Digitaliseringsdirektoratet
Norwegian Digitalisation Agency

Kartverket

Data quality in an e-Government perspective

3rd International Workshop on Spatial Data Quality, 2020-01-28~29, Malta

Jim J. Yang & Anne Karete Hvidsten, *Norwegian Digitalisation Agency*
Morten Borrebæk, *Norwegian Mapping Authority*

About us

- Norwegian Digitalisation Agency
  - The Norwegian Digitalisation Agency is the Norwegian government’s foremost tool for faster and more coordinated digitalization of the Norwegian public sector.
  - A role as rule setter and supplier, responsible for, including: national common IT solutions and building blocks, national interoperability framework and standards.
- Norwegian Mapping Authority
  - The Norwegian Mapping Authority collates, systemises, manages and communicates public geographical information.
  - Responsible for, including: National Geodetic Frame, positioning services, digital maps, Land registry, Property information, Place names, PRIMAR ENC Service and standards.

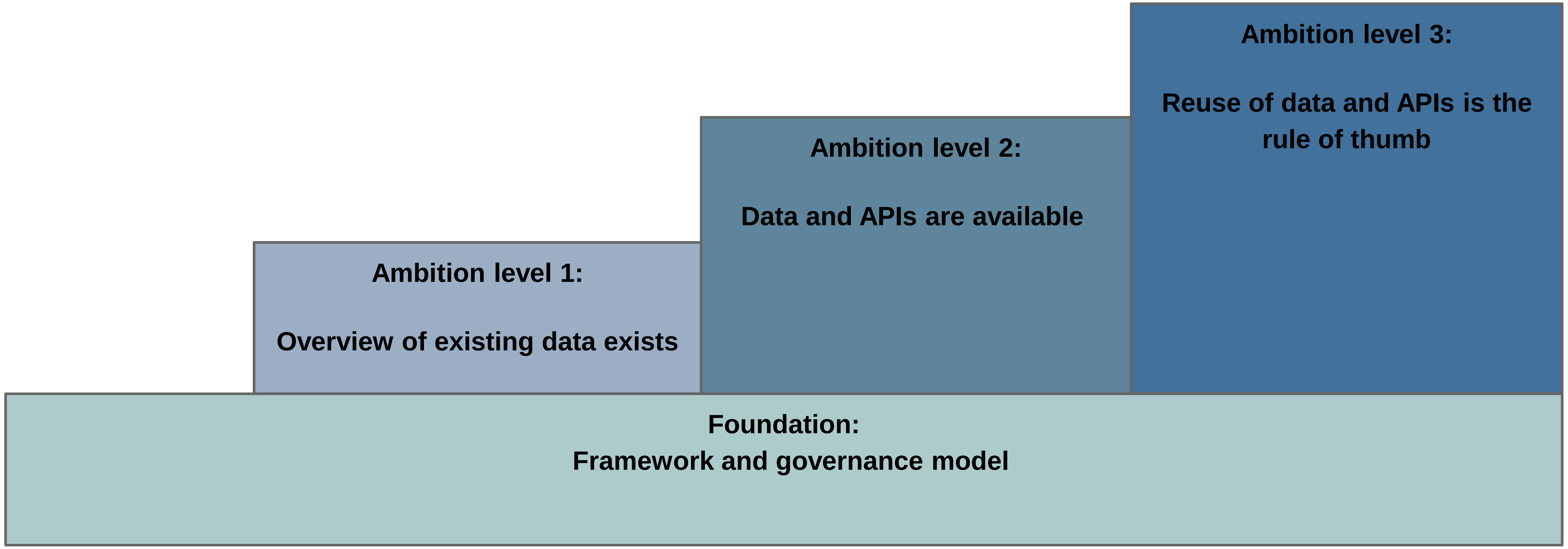
Outline

- e-Government and data sharing and reuse
- Machine-readable data quality descriptions
- Common definitions of data quality metrics etc.
- Mapping to ISO 19157
- Summary and future work

Data sharing and reuse, e-Gov

The ambition

- Digital Agenda for Norway
- Digitalization strategy
- The digitalization circular



Ambition level 3:

Reuse of data and APIs is the rule of thumb

Ambition level 2:

Data and APIs are available

Ambition level 1:

