Building and Mining Knowledge Graphs

(KEN4256)

Lecture 5: Knowledge Graph Quality

Maastricht University
Institute of Data Science

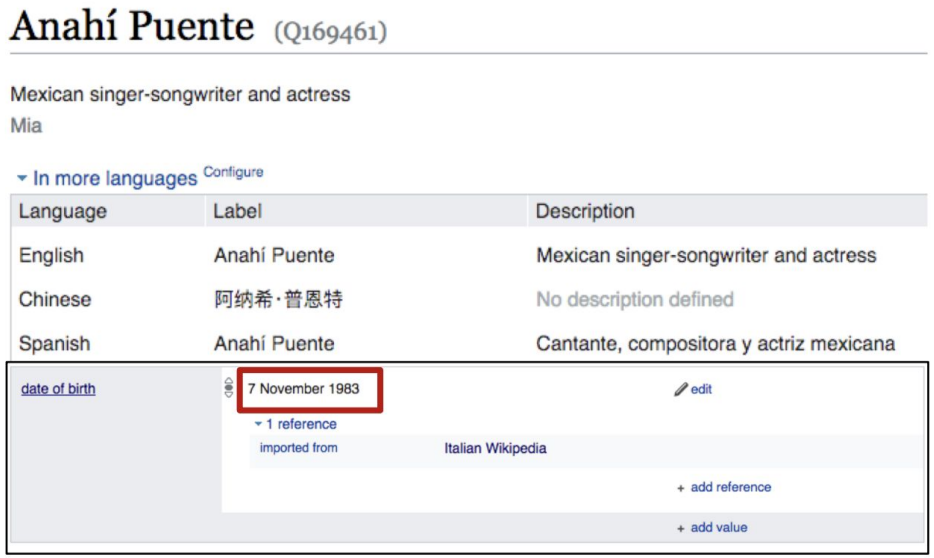
id: KEN4256_L5
version: 1.2024.0
created: February 17, 2019
last modified: March 26, 2024
published on: March 26, 2024

Quality issues in Data Integration

IMDB

Same entity?

WikiData



Which BirthDate is correct?

Data Quality Assessment & Goal

- **Goal:** assess and ultimately improve the quality of a KG
- **Process:** diagnose and fix data quality issues in a KG
  - **Root cause analysis**: identify source of quality issues
    - in the source data
    - arising due to data integration
    - incorrect use of vocabularies
    - linking data from untrustworthy sources

What is Data Quality?

Data Quality: a multi-dimensional concept with a popular definition ‘fitness for use’*.

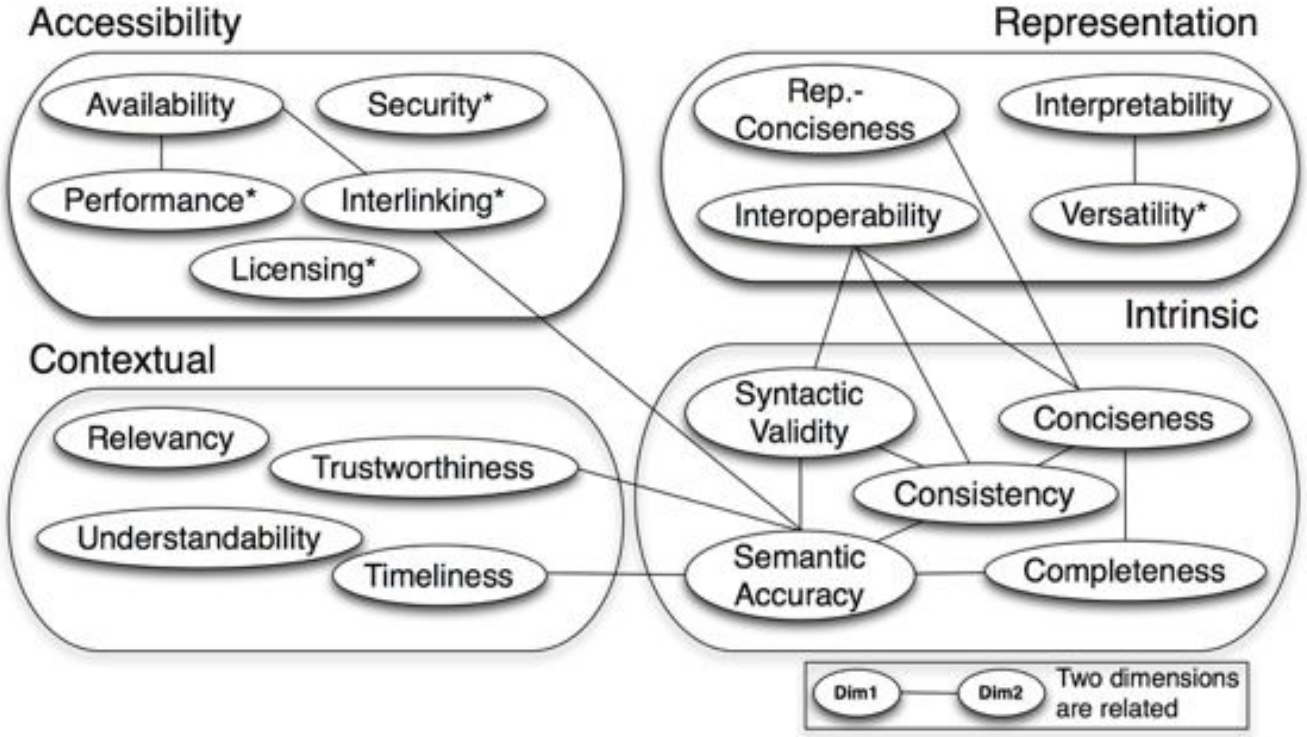
*Juran et al., The Quality Control Handbook, 1974

