

# daQ, an Ontology for Dataset Quality Information

Jeremy Debattista, Christoph Lange, Sören Auer

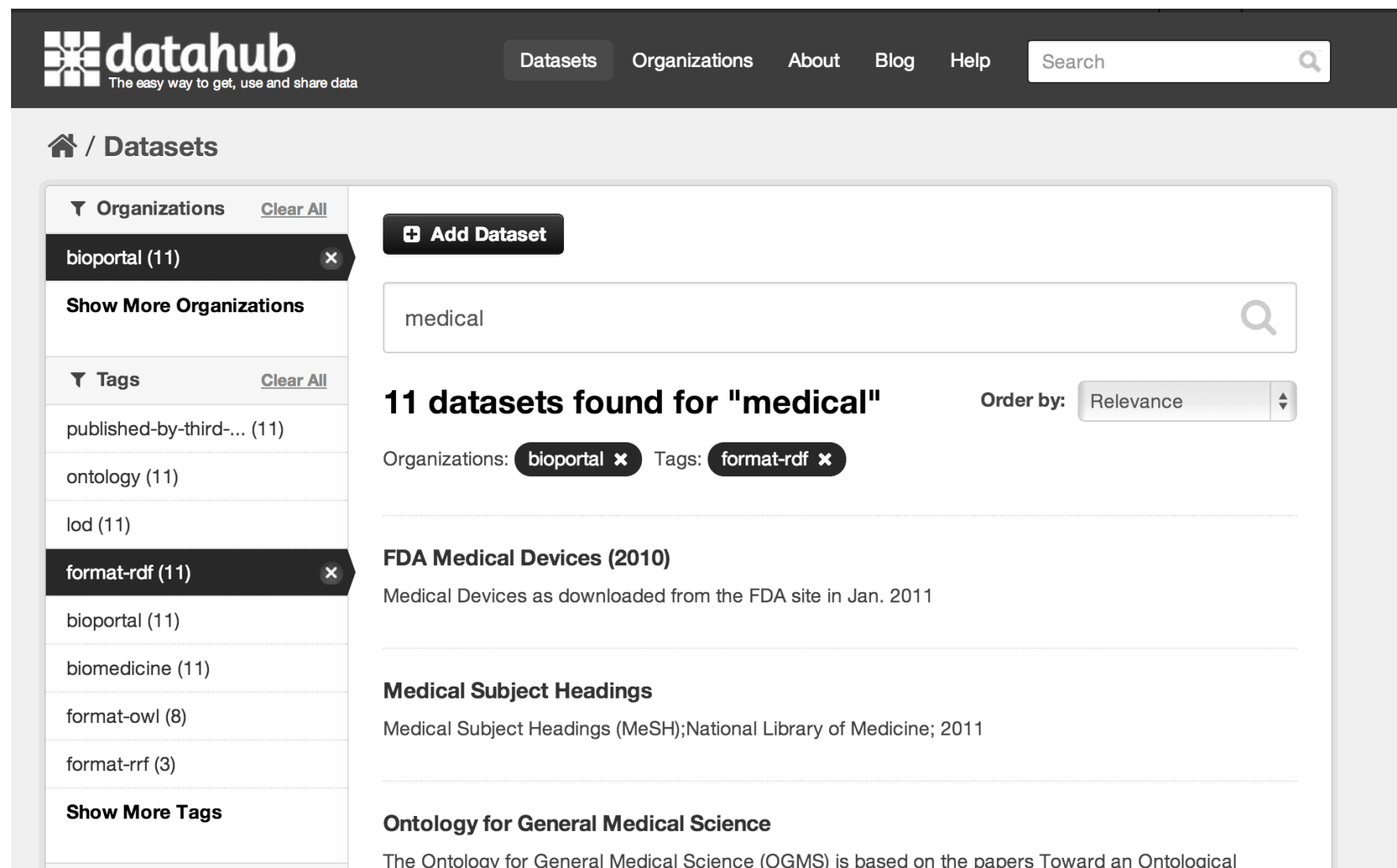
# Motivation

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?