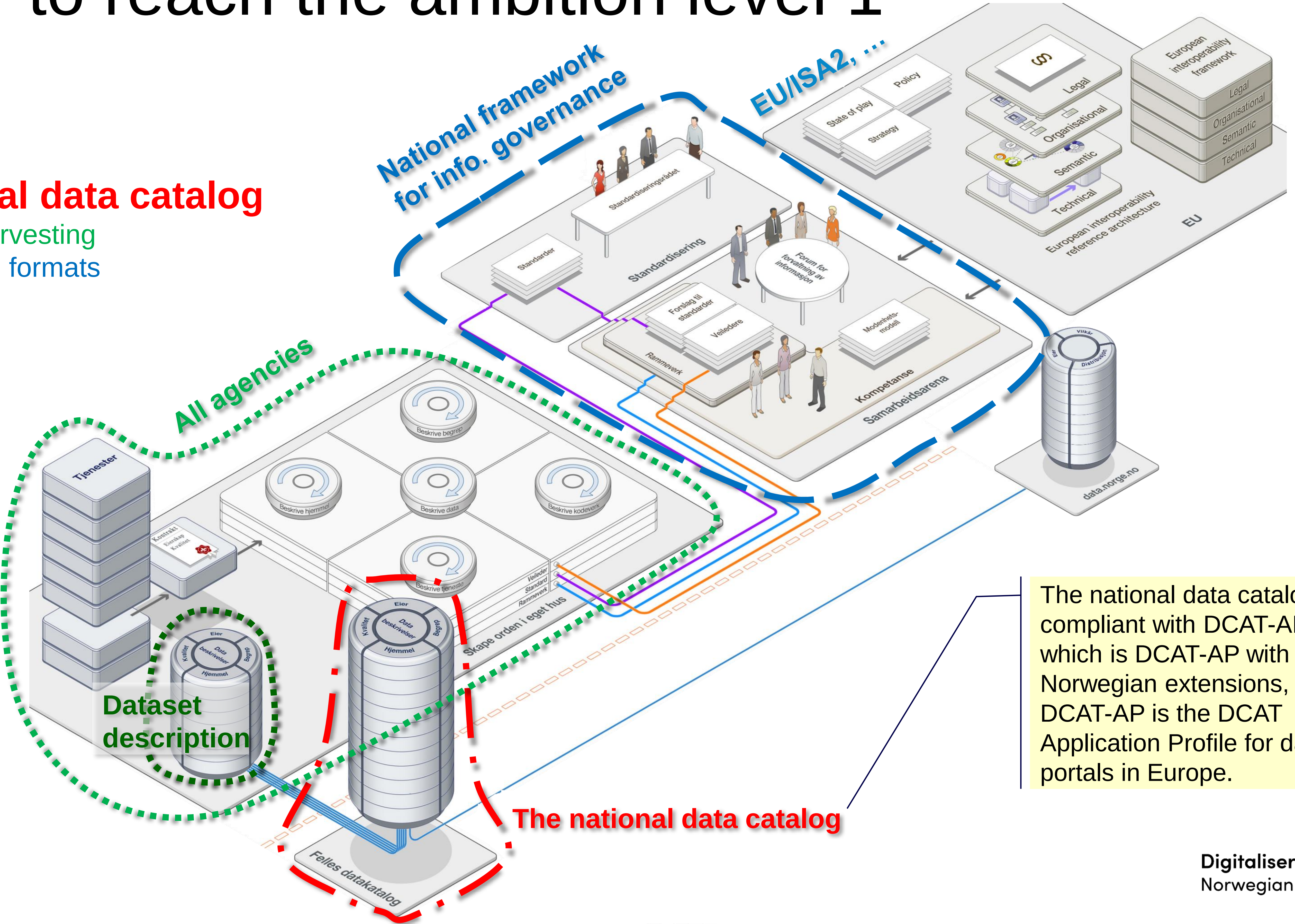
Overview of existing data exists

Foundation:
Framework and governance model

In order to reach the ambition level 1

The national data catalog

- Automatic harvesting
- Standardized formats



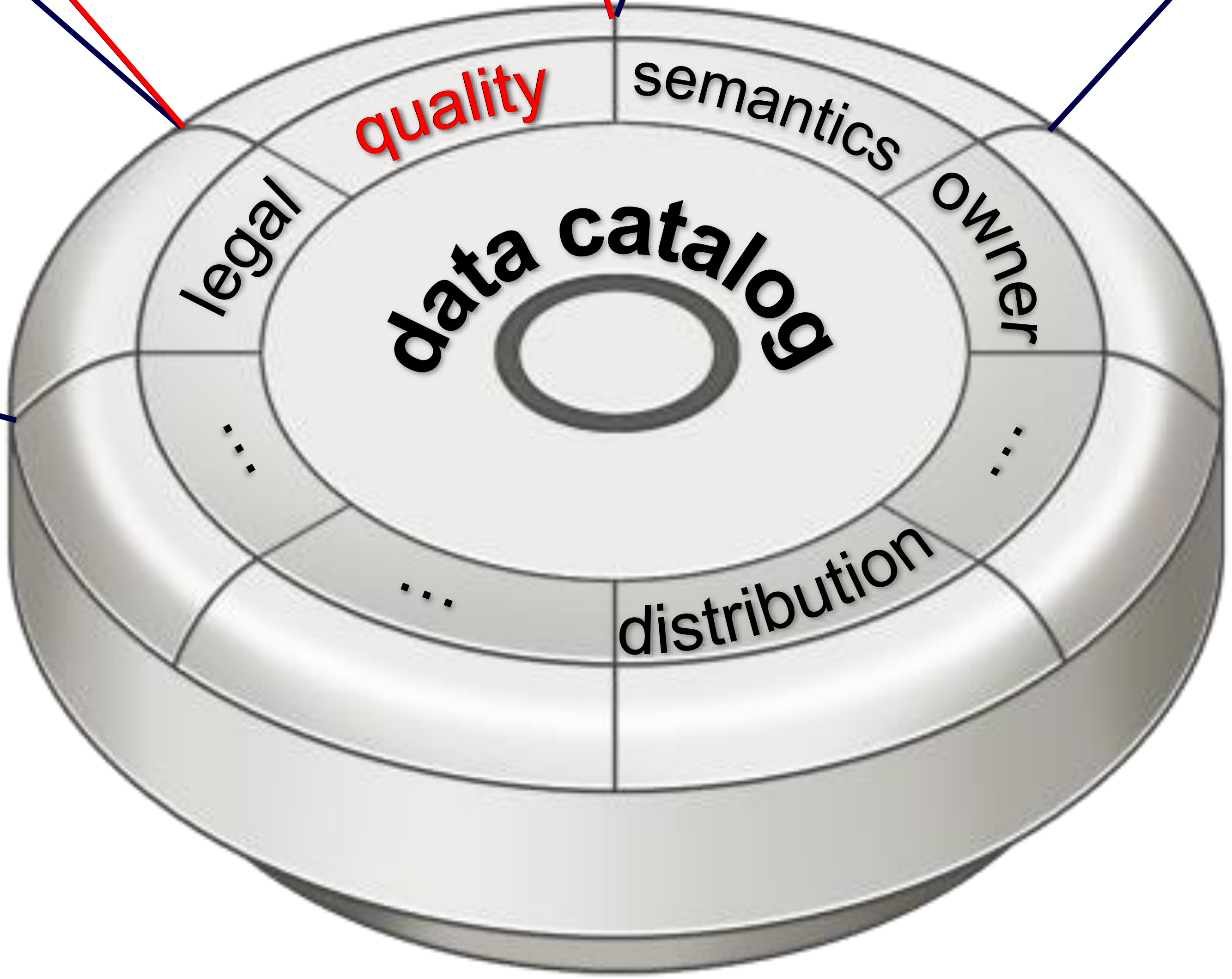
The national data catalog is compliant with DCAT-AP-NO which is DCAT-AP with Norwegian extensions, and DCAT-AP is the DCAT Application Profile for data portals in Europe.