Dimensions and metrics of Data Quality

Dimension: characteristics of a dataset.

Metric: procedure for *measuring* a quality dimension.

Data Quality Dimensions for KGs



Quality assessment for linked data: A survey. *A Zaveri, A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93*

KG Quality Dimensions - Accessibility

Availability: *extent to which data is present, obtainable and ready for use*

Metrics:

- Check whether server responds to a SPARQL query
- Check whether an RDF dump is provided and can be downloaded
- Check whether the URI returns useful data (i.e. RDF)
- Check whether all URIs contained within data are derefenceable.

Availability: de-referenceability

proportion of URIs that return HTTP 200 and a valid HTML/RDF

Graph 1:

1. <http://examples.com/Vincent_van_Gogh> <http://dbpedia.org/ontology/birthPlace> <http://examples.com/Zundert> .
2. <http://examples.com/Vincent_van_Gogh> <http://examples.com/created> “Starry Night”^^xsd:string .

Graph 2:

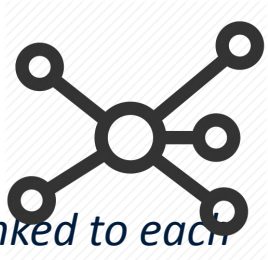
1. <https://www.wikidata.org/wiki/Q5582> <http://dbpedia.org/ontology/birthPlace> <https://www.wikidata.org/wiki/Q9883> .
2. <https://www.wikidata.org/wiki/Q5582> <http://dbpedia.org/ontology/created> <https://www.wikidata.org/wiki/Q45585> .

Which graph has higher de-referenceability?

Availability: accessibility of endpoints

https://yummydata.org (uptime)
https://lod-cloud.net/datasets

KG Quality Dimensions - Accessibility



Interlinking: *degree to which entities that represent the same concept are linked to each other, be it within or between two or more data sources or KGs*

Metrics:

- detection of the existence and use of external URIs (target dataset)
- detection of all local in-links or back-links: all triples from a dataset that have the resource's URI as the object

Interlinking: internal

metric*: degree - number of edges around a node*

Graph 1:

1. <http://mygraph.org/America> <http://dbpedia.org/property/population> “330,000,000”^^xsd:integer .
2. <http://mygraph.org/Florida> <http://dbpedia.org/ontology/isPartOf> <http://mygraph.org/USA> .

Graph 2:

1. <http://mygraph.org/America> <http://dbpedia.org/property/population> “330,000,000”^^xsd:integer .
2. <http://mygraph.org/Florida> <http://dbpedia.org/ontology/isPartOf> <http://mygraph.org/USA> .
3. <http://mygraph.org/America> <http://www.w3.org/2002/07/owl#sameAs> <http://mygraph.org/USA> .

Which graph has higher degree of interlinking?

Interlinking: external

metric: *number of sameAs chains*

Graph 1:

1. <http://mygraph.org/America> <http://dbpedia.org/property/population> “330,000,000”^^xsd:integer .
2. <http://mygraph.org/Florida> <http://dbpedia.org/ontology/isPartOf> <http://mygraph.org/USA> .

