

daQ, an Ontology for Dataset Quality Information

Jeremy Debattista, Christoph Lange, Sören Auer

Presenter: Claus Stadler

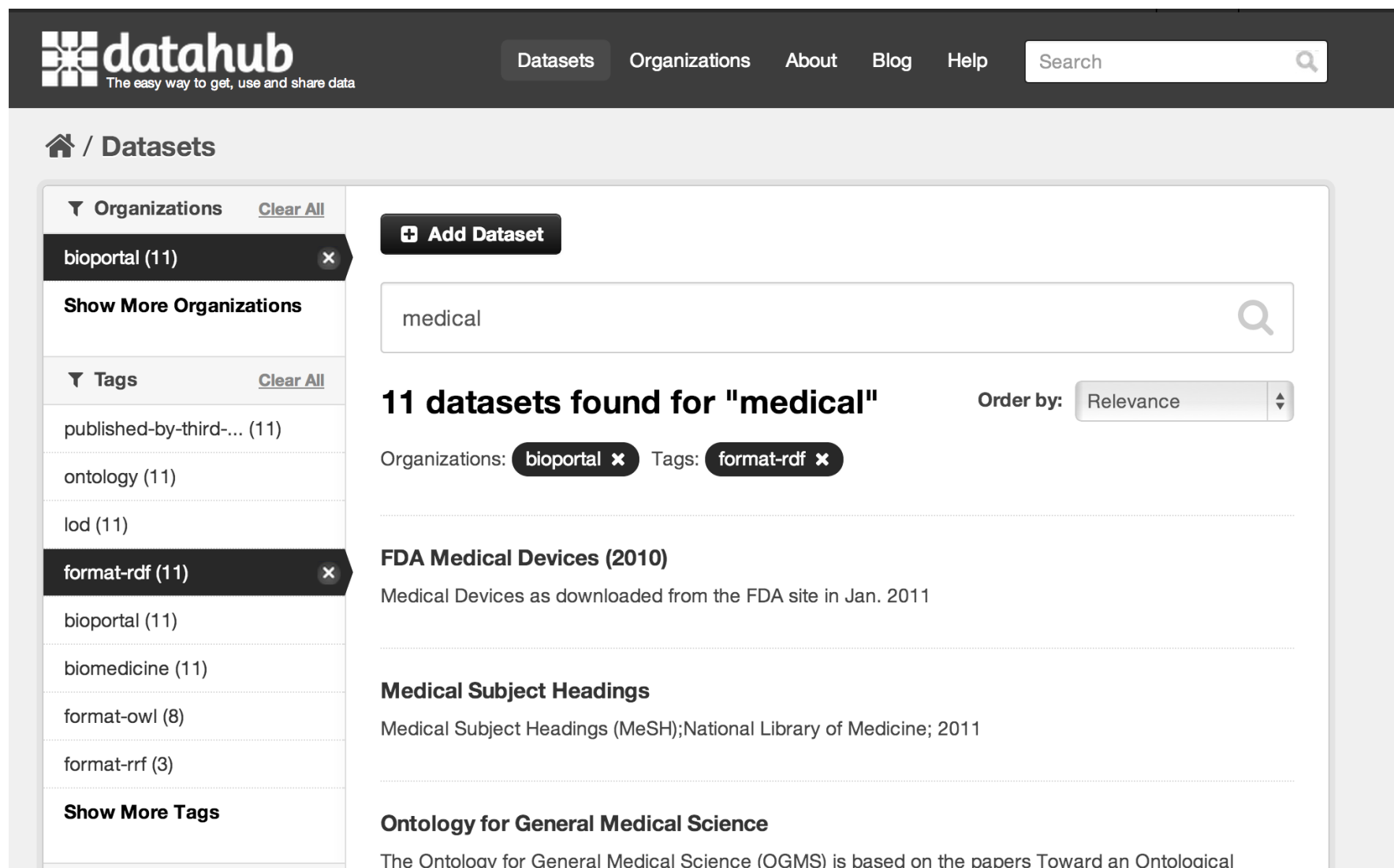
Motivation

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?



