu-budget/dsd.ttl

## Appendix D - Sample data

This appendix (eu-budget-data.ttl) contains sample data in RDF (Turtle) using the example data structure definition from Appendix C.

https://github.com/openbudgets/data-model/blob/master/budget/eu-budget/example-data.ttl