Graph 2:

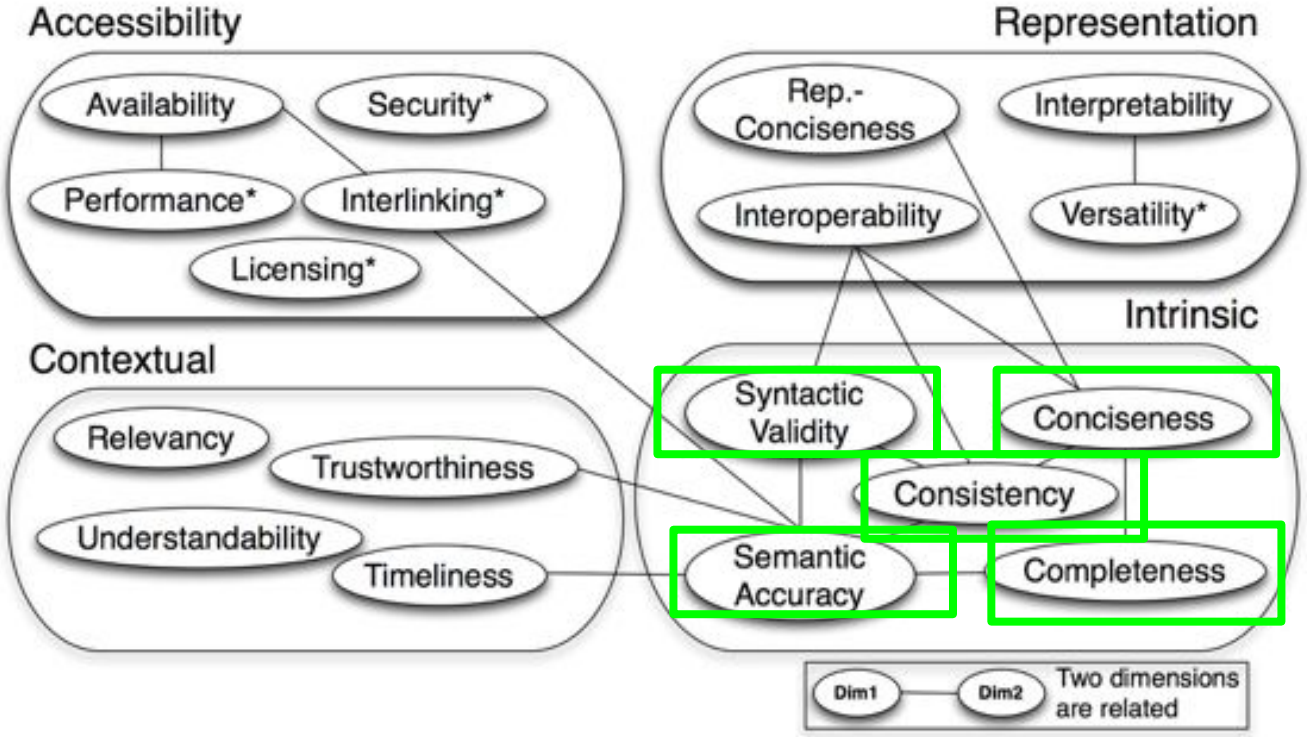
1. <http://mygraph.org/America> <http://dbpedia.org/property/population> “330,000,000”^^xsd:integer .
2. <http://mygraph.org/Florida> <http://dbpedia.org/ontology/isPartOf> <http://mygraph.org/USA> .
3. <http://mygraph.org/America> <http://www.w3.org/2002/07/owl#sameAs> <http://geonames.org/country/USA> .

External graph:

1. <http://geonames.org/country/USA> <http://geonames.org/property/area> “9, 834, 000”^^xsd:double .

Which graph has higher degree of interlinking?

Data Quality Dimensions for KGs



Quality assessment for linked data: A survey. *A Zaveri, A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93*

KG Quality Dimensions - Intrinsic

Syntactic Validity: *degree to which an RDF document conforms to the specification of the serialization format.*

Metrics:

(i) syntactic rules (type of characters allowed and/or the pattern of literal values)

(ii) use of explicit definition of the allowed values for a datatype

Process:

detecting syntax errors using (i) validators, (ii) via crowdsourcing

Example:
Literals (e.g. date) are tagged with appropriate data type.

KG Quality Dimensions - Intrinsic

Semantic Accuracy: *The degree to which data has attributes that correctly represent the real-life phenomena.*

Metrics:

- no incorrect values
- no misuse of properties
- no inaccurate annotations, labellings or classifications
- outliers

Example:
ex:John schema:age “-1”^^xsd:integer

KG Quality Dimensions - Intrinsic

Consistency: *The degree to which data is consistent with (has no violation) semantic rules defined.*

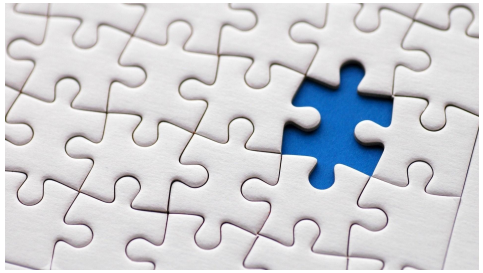
Metrics:

- correct domain and range definition
- no misplaced classes or properties

Example:

ex:John :drives ex:Tesla .
ex:Tesla rdf:type Person .