The screenshot shows the DataHub website interface. At the top, there's a navigation bar with the DataHub logo, a search bar, and links for Datasets, Organizations, About, Blog, and Help. 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)' selected. The 'Tags' filter shows 'format-rdf (11)' selected. The main search area has a search bar with the text 'medical' and a magnifying glass icon. Below the search bar, it says '11 datasets found for "medical"'. To the right of this, there's a dropdown menu for 'Order by' set to 'Relevance'. Below the search results, there are three dataset entries: 'FDA Medical Devices (2010)', 'Medical Subject Headings', and 'Ontology for General Medical Science'. Each entry has a brief description.

datahub
The easy way to get, use and share data

Datasets Organizations About Blog Help Search

/ Datasets

Organizations [Clear All](#)

bioportal (11) [×](#)

Show More Organizations

Tags [Clear All](#)

published-by-third-... (11)

ontology (11)

lod (11)

format-rdf (11) [×](#)

bioportal (11)

biomedicine (11)

format-owl (8)

format-rrf (3)

Show More Tags

[+ Add Dataset](#)

medical

11 datasets found for "medical" Order by: Relevance

Organizations: [bioportal](#) [×](#) Tags: [format-rdf](#) [×](#)

FDA Medical Devices (2010)
Medical Devices as downloaded from the FDA site in Jan. 2011

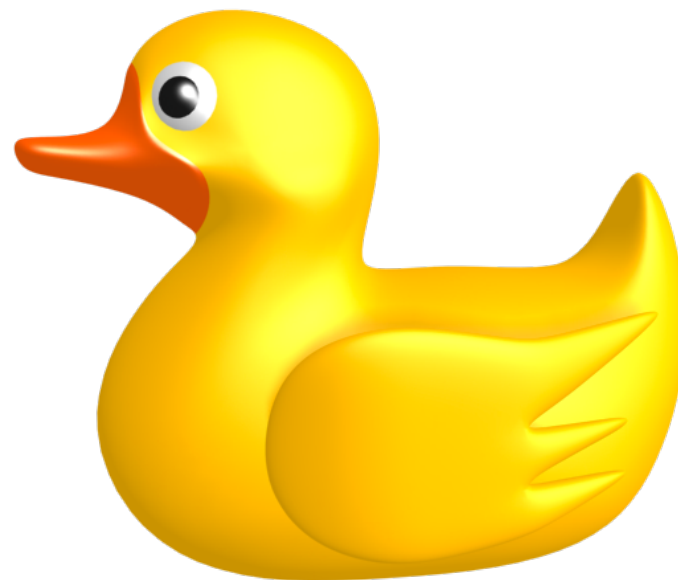
Medical Subject Headings
Medical Subject Headings (MeSH); National Library of Medicine; 2011

Ontology for General Medical Science
The Ontology for General Medical Science (OGMS) is based on the papers Toward an Ontological

<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\')