In order to evaluate if a dataset is reusable

Descriptions of
- purpose of processing
- legal basis for processing
- legal basis for non-disclosure
- legal basis for disclosure
- …

Descriptions of
- accuracy
- completeness
- consistency
- …

Descriptions of
- data elements
- concepts
- reference data
- data models
- …



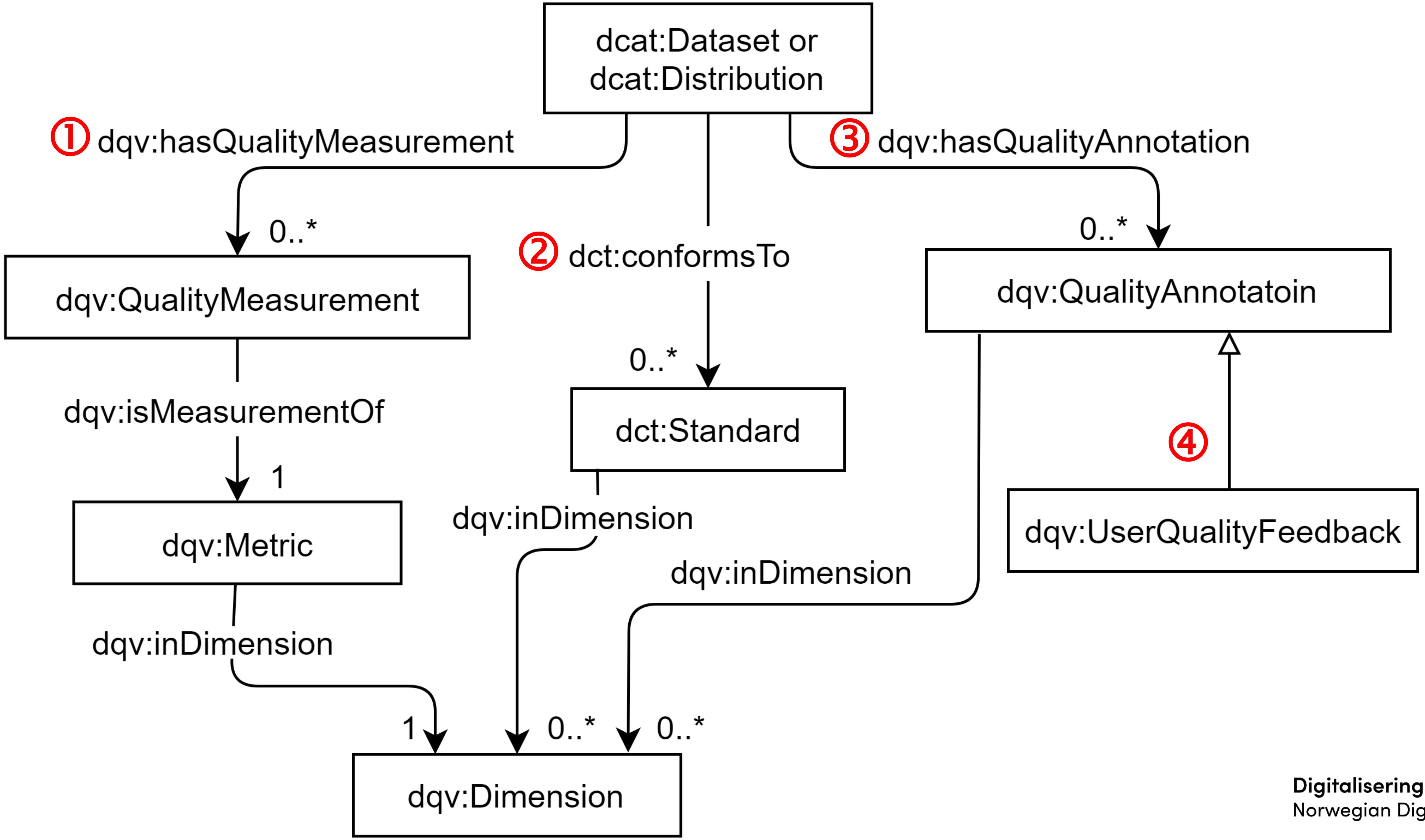
! Challenge #1:
DCAT-AP that DCAT-AP-NO is based on, does not cover all the aspects