KG Quality Dimensions - Intrinsic



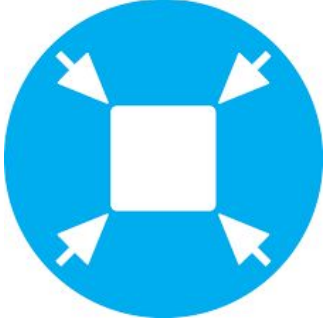
Completeness: *degree to which all required information is present in a particular dataset.*

Metrics:

- Schema - ontology completeness (do we have types for all entities?)
- Property - missing values for a specific property?
- Population - % of all real-world objects of a particular type
- Interlinking - degree to which instances in the dataset are interlinked

Example:
Does the KG have the GDP for all **countries,** and all **years**?

KG Quality Dimensions - Intrinsic



Conciseness: *degree to which the irrelevant and duplicate schema and data elements are avoided.*

Metrics:

- Intensional conciseness - refers to the case when the data set does not contain redundant schema elements (properties and classes). Only essential properties and classes are included in the schema;
- Extensional conciseness -refers to the case when the data set does not contain redundant objects (instances).

Example:
ex:NL rdfs:label “Netherlands”
ex:NL schema:name “Netherlands”

Which data quality issues do you see in the given RDF snippet?

@prefix ex: <http ://example.org/ontology/> .

ex:Italy ex:hasCapital ex:Milan.

ex:Italy ex:areaTotal ‘‘301338’’^^xsd:string .

ex:Italy rdfs:name “Italy”@en

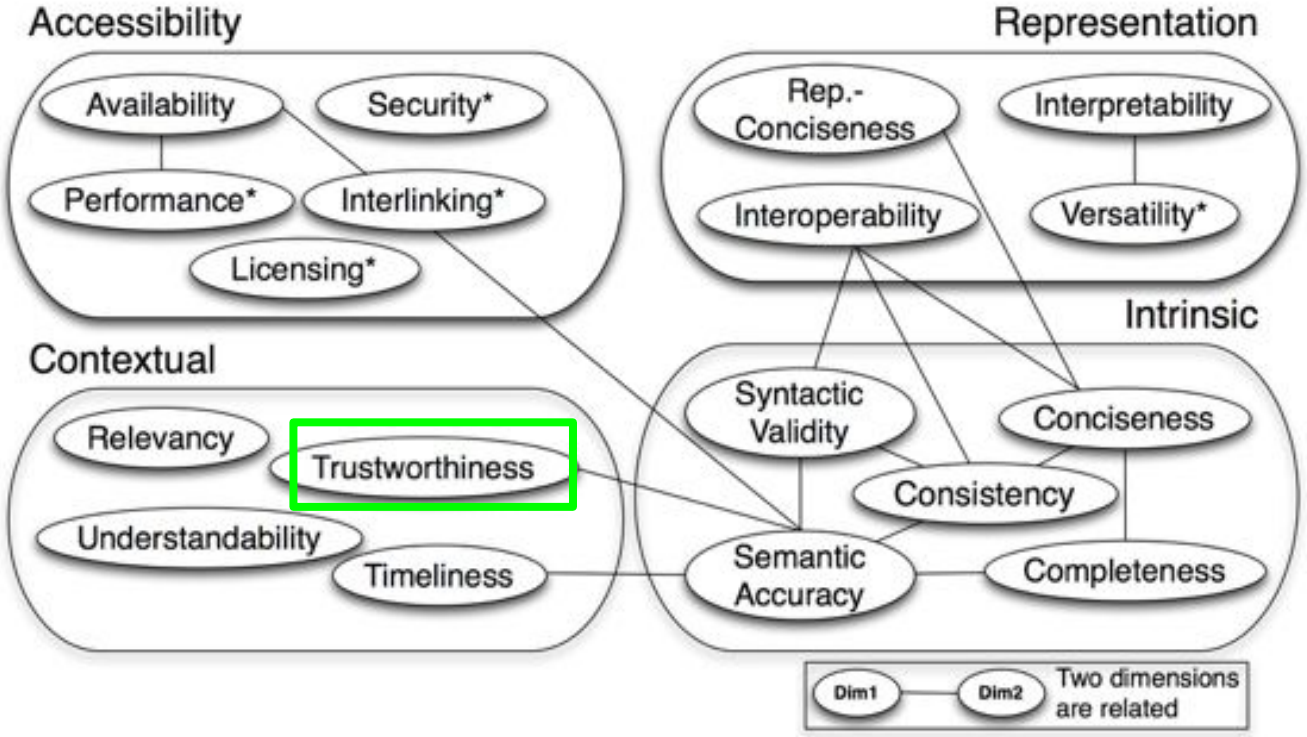
ex:Italy ex:hasCapital ex:Rome.

ex:Rome a :Place.

