daQ, an Ontology for Dataset Quality Information

Jeremy Debattista, Christoph Lange, Sören Auer

Presenter: Claus Stadler
Motivation

What are the *quality aspects* of a dataset for a particular domain?

- Quality of data is *subjective*
- Different domains require different quality attributes
- Data quality is commonly defined as *fitness for use*
Motivation (ii)

How can we *find* a good quality dataset?

http://www.datahub.io
The daQ is a light-weight, **extensible** vocabulary for **attaching** the results of quality **benchmarking** of a linked open dataset to that **dataset**

**daQ** (pronounced \ˈdək\)
Use Cases

Publishers are interested in publishing good quality data. But how can they convince the consumer?

- is the published data fit to use for its domain?
- how can publishers calculate the quality of a dataset and have this metadata part of it?
Use Cases (ii)

Consumers are interested in finding dataset which are *fit to use* in their domain.

- how can consumers *discover* certain aspects of a potential dataset?
- how can consumers *retrieve* datasets?
6th Star?

As a Consumer you can do all that ★★★★★ enables you to do, and additionally
✔ discovery good quality dataset

As a Publisher, ...
✔ make your data conform to domain quality metrics
✔ make your data more discoverable on certain quality aspects
A daq:QualityGraph is a *Named Graph*

✔ Separate aggregated metadata
✔ Digitally signed graphs using the swp:assertedBy
   (Semantic Web Publishing - Chris Bizer)

A daq:QualityGraph in theory can be computed on any resource but typically on a Dataset

http://purl.org/eis/vocab/daq
The daQ ontology is a generic framework, where classes and properties are defined in an abstract manner.
A category represents the highest level of quality assessment.
A dimension groups one or more metrics
Metric

The *smallest* unit of measuring a quality dimension
Using the daQ

T-Box

Accessibility  hasAvailabilityDimension  Category
hasRDFAccessibilityMetric  rdf:type  daq:hasAvailabilityDimension
RDF AvailabilityMetric
hasDoubleValue  rdf:type  daq:hasRDFAccessibilityMetric
xsd:double

A-Box

ex:access  daq:hasAvailabilityDimension
ex:avail  daq:hasRDFAccessibilityMetric
ex: rdfavail  daq:dateComputed
1.0  daq:hasDoubleValue
2014-01-23T14:53:00
Concluding Remarks

The daQ is a light-weight, *extensible* vocabulary for attaching the results of quality *benchmarking* of a linked open dataset to that *dataset*.

*Next Steps:*
- Extend the daQ framework with more concepts
- Represent more concrete quality metrics
- Dataset Retrieval based on Quality Metrics - extend a portal such as CKAN
Discussion

How can we sign the (dataset, qualitygraph) pair to make sure that:

a) the Quality Graph has not been tempered with
b) the Dataset is unchanged from the state in which the quality graph has been computed on?

Jeremy Debattista
jeremy.debattista@iais-extern.fraunhofer.de

Christoph Lange
math.semantic.web@gmail.com