Modelování dat pomocí slovníku DataCube

Doc. Ing. Vojtěch Svátek, Dr.

Zimní semestr 2017

http://nb.vse.cz/~svatek/rzzw.html

Motivace

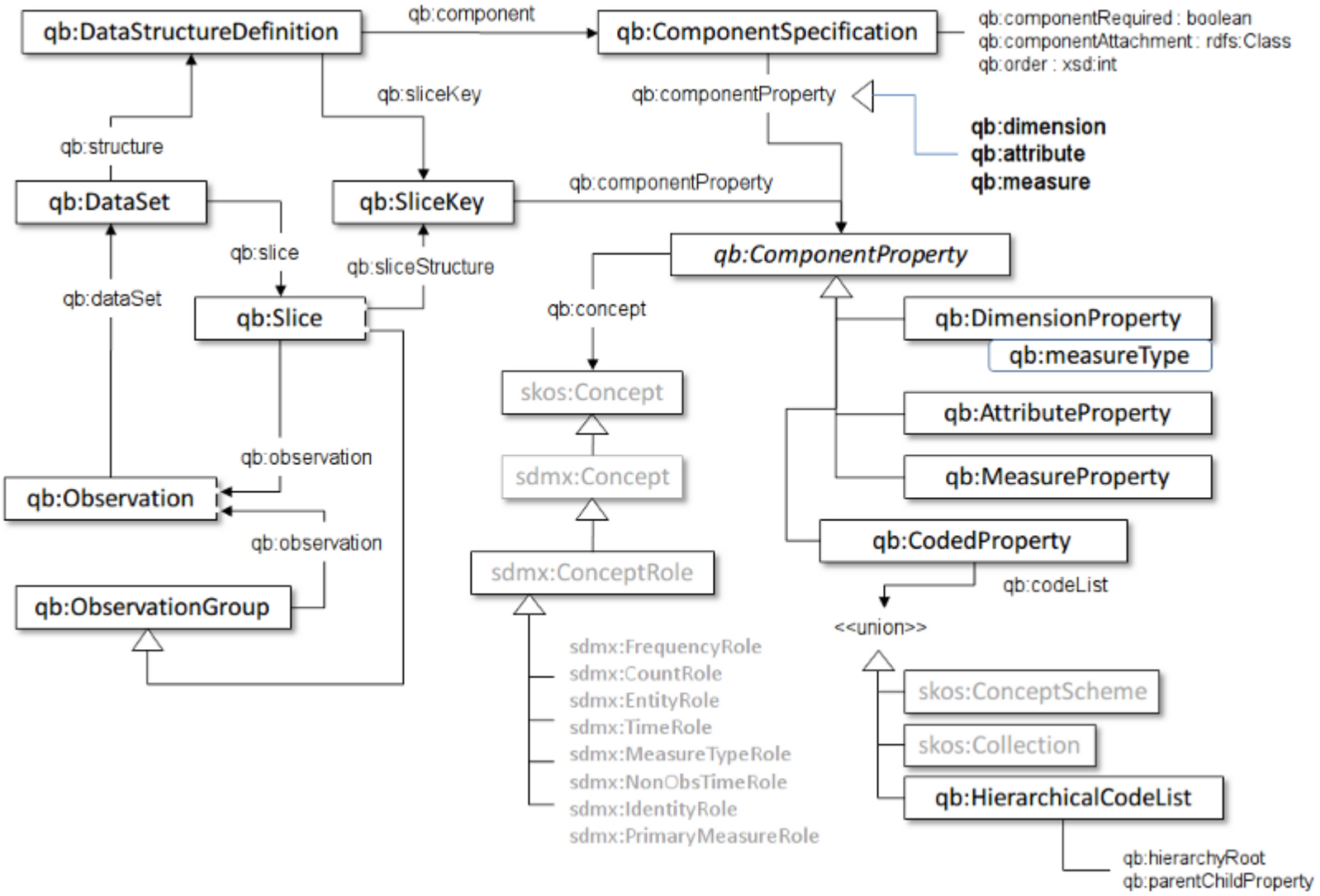
- Některá data jsou na webu zachycena na podrobné úrovni
  - Např. encyklopedická data na Wikipedii → DBpedia
  - Jejich agregace můžeme v RDF navrhovat pomocí SPARQL
- Jiná podrobná data jsou ale často důvěrná, ev. na jednotlivé úrovni (tzv. „mikrodata“ ve statistickém smyslu) málo významná; zveřejňují se proto až na úrovni agregací
  - Mluvíme zpravidla o „statistických“ datech
  - Agregované hodnoty (počty, součty, průměry apod.) se označují jako *míry*, a jsou vymezené různými *dimenzemi* tvořícími „datovou kostku“; v souhrnu proto mluvíme o *vícerozměrných* datech
  - I je můžeme nadále agregovat a kombinovat s jinými zdroji. Vzhledem k tomu, že primární data jsou zpravidla (ve srovnání např. s Wikipedií) relativně přesná a úplná, jejich využití pro agregované výstupy je pro věrohodné datové vizualizace vhodnější (nehledě na úsporný výpočet)