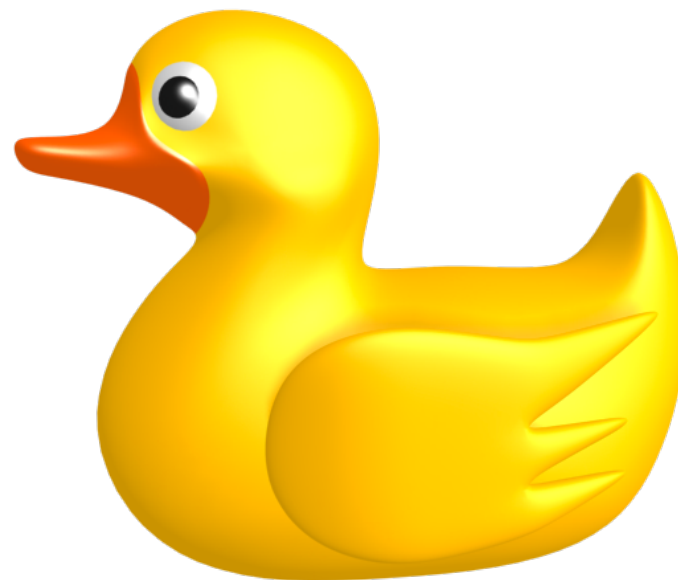


The screenshot shows the DataHub website interface. At the top, there's a navigation bar with the DataHub logo, the tagline 'The easy way to get, use and share data', and links for Datasets, Organizations, About, Blog, and Help. A search bar is also present. Below the navigation bar, the main content area is titled '/ Datasets'. On the left, there's a sidebar with filters for Organizations and Tags. The 'Organizations' filter shows 'bioportal (11)' and a 'Show More Organizations' link. The 'Tags' filter shows 'published-by-third-...' (11), 'ontology (11)', 'lod (11)', 'format-rdf (11)', 'bioportal (11)', 'biomedicine (11)', 'format-owl (8)', and 'format-rrf (3)', along with a 'Show More Tags' link. The main content area features an 'Add Dataset' button and a search bar containing the word 'medical'. Below the search bar, it states '11 datasets found for "medical"' and 'Order by: Relevance'. There are also filters for 'Organizations: bioportal x' and 'Tags: format-rdf x'. The first two results are 'FDA Medical Devices (2010)' and 'Medical Subject Headings'. The third result is 'Ontology for General Medical Science'.