A working group, for coping with challenge #1

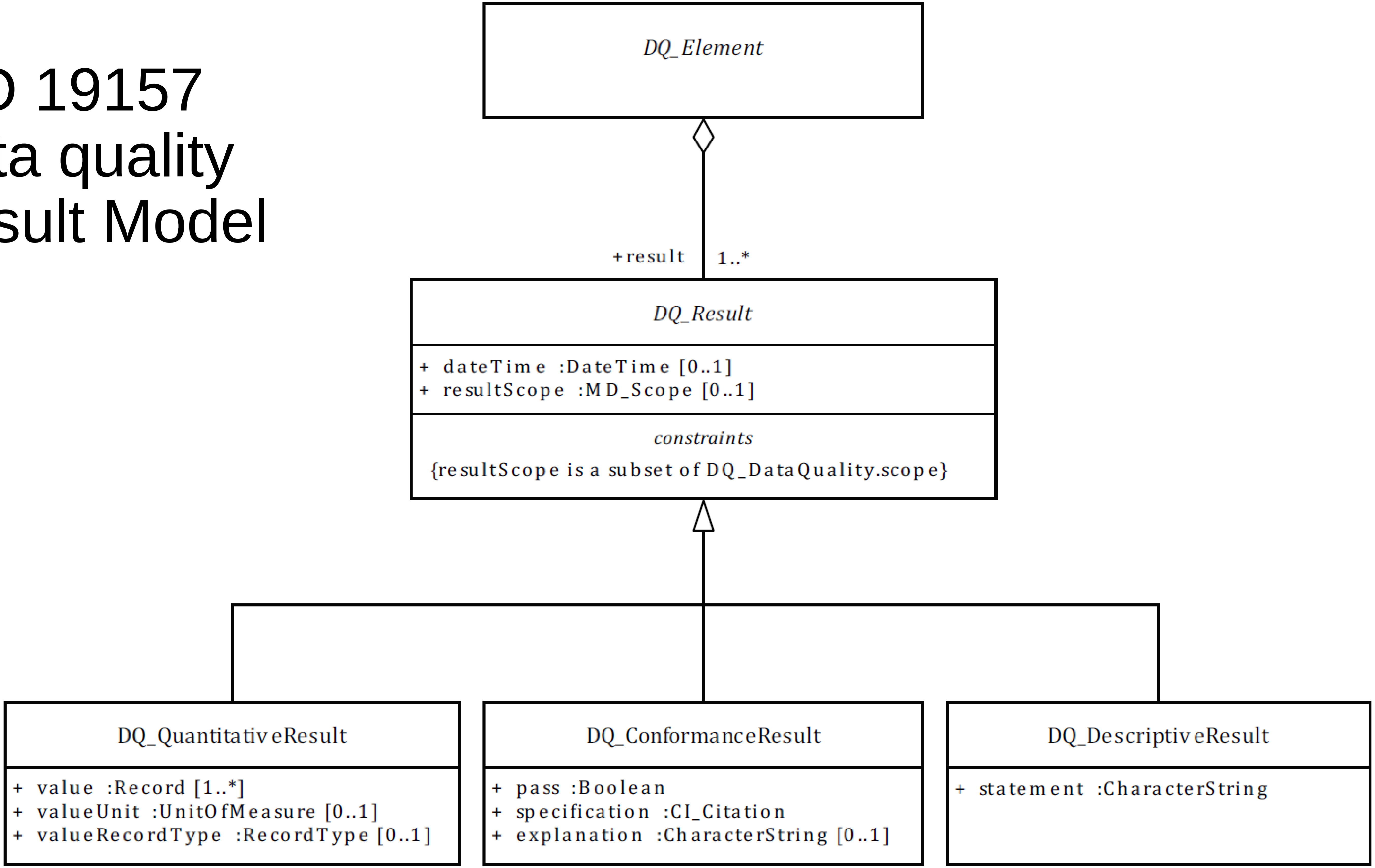
- to establish standards/specifications for machine-readable description of quality of datasets
- suggested to
  - extend DCAT-AP-NO with W3C/DQV (Data Quality Vocabulary)
  - start with:
    1. description of quantitative data quality
    2. description of data quality that conforms to given standards/specifications
    3. description of data quality in plain text
    4. user feedback on data quality, in plain text

Standardized machine-readable DQ descriptions

Using DQV (Data Quality Vocabulary, https://www.w3.org/TR/vocab-dqv/)