:hasCapital rdfs:domain :Country .

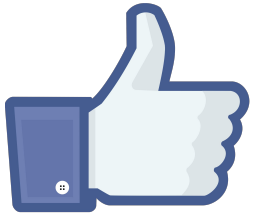
:hasCapital rdfs:range :City.

Data Quality Dimensions for KGs



Quality assessment for linked data: A survey. *A Zaveri, A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93*

KG Quality Dimensions - Contextual

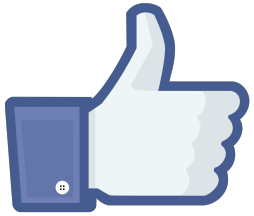


Trustworthiness: *degree to which the information is accepted to be correct, true, real and credible.*

Metrics:

- Does the KG contain triples that capture the **provenance** of each assertion (triples)? Who, when, where, how?
- Does it use provenance specifying schemas/ontologies (PROV-O, HCLS)
- Majority vote / opinion-based method: how many KG contributors have annotated this assertion / triple to state that they trust it?

KG Quality Dimensions - Contextual

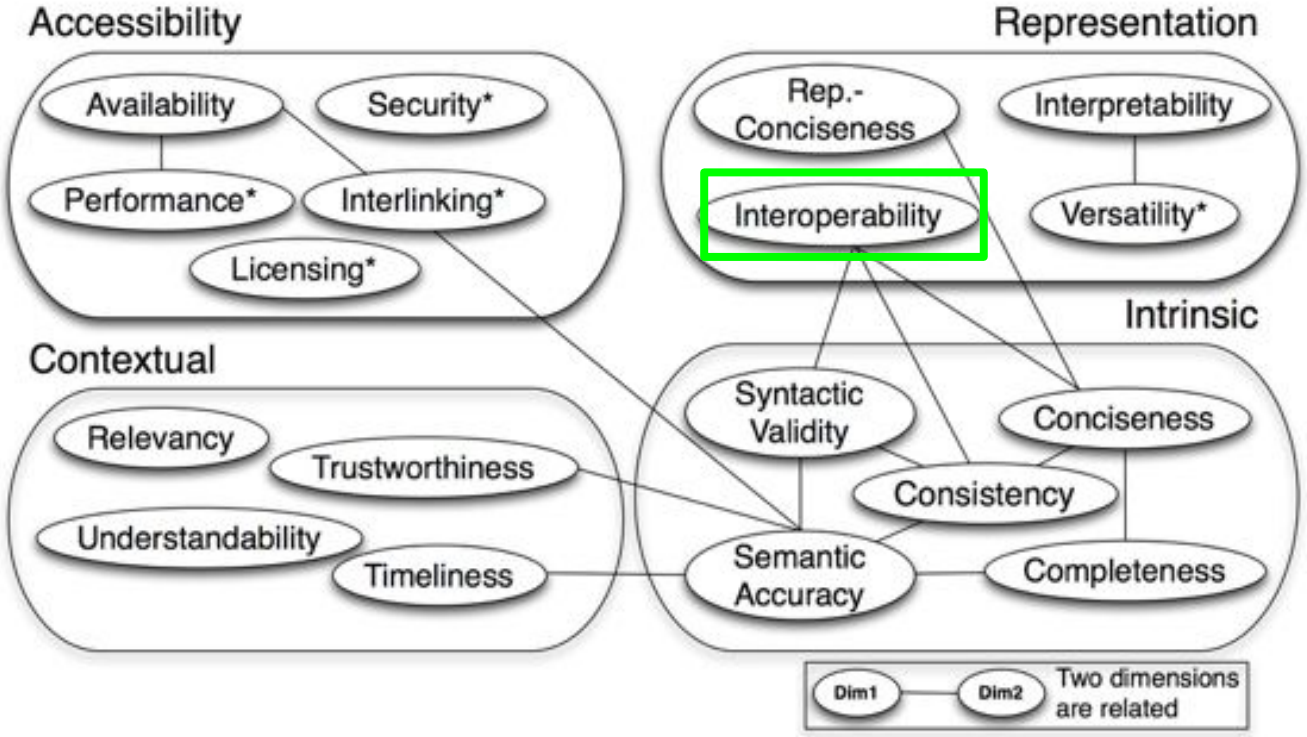


Trustworthiness: *degree to which the information is accepted to be correct, true, real and credible.*

```
:assertion {
   ex:trastuzumab ex:is-indicated-for ex:breast-cancer .
}

:provenance {
   :assertion prov:generatedAtTime "2012-02-03T14:38:00Z"^^xsd:dateTime  .
   :assertion prov:wasDerivedFrom :experiment .
   :assertion prov:wasAttributedTo :experimentScientist .
}
```