<http://www.datahub.io>

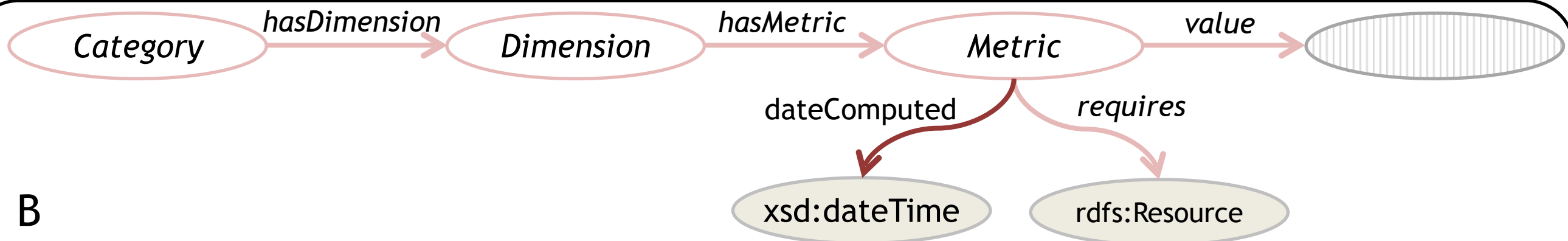
# Dataset Quality Ontology

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



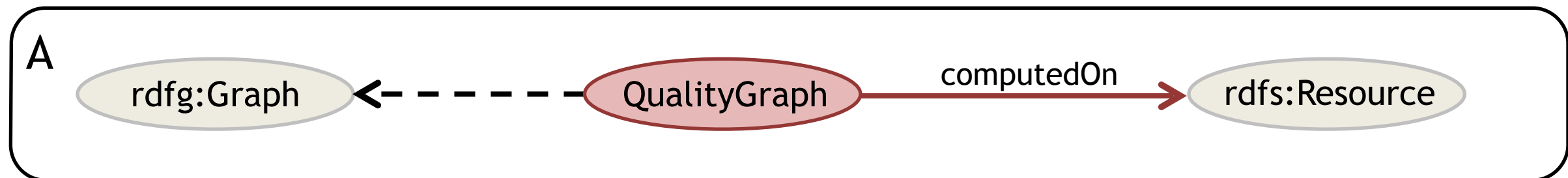
**daQ (pronounced \'dæk\')**

# daQ Ontology - Overall Framework



<http://purl.org/eis/vocab/daq>

# daQ Ontology

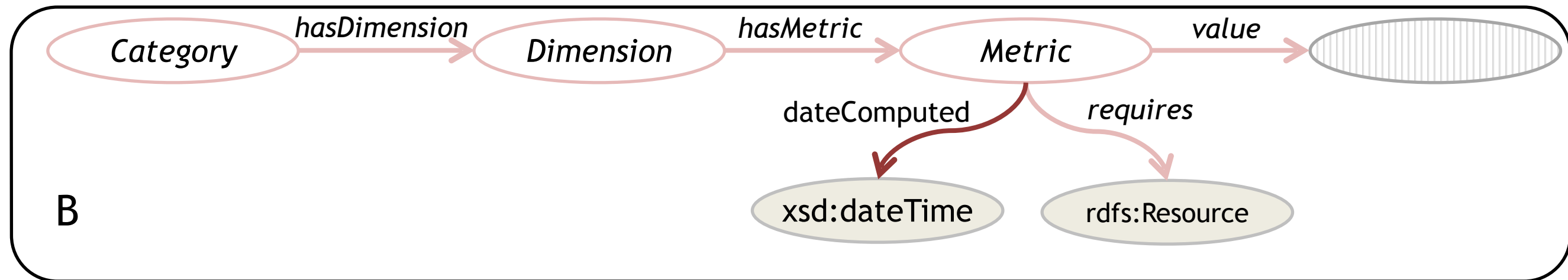


A daq:QualityGraph is a *Named Graph*

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

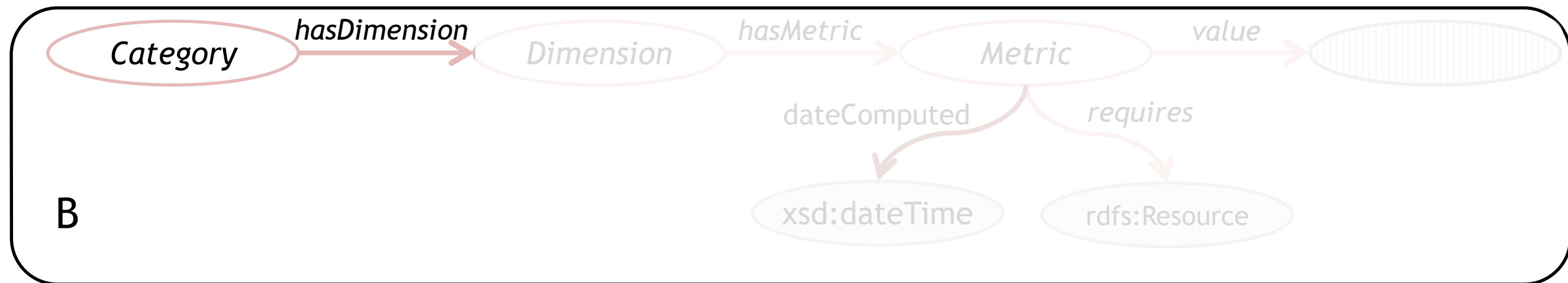
A daq:QualityGraph in theory can be computed on any resource but typically on a Dataset