ISO 19157
Data quality
Result Model

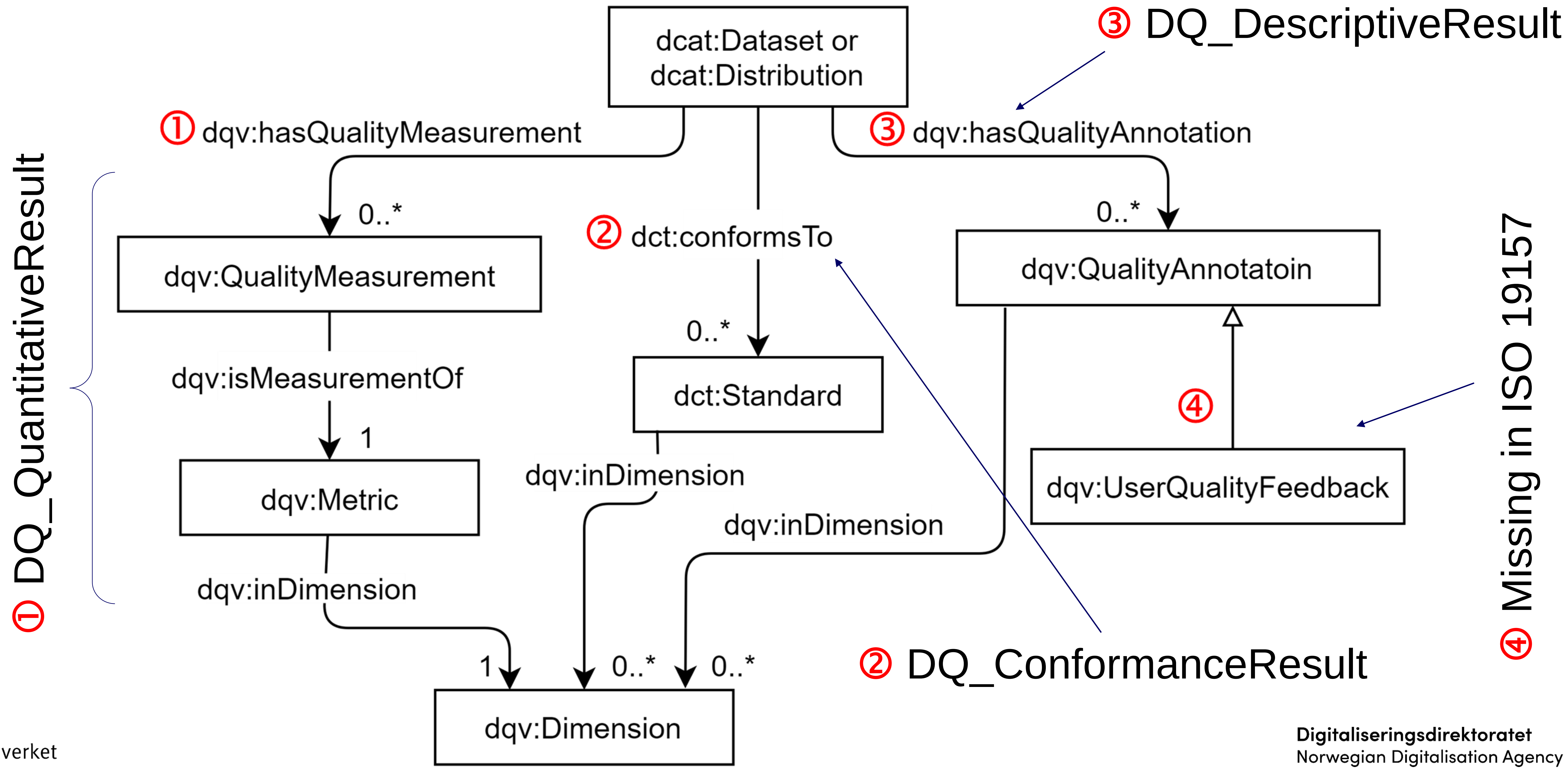


DQ_Element
+result
1..*
DQ_Result
+ dateTime :DateTime [0..1]
+ resultScope :MD_Scope [0..1]
constraints
{resultScope is a subset of DQ_DataQuality.scope}
DQ_QuantitativeResult
+ value :Record [1..*]
+ valueUnit :UnitOfMeasure [0..1]
+ valueRecordType :RecordType [0..1]
DQ_ConformanceResult
+ pass :Boolean
+ specification :CI_Citation
+ explanation :CharacterString [0..1]
DQ_DescriptiveResult
+ statement :CharacterString

Kartverket

Digitaliseringsdirektoratet
Norwegian Digitalisation Agency

Describing ISO 19157 DQ_Result using DQV



① DQ_QuantitativeResult

dcat:Dataset or
dcat:Distribution

① dqv:hasQualityMeasurement

0..*

dqv:QualityMeasurement

dqv:isMeasurementOf

1

dqv:Metric

dqv:inDimension

1

② dct:conformsTo

0..*

dct:Standard

dqv:inDimension

0..*

③ DQ_DescriptiveResult

③ dqv:hasQualityAnnotation

0..*

dqv:QualityAnnotatoin

④

dqv:UserQualityFeedback

dqv:inDimension

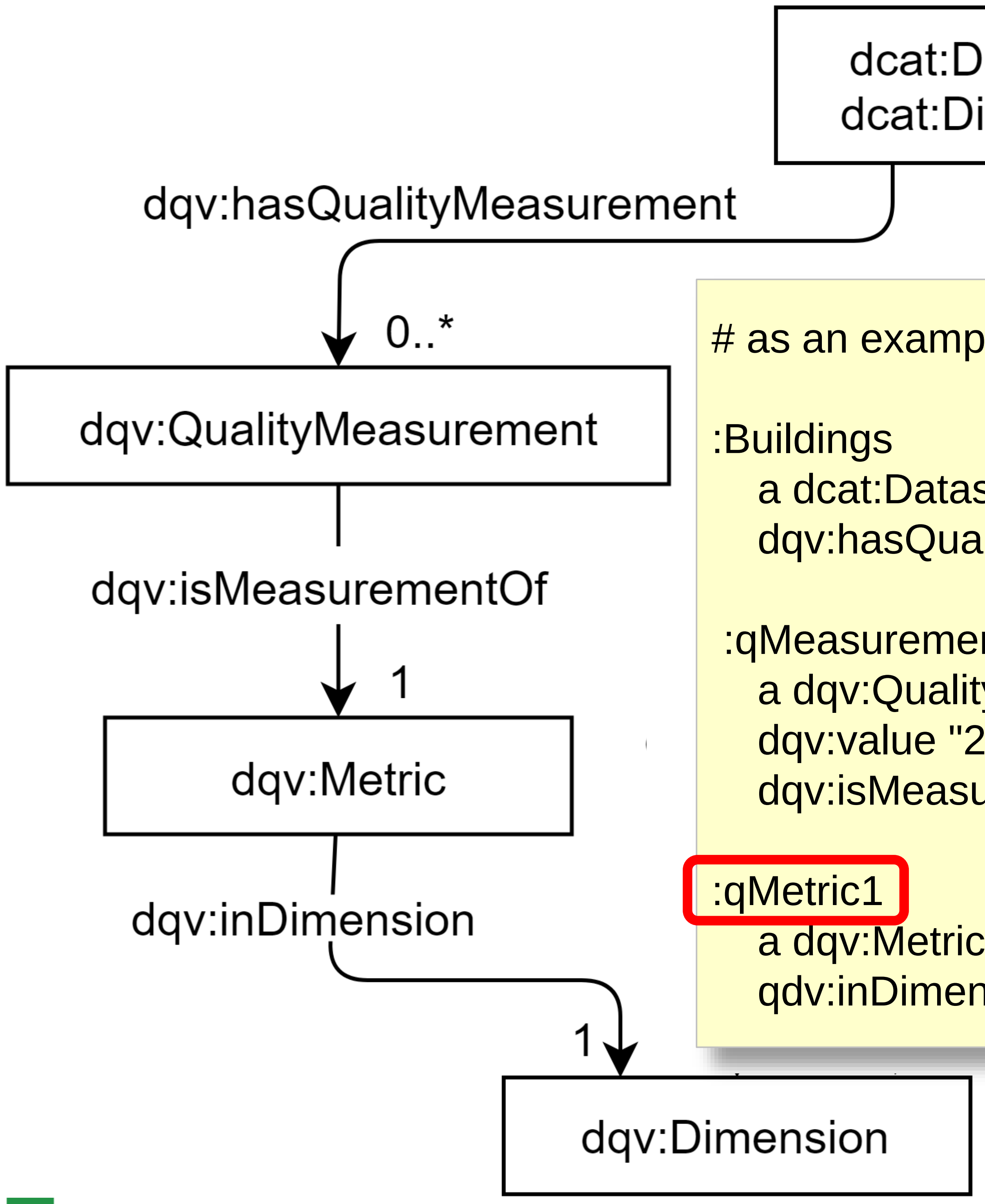
0..*

dqv:Dimension

② DQ_ConformanceResult

④ Missing in ISO 19157

An example – machine-readable description



```
# as an example (in RDF)

:Buildings
    a dcat:Dataset ;
    dqv:hasQualityMeasurement : qMeasurement1 .

 :qMeasurement1
    a dqv:QualityMeasurement ;
    dqv:value "2"^^xsd:integer ;
    dqv:isMeasurementOf :Metric1 .

:qMetric1
    a dqv:Metric ;
    qdv:inDimension :qDimension1 .
```

?