Data Quality Dimensions for KGs



Quality assessment for linked data: A survey. *A Zaveri, A Rula, A Maurino, R Pietrobon, J Lehmann, S Auer. Semantic Web 7 (1), 63-93*

KG Quality Dimensions - Representational

Interoperability - *degree to which the format and structure of the information conforms to previously returned information as well as data from other sources.*

Metrics:

- Reuse of well known vocabularies

Example:

ex:myKG ex:hasProvenance <https://d2s.semanticscience.org> .

vs

ex:myKG prov:wasGeneratedBy <https://d2s.semanticscience.org> .

Tools for KG Quality Assessment

SPARQL

SPARQL can be used to verify data quality issues.

RDF Graph

```
ex:ValidCountry a ex:Country ;
      ex:germanLabel "Spanien"@de .

ex:InvalidCountry a ex:Country ;
      ex:germanLabel "Spain"@en .
```

SPARQL

```
SELECT ?this (ex:germanLabel AS ?path) ?value
WHERE {
  ?this ex:germanLabel ?value .
  FILTER (!isLiteral(?value) || !langMatches(lang(?value), "de"))
}
```

SHACL

W3C Standard to validate RDF graphs.
SHACL offers a syntax to readily construct RDF constraints, and these can be implemented as SPARQL queries.

Some tools exist to construct and validate SHACL rules
https://shacl-play.sparna.fr/play/
https://shacl.org/playground/
https://forms.hypermedia.app/playground/
https://www.itb.ec.europa.eu/shacl/any/upload

SHACL conformance can be spotty, so check the test suite:

Example data graph

```
ex:Alice a ex:Person .
ex:Bob ex:address [ a ex:PostalAddress ; ex:city ex:Berlin ] .
ex:Carol ex:address [ ex:city ex:Cairo ] .
```

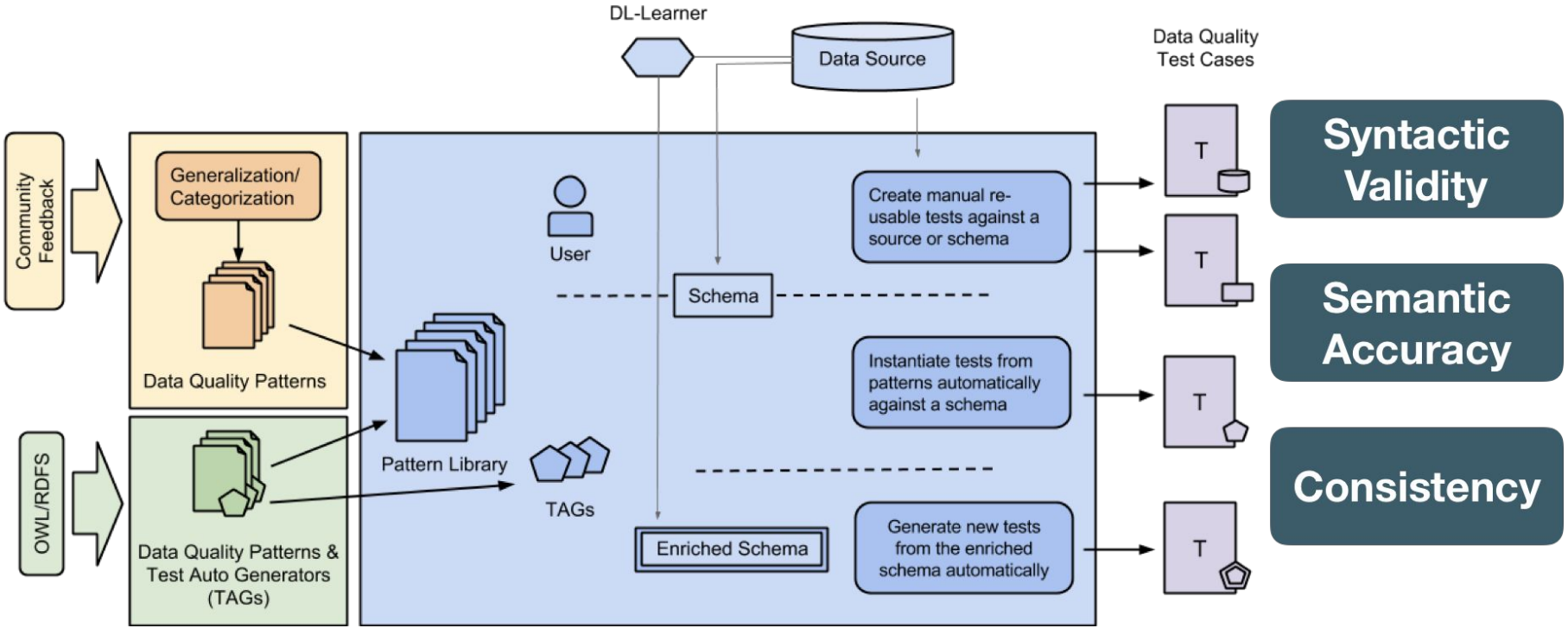
Example shapes graph

```
ex:ClassExampleShape
   a sh:NodeShape ;
   sh:targetNode ex:Bob, ex:Alice, ex:Carol ;
   sh:property [
      sh:path ex:address ;
      sh:class ex:PostalAddress ;
   ] .
```