Use Cases

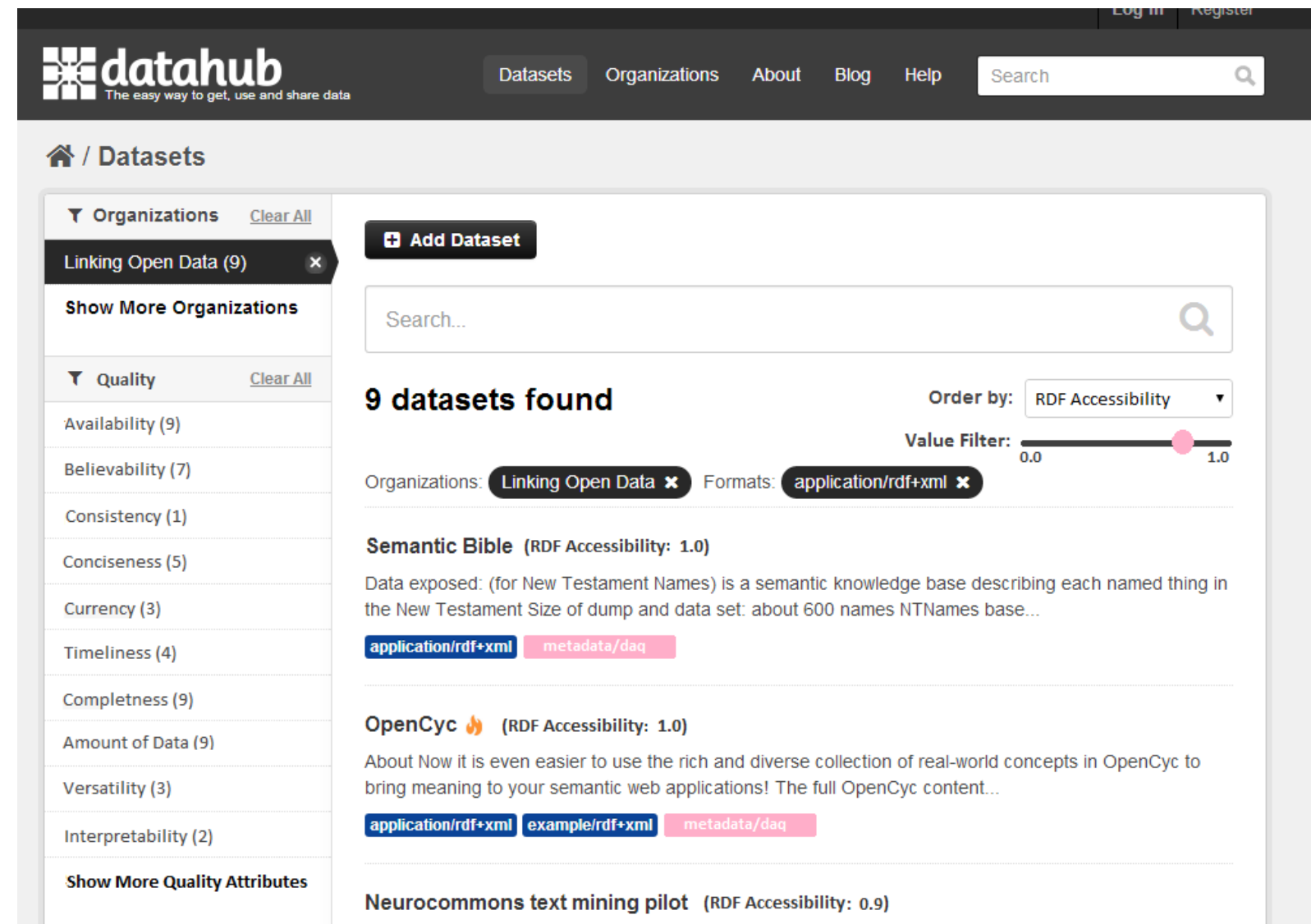
Publishers are interested in *publishing good quality* data. But how can they *convince* the consumer?

- 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)