In plain English:

"Buildings" is a dataset (as defined in DCAT); it has a quality measurement called "qMeasurement1".

"qMeasurement1" is a quality measurement (as defined in DQV); it has an integer value "2"; it is a measurement of "qMetric1".

"qMetric1" is a metric (as defined i DQV); it is in a dimension called "qDimension1" .

A better approach – predefined metrics etc.

```
# yet another better example (in RDF)

:Buildings
   a dcat:Dataset ;
   dqv:hasQualityMeasurement :qMeasurement1 .

 :qMeasurement1
   a dqv:QualityMeasurement ;
   dqv:value “2”^^xsd:integer ;
   dqv:isMeasurementOf dqvno:NumberOfMissingObjects .
```

```
dqvno:NumberOfMissingObjects
      a dqv:Metric ;
      skos:definition "number of missing objects in relation to the
      number of objects that should be present in the dataset"@en ;
      dqv:expectedDataType xsd:integer ;
      dqv:inDimension iso:completeness .

iso:completeness
      a dqv:dimension ;
      skos:definition "the degree to which ...”@en ;
      dct:source “ISO 25012:2008 Software engineering ...”@en .
```

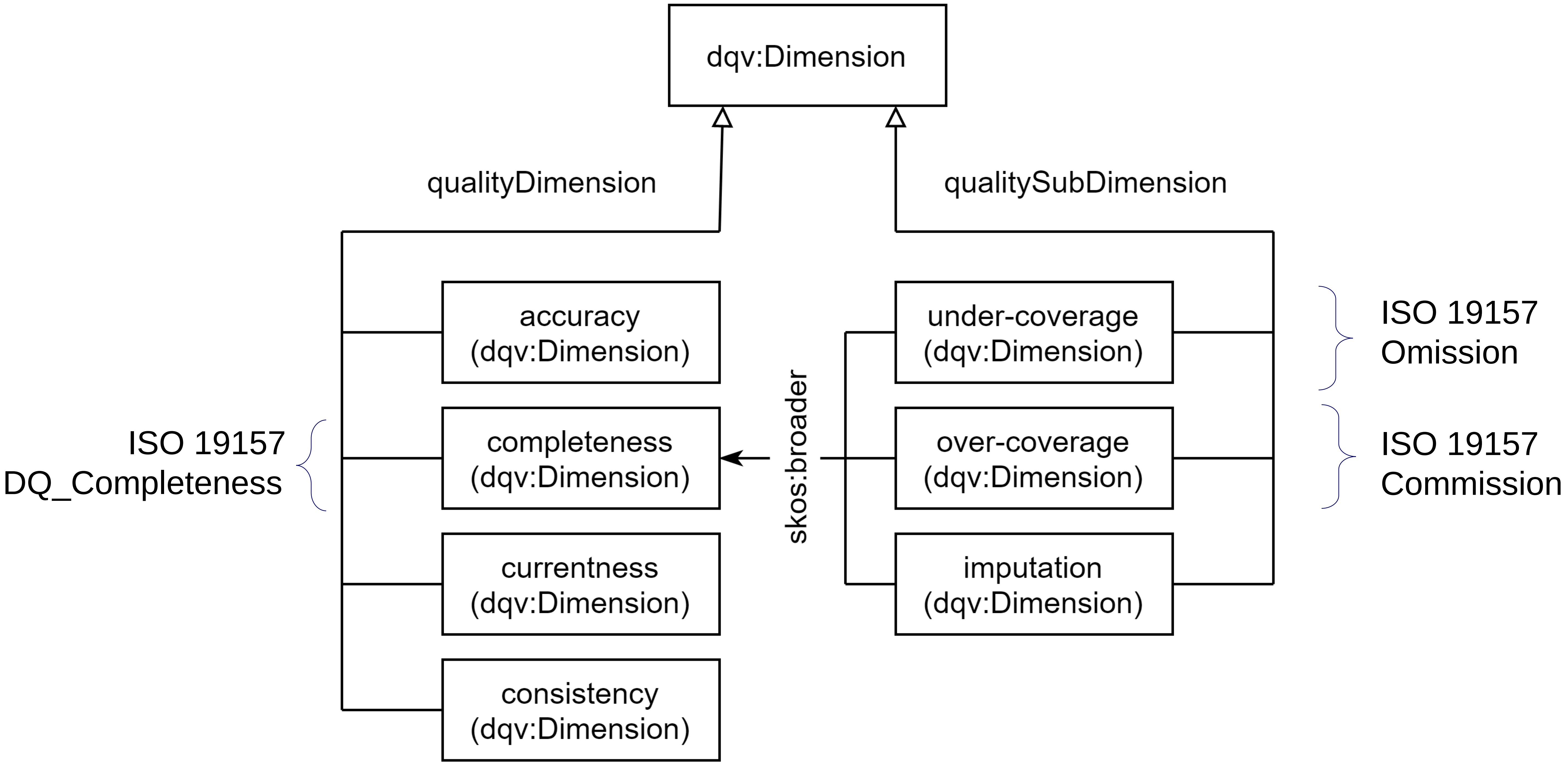
! Challenge #2:
Which pre-definitions?

pre-defined, as a controlled vocabulary

A working group, for coping with challenge #2

- to establish a set of common definitions
  - definitions of quality metrics, i.e., quantitative quality descriptions
  - quality metrics for datasets that are (to be) made available (i.e., not for the production phase, e.g. not "punctuality")
  - quality metrics not already defined in other standardized vocabularies (e.g. not "data updating frequency")
  - not sector/domain specific
  - only inherent data quality metrics (i.e., not system dependent, e.g. not "accessibility")
- also had to define relevant quality dimensions etc. (since metrics should be related to quality dimensions)