Základní pojmy

- Statistický *dataset* = množina pozorování („observation“)
- Pozorování jsou v rámci datasetu strukturována pomocí *komponent*: dimenzí, atributů a měr
  - Hodnoty *dimenzí* identifikují, o jakou část prostoru pozorování jde; obvykle je jednou z dimenzí *čas*
  - Hodnoty *měr* vyjadřují výsledek pozorování
  - *Atributy* poskytují upřesňující informace k mírám, např. jejich jednotku nebo platnost (předběžná/finální hodnota npod.)
- Viz struktura datových skladů (hvězda / sněhová vločka)
- Jednotlivá pozorování lze na základě hodnot dimenzí seskupovat do *řezů* (slices)
  - Řez vybírá pro většinu dimenzí konkrétní hodnotu; u zbylých dimenzí obsahuje všechny kombinace hodnot

Slovník Data Cube

- „Data Cube Vocabulary“ (DCV)
- Specifikace viz http://www.w3.org/TR/vocab-data-cube
- Obvyklý prefix „*qb:*“
- Opírá se o standard pro publikování statistických dat SDMX, https://sdmx.org/


qb:DataStructureDefinition
qb:component
qb:ComponentSpecification
qb:componentRequired : boolean
qb:componentAttachment : rdfs:Class
qb:order : xsd:int
qb:sliceKey
qb:componentProperty
qb:dimension
qb:attribute
qb:measure
qb:structure
qb:DataSet
qb:SliceKey
qb:componentProperty
qb:ComponentProperty
qb:slice
qb:sliceStructure
qb:dataSet
qb:Slice
qb:concept
qb:DimensionProperty
qb:measureType
skos:Concept
qb:AttributeProperty
qb:observation
sdmx:Concept
qb:MeasureProperty
qb:Observation
qb:observation
qb:CodedProperty
sdmx:ConceptRole
qb:codeList
qb:ObservationGroup
<<union>>
sdmx:FrequencyRole
sdmx:CountRole
sdmx:EntityRole
sdmx:TimeRole
sdmx:MeasureTypeRole
sdmx:NonObsTimeRole
sdmx:IdentityRole
sdmx:PrimaryMeasureRole
skos:ConceptScheme
skos:Collection
qb:HierarchicalCodeList
qb:hierarchyRoot
qb:parentChildProperty

Data structure definition

- Popis struktury statistických dat v RDF se vyjadřuje *definicí datové struktury* („data structure definition“ - DSD) jako instance `qb:DataStructureDefinition`
- Ta odkazuje na specifikace jednotlivých komponent pomocí vlastnosti `qb:component`
- V rámci specifikace v DSD lze komponentám určit
  - pořadí v rámci DSD: `qb:order` (pro dimenze)
  - zda jsou povinné: `qb:componentRequired` (pro atributy)
  - na co se aplikují: `qb:componentAttachment` (pro dimenze a atributy)

Propojení DSD na komponenty

- Specifikace pak odkazuje na komponentu samotnou pomocí vlastnosti `qb:componentSpecificaion` nebo jejích podvlastností `qb:dimension`, `qb:measure` nebo `qb:attribute`

Příklad DSD – demografie

(ze specifikace DCV, zjednodušeno)

```
eg:dsd-le3 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:componentRequired "true"^^xsd:boolean;
qb:componentAttachment qb:DataSet ] .
```