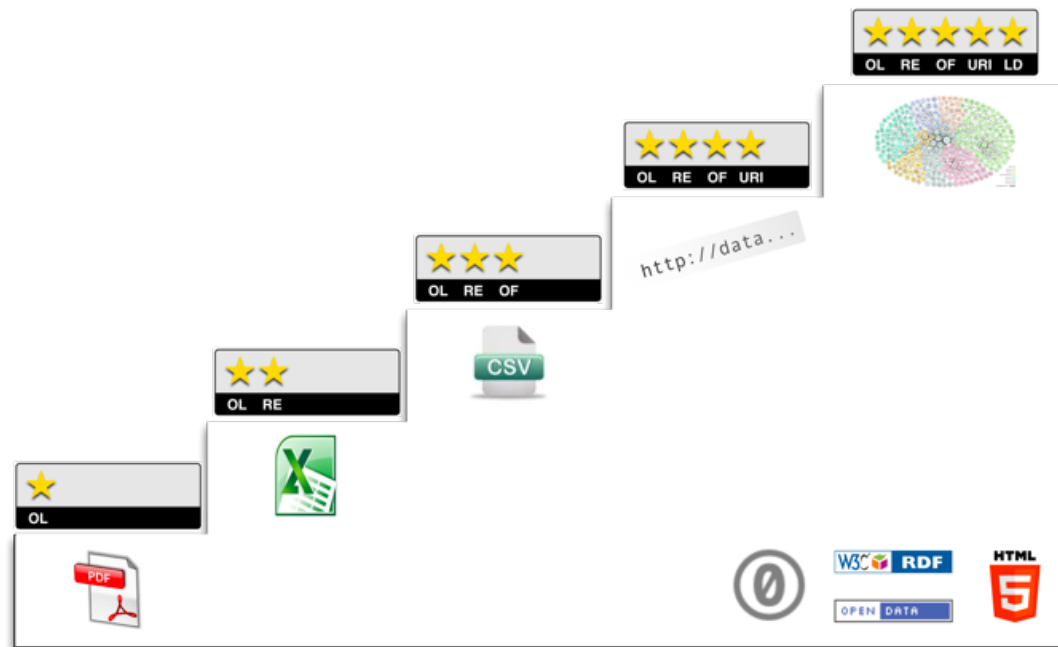
Consumers are interested in finding dataset which are *fit to use* in their domain.

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

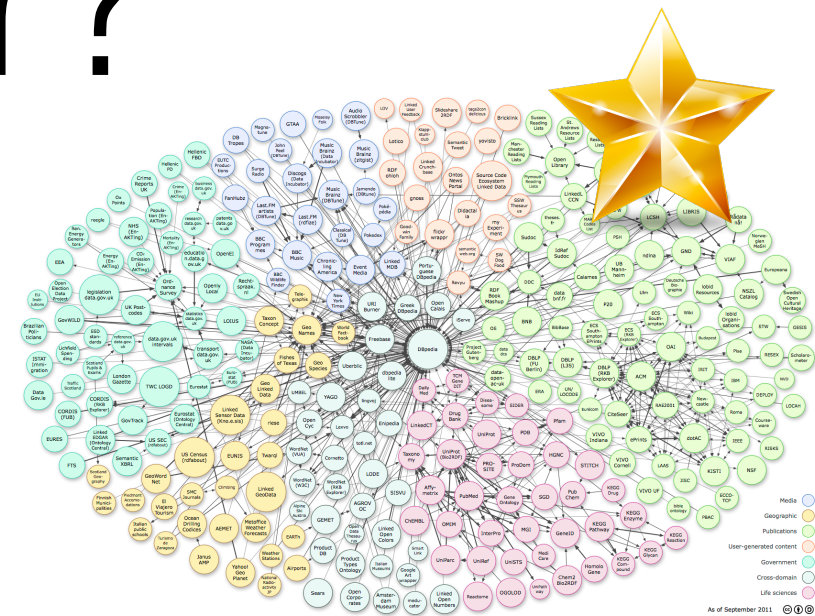


The screenshot shows the DataHub website interface. The top navigation bar includes links for Datasets, Organizations, About, Blog, and Help, along with a search bar. The main content area is titled 'Datasets' and features a sidebar with filters for Organizations and Quality. The Quality filter is expanded, showing various attributes like Availability, Believability, Consistency, Conciseness, Currency, Timeliness, Completeness, Amount of Data, Versatility, and Interpretability. The main search results area displays '9 datasets found' and includes a search bar, an 'Add Dataset' button, and a 'Value Filter' slider. The first two results are 'Semantic Bible' and 'OpenCyc', both with an RDF Accessibility of 1.0. The third result is 'Neurocommons text mining pilot' with an RDF Accessibility of 0.9.

6th Star?



