Validation: requirements and approaches

Dave Reynolds, Epimorphics Ltd
@der42

Validation requirements
based on experiences with data.gov.uk Linked Data

- Most current Linked Data in data.gov.uk is:
  - described using a range of vocabularies and documentation
  - validated , if at all, by publisher using internal/ad hoc tooling
- Emerging requirement for shared validation approach:
  - to enable interoperability
  - so publishers know the shape of data required
  - publishing tools can e.g. auto-populate forms
  - consuming tools know what to expect
- Key requirements:
  - declarative – easily inspectable by tools
  - declared – can locate the structure definition for a data set
  - accessible to mortals

A spread of requirements

- regular data
  - statistics, financial, environmental measurements, ...
- irregular data
  - organizational structure, strategic plans, ...
- controlled terms
  - code lists, regulated entities, geographic regions, ...

Regular data

- use Data Cube vocabulary
  - http://www.w3.org/TR/vocab-data-cube/
  - meets the requirements:
    - declarative specification of structure - Data Structure Definition (DSD)
    - declared: all observations link to DataSet link to DSD
    - fairly understandable:

```
:complianceDsd      a    qb:DataStructureDefinition;
    rdfs:label      "complianceDsd"@en;
    qb:component    [qb:dimension   :bathingWater],
                    [qb:dimension   :samplingPoint],
                    [qb:dimension   :sampleYear],
                    [qb:measure     :complianceClassification],
                    [qb:attribute   :inYearDetail];
    qb:sliceKey     :complianceByYearKey,
                    :complianceBySamplingPointKey .
```

But how to validate a data cube?

- Specification now defines “well-formed” cubes
  - closed world notion of compliance with DSD
  - integrity constraints specified by a set of SPARQL queries

- Lessons:
  - SPARQL was sufficient to express all the required ICs
  - some of the queries are convoluted and non-obvious
  - at least one is quadratically slow unless optimizer is magic
  - Useful compromise
    - SPARQL doesn’t meet requirements of inspectable and understandable
    - but tools and humans can operate at the DSD level

Irregular data

- typically mix-and-match range of vocabularies
  - declare usage via void:vocabulary
- target users find OWL impenetrable
- requirement for “vocabulary profiles”
  - closed-world constraints on properties (cardinalities, ranges)
  - expressivity of closed-world OWL would be sufficient
  - but need a presentation layer to simplify authoring and consumption – OSLC resource shapes?
  - discovery mechanism

Controlled terms

- the other 80% of the problem
  - common resource shapes the easy part
  - interoperability means re-using terms for things in the domain
- sets of controlled terms (URI sets, code lists etc)
  - can be very large
  - often managed by third parties independent of data publisher and vocabulary definer
  - can be dynamic
  - typically handled by some form of *registry*
    - governed, closed-world, lists of approved terms at point in time
- implication
  - need ability to validate against external services such as registries