POTENTIAL DEFINITION IN SPARQL (Must evaluate to true for each value node $value)

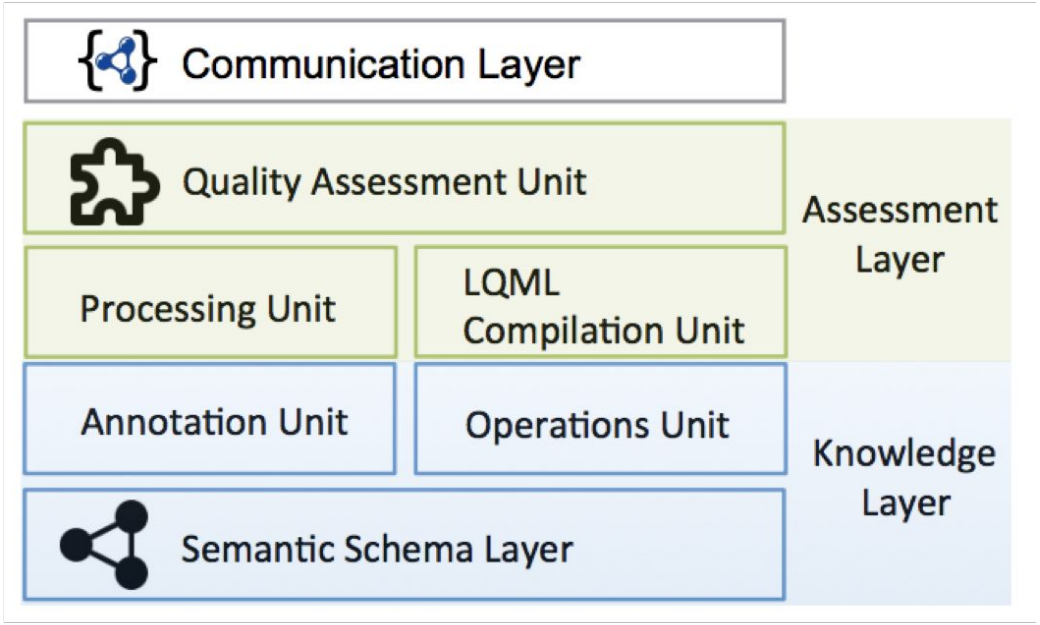
```
ASK {
   $value rdf:type/rdfs:subClassOf* $class .
}
```

DQA Tool - RDFUnit



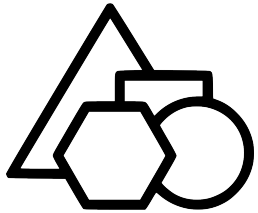
http://aksw.org/Projects/RDFUnit.html

DQA Tool - Luzzu



http://eis-bonn.github.io/Luzzu/index.html

DQA Tool - Shape Expressions



Shape Declaration

```
PREFIX :  <http://example.org/>
PREFIX schema: <http://schema.org/>
PREFIX xsd:
<http://www.w3.org/2001/XMLSchema#>

:User {
  schema:name         xsd:string  ;
  schema:birthDate    xsd:date?  ;
  schema:gender       [schema:Male
                      schema:Female ]
                      OR xsd:string ;
  schema:knows        IRI @:User*
}
```