<http://www.5stardata.info>



OL RE OF URI LD **DAQ**

As a Consumer you can do all that ★★★★★ enables you to do, and additionally

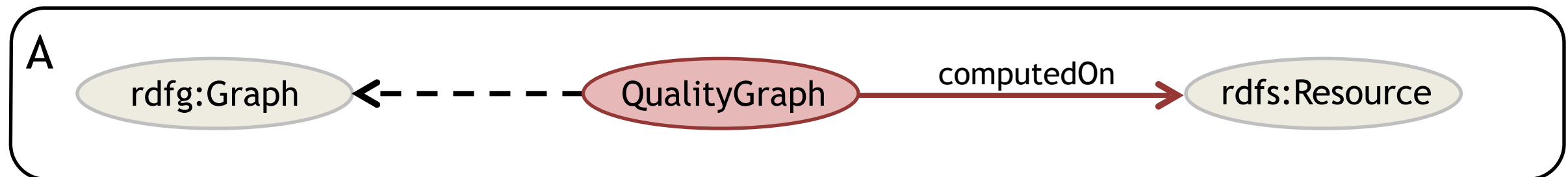
✓ discovery good quality dataset

As a Publisher, ...

✓ make your data conform to domain quality metrics

✓ make your data more discoverable on certain quality aspects

daQ Ontology



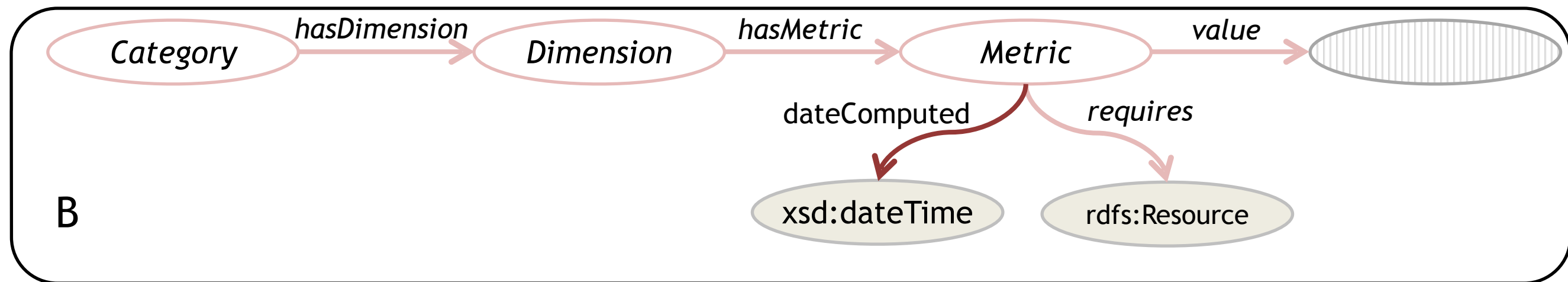
<http://purl.org/eis/vocab/daq>

A daq:QualityGraph is a *Named Graph*

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

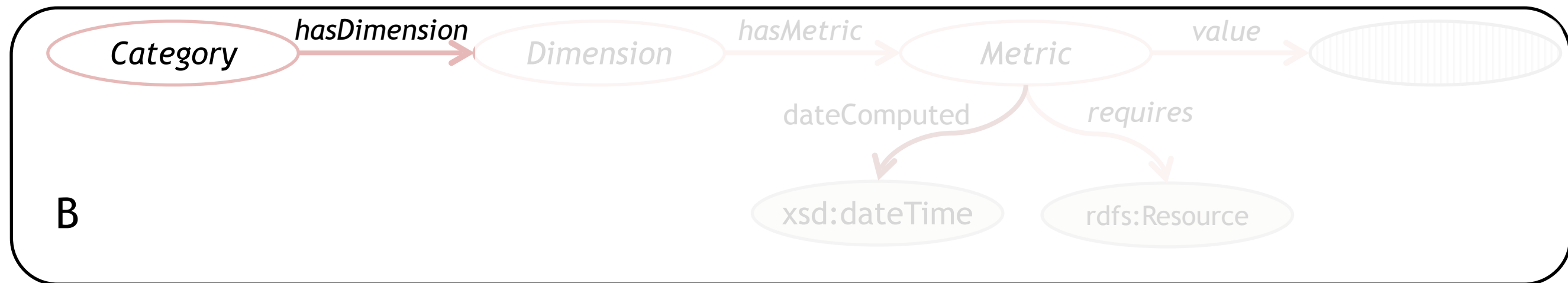
A daq:QualityGraph in theory can be computed on any resource but typically on a Dataset

daQ Ontology (ii)