# daQ Ontology (ii)



The daQ ontology is a generic *framework*, where classes and properties are defined in an *abstract manner*

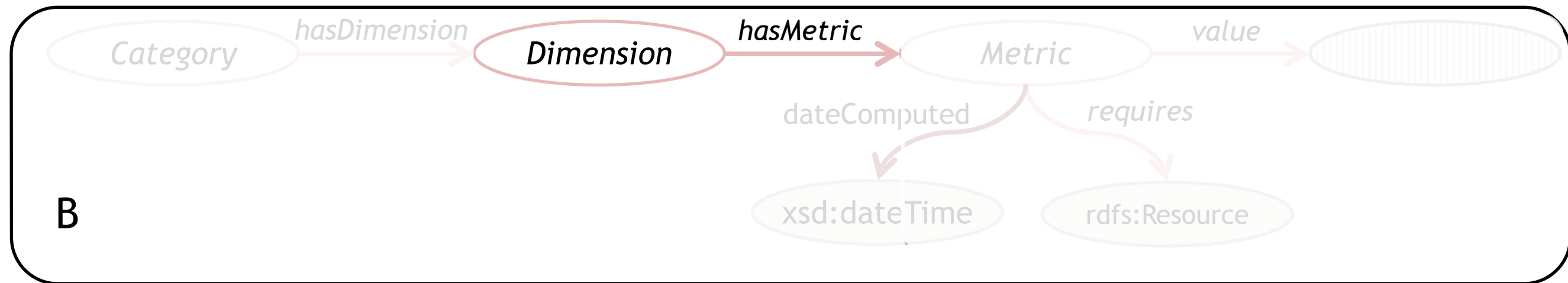
# Category



A category represent the highest level of quality assessment

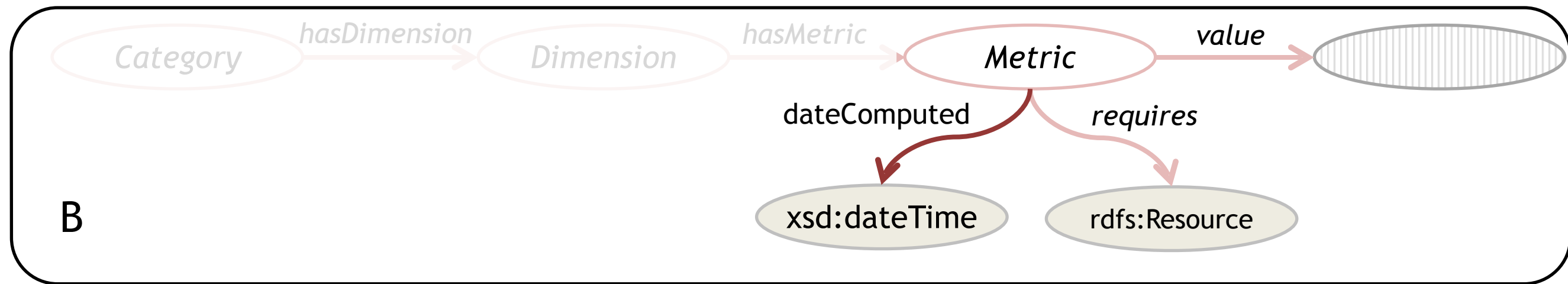


# Dimension



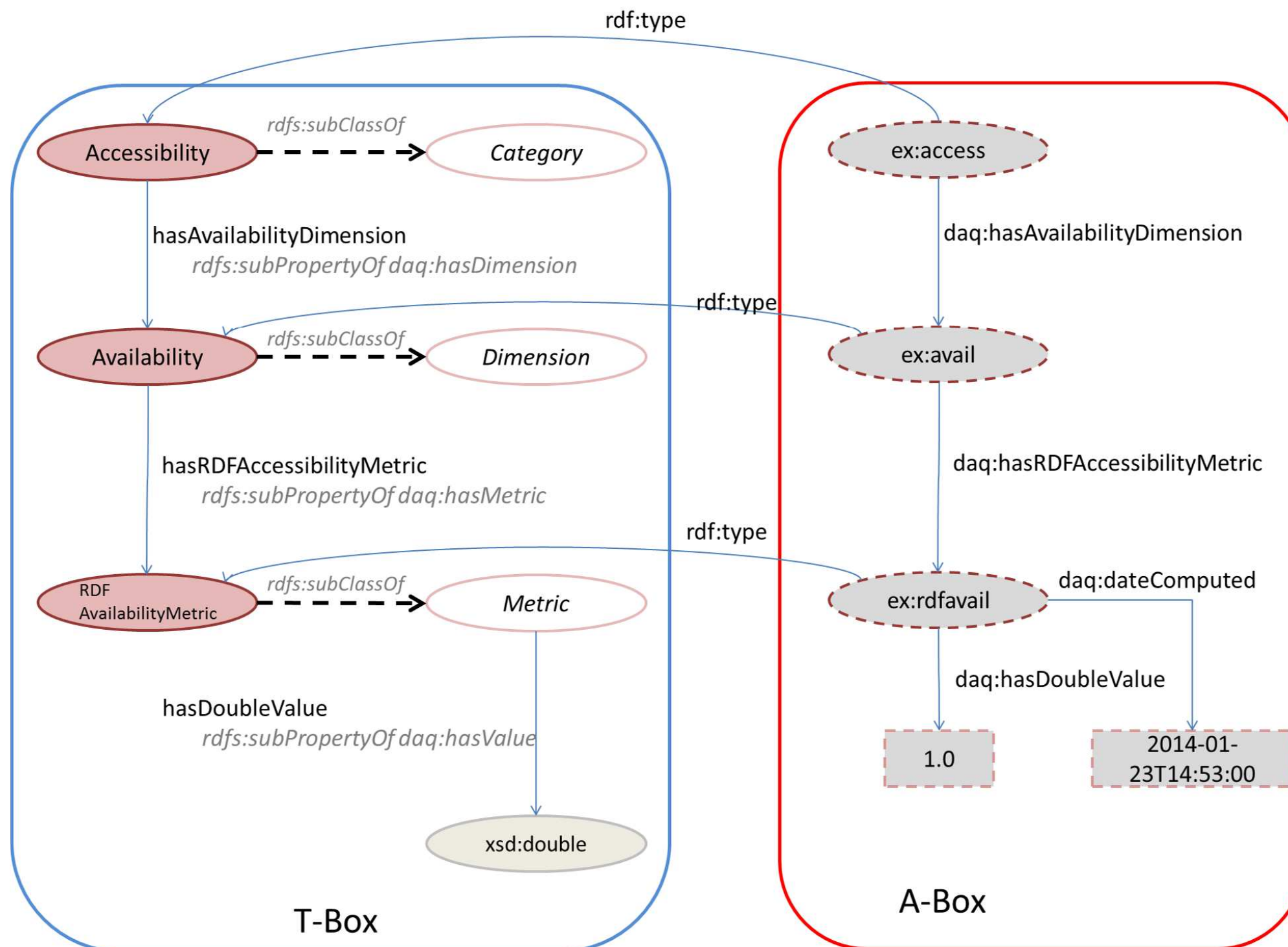
A dimension groups one or more metrics

# Metric



The *smallest* unit of measuring a quality dimension

# Using the daQ



# Concluding Remarks

*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](mailto:jeremy.debattista@iais-extern.fraunhofer.de)

Christoph Lange  
[math.semantic.web@gmail.com](mailto:math.semantic.web@gmail.com)