Describing ISO 19157 DQ_Element using dqv:Dimension



dqv:Dimension
qualityDimension
qualitySubDimension
accuracy (dqv:Dimension)
completeness (dqv:Dimension)
currentness (dqv:Dimension)
consistency (dqv:Dimension)
skos:broader
under-coverage (dqv:Dimension)
over-coverage (dqv:Dimension)
imputation (dqv:Dimension)
ISO 19157 DQ_Completeness
ISO 19157 Omission
ISO 19157 Commission

SKOS: Simple Knowledge Organization System

Kartverket

Digitaliseringsdirektoratet
Norwegian Digitalisation Agency

Pre-defined metrics, subdimensions and dimensions

Quliaty dimension	Quality subdimension	Quality metric (with data type)	Example
completeness	under-coverage	missing objects (boolean)	"false" (all buildings present)
		number of missing objects (integer)	"2" (two buldings missing)
		rate of missing objects (percentage)	"0.02%" (0.02% of buildings missing)
		number of objects with missing value for a given property (integer)	"2" (two buldings with missing values for "usable area")
		rate of objects with missing value for a given property (percentage)	"0.02%" (0.02% of buldings with missing values for "usable area")
	over-coverage	excess objects (boolean)	"true" (some excess buildings)
		number of excess objects (integer)	"2" (two excess buildings)
		rate of excess objects (percentage)	"0.02%" (0.02% excess buildings)
	imputation	number of objects with imputed value for a given property (integer)	"2" (two buildings with imputed values for "year of construction")
		rate of objects with imputed value for a given property (percentage)	"0.02%" (two buildings with imputed values for "year of construction")

Please conf. the published paper for the definitions

Pre-defined ... (cont.)

Quliaty dimension	Quality subdimension	Quality metric (with data type)
currentness	delay	overall time difference (xsd:duration)
consistency	consistency within the dataset	rate of objects with inconsistent properties (percentage)
		rate of objects with inconsistency between given properties (percentage)
accuracy	identifier correctness	number of objects with incorrect identifiers (integer)
		rate of objects with incorrect identifiers (percentage)
	classification correctness	number of incorrectly classified objects for a given property (integer)
		rate of incorrectly classified objects for a given property (percentage)

Please conf. the published paper for the definitions

Mapping to ISO-standards

Quliaty dimension	Quality subdimension	Quality metric (with data type)	
completeness Definition from ISO 25012:2008	under-coverage Definition from ISO 19157:2013 "omission"	missing objects (boolean)	Definitions based on ISO 19157:2013
		number of missing objects (integer)	
		rate of missing objects (percentage)	
		number of objects with missing value for a given property (integer)	
		rate of objects with missing value for a given property (percentage)	
	over-coverage Definition from ISO 19157:2013 "commission"	excess objects (boolean)	Definitions based on ISO 19157:2013
		number of excess objects (integer)	
		rate of excess objects (percentage)	
	imputation	number of objects with imputed value for a given property (integer)	
		rate of objects with imputed value for a given property (percentage)	

ISO 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model

ISO 19157:2013 Geographic information — Data quality

Mapping to ISO-standards (cont.)

Quality dimension	Quality subdimension	Quality metric (with data type)
currentness	delay	overall time difference (xsd:duration)
consistency	consistency within the dataset	rate of objects with inconsistent properties (percentage)
		rate of objects with inconsistency between given properties (percentage)
accuracy	identifier correctness	number of objects with incorrect identifiers (integer)
		rate of objects with incorrect identifiers (percentage)
	classification correctness	number of incorrectly classified objects for a given property (integer)
		rate of incorrectly classified objects for a given property (percentage)