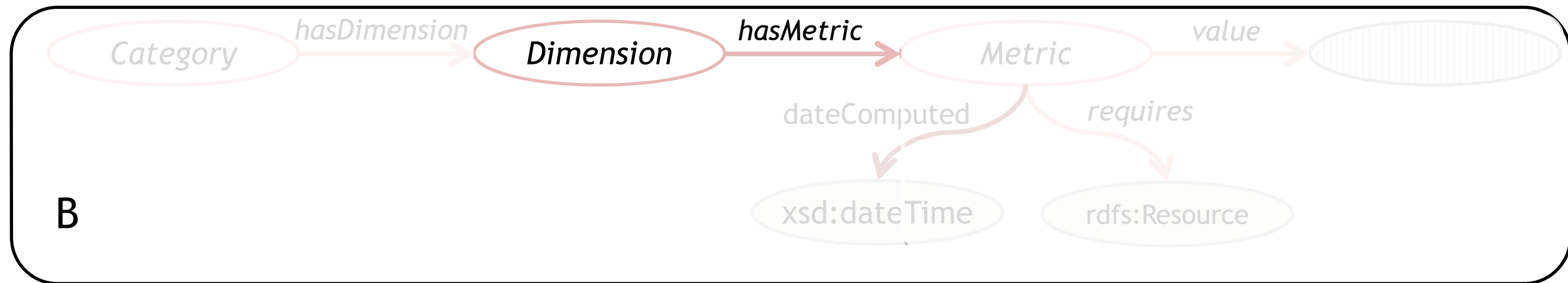
The daQ ontology is a generic *framework*, where classes and properties are defined in an *abstract manner*

Category



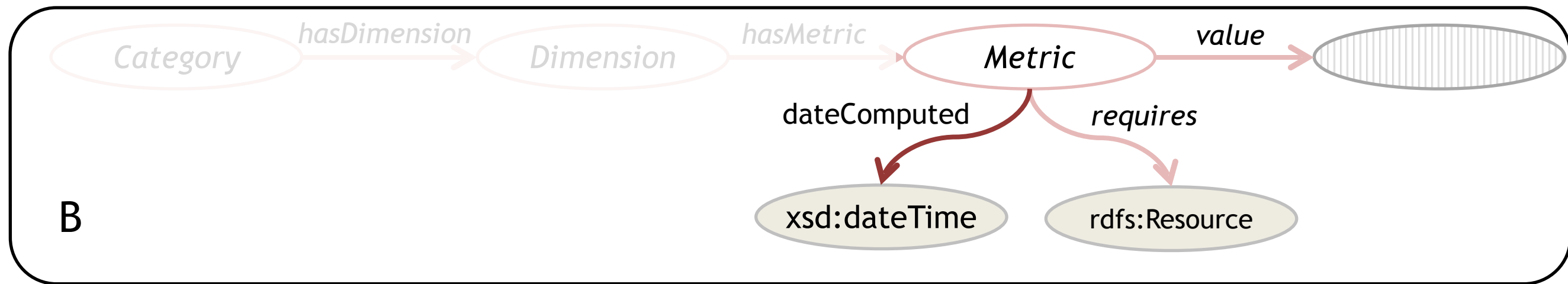
A category represent the highest level of quality assessment