Příklad DSD – rozpočty

(z projektu OpenBudgets.eu)

… qb:component
[qb:dimension obeu-dimension:budgetaryUnit ;
qb:componentAttachment qb:DataSet],
[qb:dimension obeu-dimension:budgetPhase],
[qb:dimension eu-dimension:operationCharacter],
[qb:dimension obeu-dimension:fiscalYear],
[qb:dimension eu-dimension:budgetNomenclature],
[qb:dimension eu-dimension:catpol],
[qb:attribute obeu-attribute:currency ;
qb:componentRequired "true"^^xsd:boolean ;
qb:componentAttachment qb:DataSet],
[qb:attribute eu-attribute:reserve ;
qb:componentRequired "false"^^xsd:boolean],
[qb:measure obeu-measure:amount] .

Komponenty

- Komponenty jsou z hlediska RDF *vlastnostmi*
  - Nelze jim přiřazovat instance
- Popis komponenty obvykle zahrnuje minimálně
  - obor hodnot komponenty – `rdfs:range`
  - její `rdf:type`, tj. `qb:DimensionProperty`, `qb:AttributeProperty`, nebo `qb:MeasureProperty`
  - textový popis v `rdfs:label`, zpravidla i `rdfs:comment`
- I komponenty samotné lze mapovat na pojmy, pomocí `qb:concept`
  - „Měna“ může figurovat jako dimenze (pro kurzová data), ale i jako atribut (např. pro data o tržbách)

Příklady komponent – demografie

(ze specifikace DCV, zjednodušeno)

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

Příklady komponent – rozpočty

(z projektu OpenBudgets.eu)

```
eu-dimension:budgetNomenclature a rdf:Property,
qb:CodedProperty, qb:DimensionProperty ;
rdfs:label "Activity-Based Budgeting nomenclature 2014"@en ;
rdfs:subPropertyOf obeu-dimension:classification ;
qb:codeList eu-codelist:budget-nomenclature-2014 ;
rdfs:isDefinedBy
<http://example.openbudgets.eu/ontology/dsd/eu-budget-2014> .

eu-attribute:reserve a rdf:Property, qb:AttributeProperty ;
rdfs:label "Reserve"@en ; rdfs:range xsd:decimal ;
rdfs:isDefinedBy
<http://example.openbudgets.eu/ontology/dsd/eu-budget-2014> .
```

Dimenze

- Objektové vlastnosti - jejich hodnotou tedy nemůže být literál!
- Obor hodnot je definován pomocí kódovníku; ten může být implicitní nebo explicitní
- Explicitní kódovník
  - odkaz na něj se specifikuje vlastností `qb:codeList`
  - je nejčastěji vyjádřený pomocí schématu SKOS
  - pokud je rdfs:range dimenze skos:Concept, explicitní odkaz na kódovník je povinný

Dataset

- Konkrétní dataset je instancí `qb:DataSet`
  - Odkazuje na DSD pomocí `qb:structure`
- Jednotlivá pozorování jsou instancemi `qb:Observation`
  - Odkazují na dataset pomocí `qb:dataSet`
  - Musí vždy zahrnovat všechny dimenze i míry deklarované v DSD
    - atributy musí být přítomny pouze pokud jsou deklarované jako *required*
    - dimenze a atributy mohou být specifikovány hromadně pomocí `qb:componentAttachment` (tzv. denormalizace datasetu… z hlediska databázového by ale šlo spíš o normalizaci!)

Příklad datasetu a pozorování

(ze specifikace DCV, zjednodušeno)

```
eg:dataset-le1 a qb:DataSet;
    rdfs:label "Life expectancy"@en;
qb:structure eg:dsd-le .

eg:o1 a qb:Observation;
    qb:dataSet                  eg:dataset-le1 ;
    eg:refArea                  ex-geo: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 .
```

Přepoužití jiných slovníků

- Viz http://lov.okfn.org/dataset/lov/vocabs/qb
  - SKOS: různé taxonomie pojmů
  - Dublin Core Terms: metadata
  - Nepřímo další
    - VoiD: údaje o přístupnosti dat
    - FOAF: „agenti“
    - ORG: organizace