Shape Validation

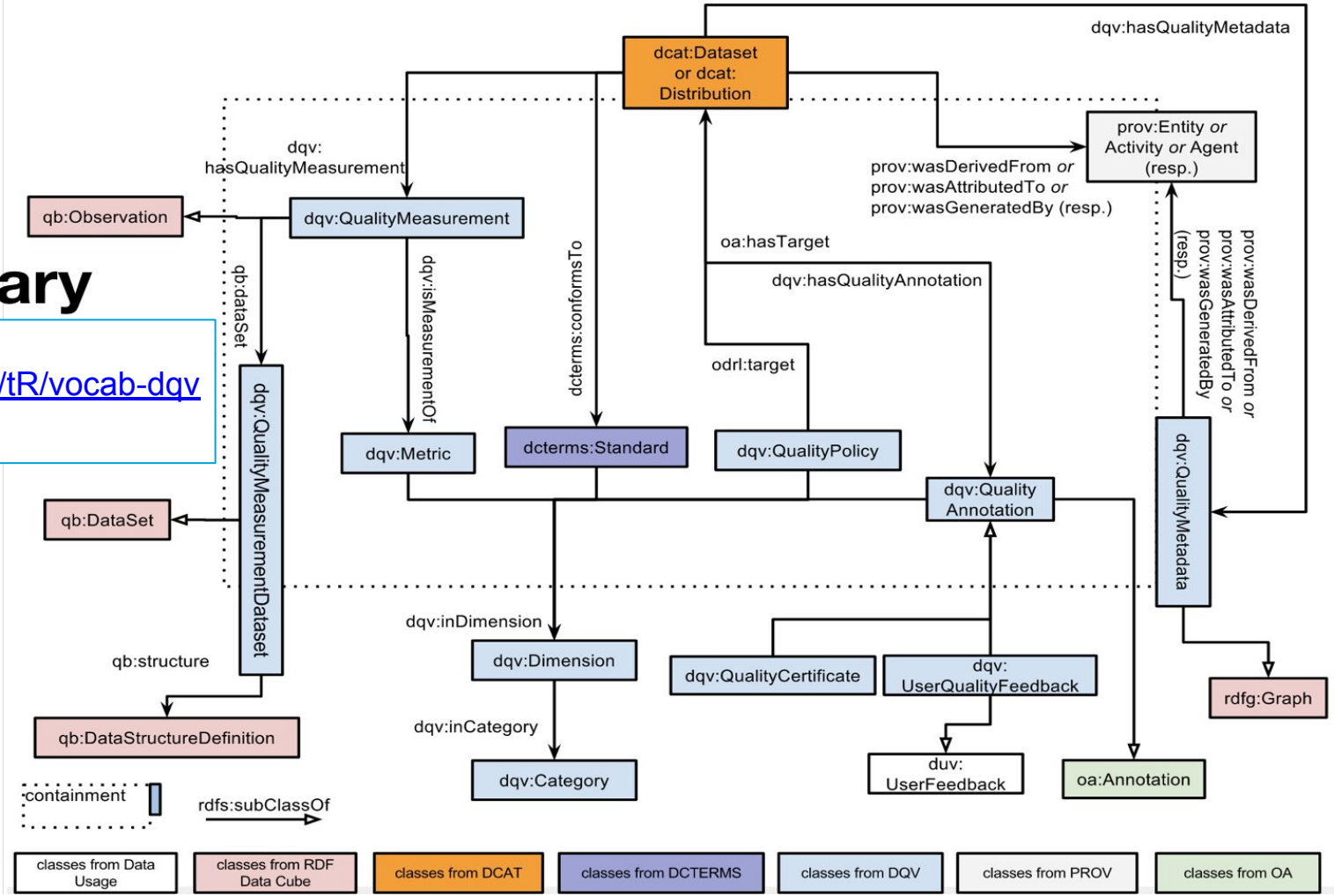
```
alice schema:name        "Alice" ;          #  Passes as a :User
      schema:gender      schema:Female ;
      schema:knows       :bob .

:bob  schema:gender    schema:Male ;     #  Passes as a :User
      schema:name        "Robert";
      schema:birthDate    "1980-03-10"^^xsd:date .


:dave  schema:name         "Dave";       #  Fails as a :User
       schema:gender       "XYY";
       schema:birthDate    1980 .        #  1980 is not an xsd:date *)
```

W3C Data Quality Vocabulary

https://www.w3.org/tR/vocab-dqv



Summary

- Data (KG) quality has important **social and financial consequences** because the data in KGs drive applications for decision-making.
- Data (KG) quality is a **vast research area** studying various dimensions and aspects of data quality.
- Each dimension of quality **can be assessed and measured** using standard or KG engineer-specified metrics.
- **Software tools can assist** in automated assessment of data quality. However, it can also be performed semi-automatically with **machines in combination with humans** (e.g. crowdsourcing)
- Data quality assessment **can improve the quality of your KGs**, which in turn **improves the trustworthiness of any analyses and systems** built on top of these KGs.