Dimension



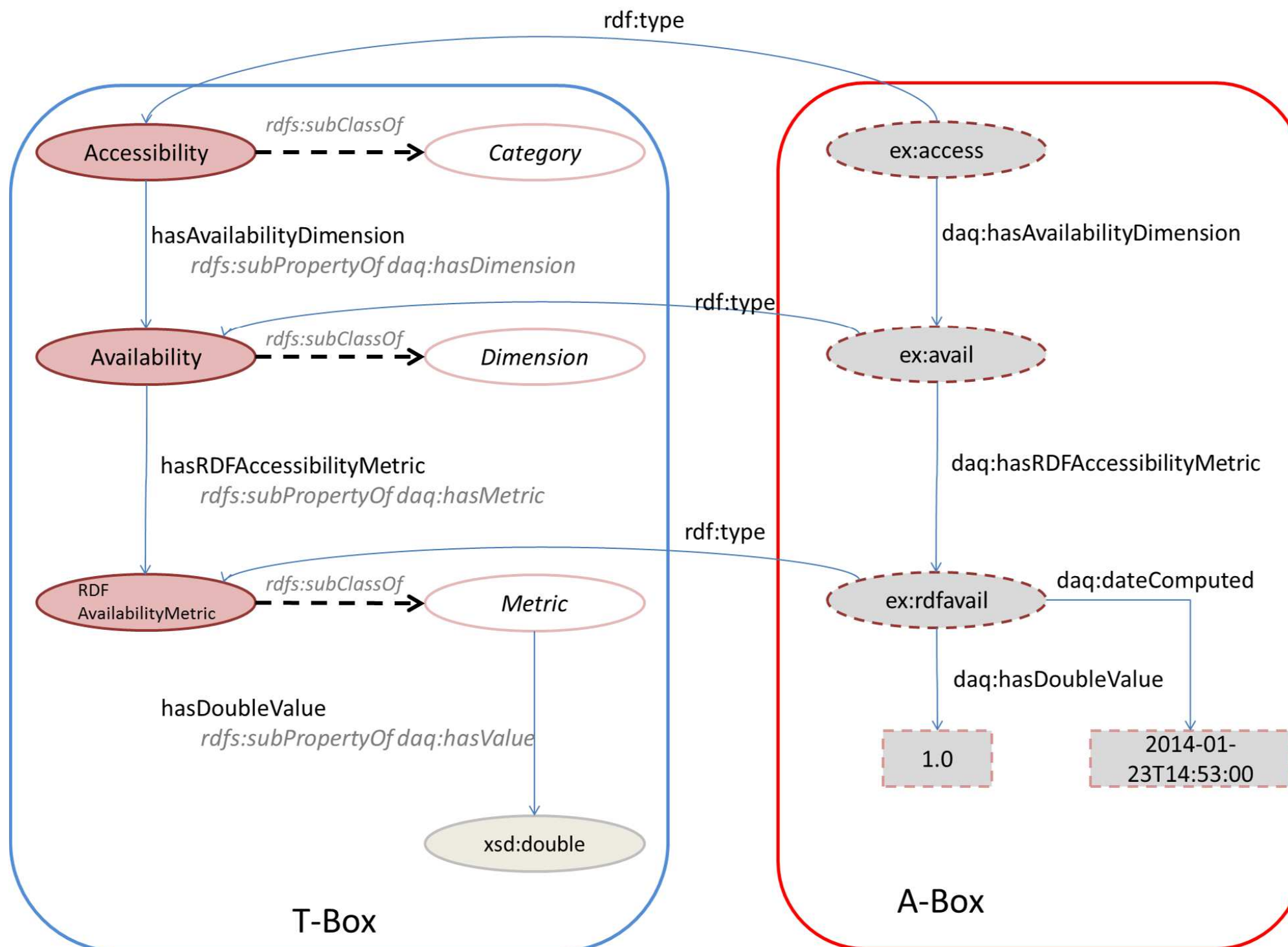
A dimension groups one or more metrics

Metric



The *smallest* unit of measuring a quality dimension

Using the daQ



Concluding Remarks

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