Definitions from ISO 25012: 2008

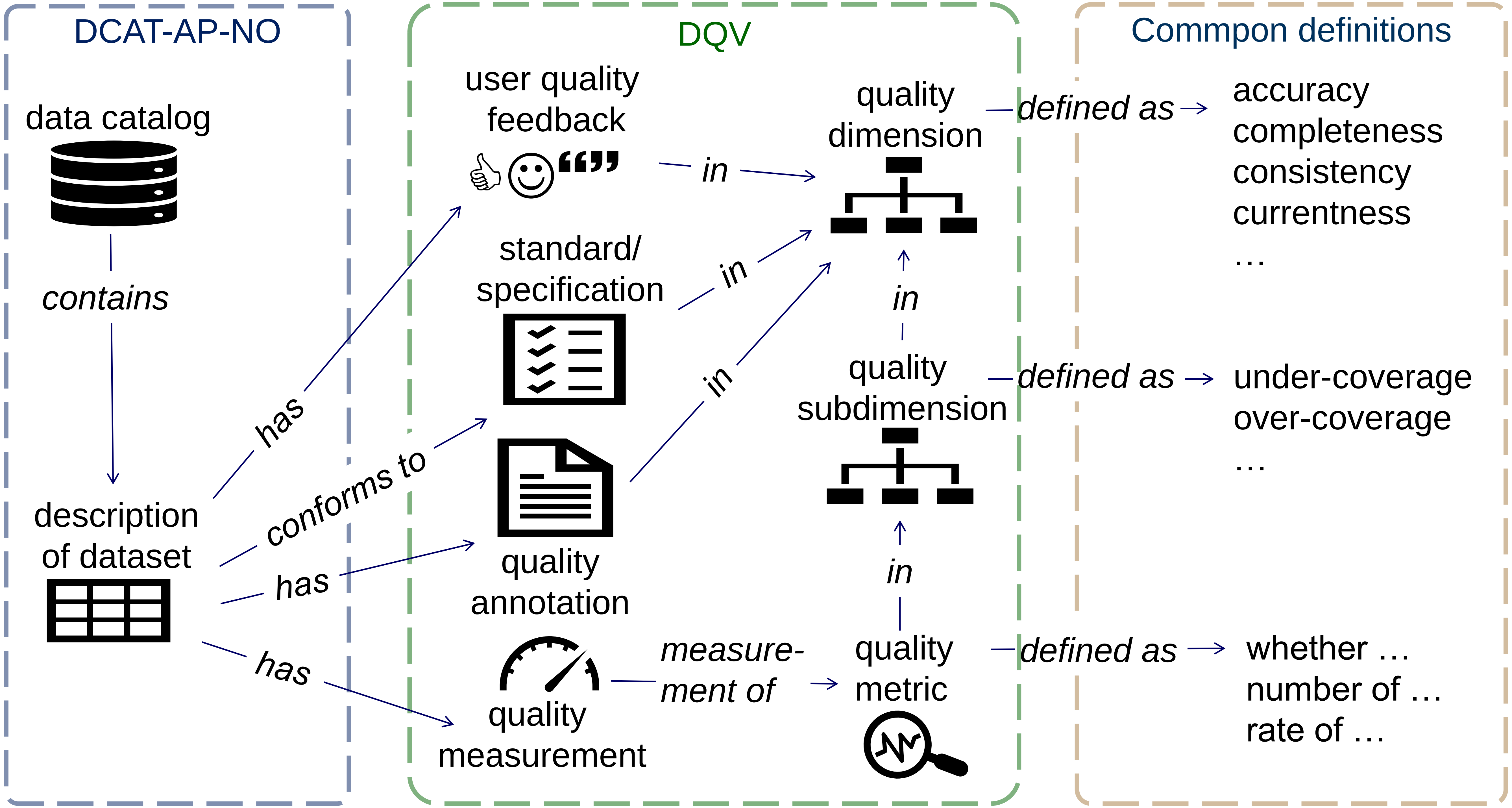
Definition from ISO 19157:2013

Definitions based on ISO 19157:2013

ISO 25012:2008 Software engineering — Software product Quality Requirements and Evaluation (SQuaRE) — Data quality model

ISO 19157:2013 Geographic information — Data quality

Summary – standardized, machine-readable, unified



Kartverket

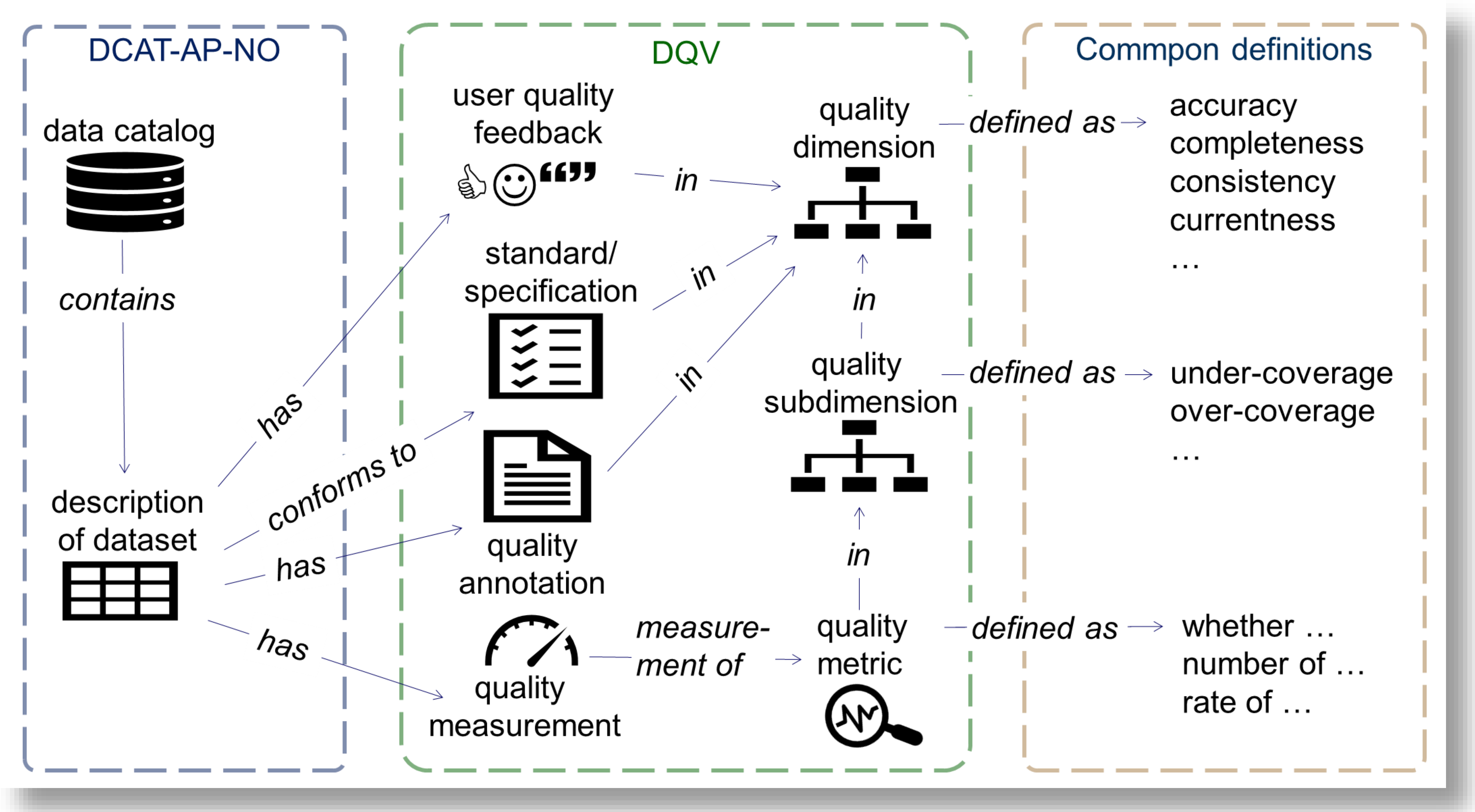
Digitaliseringsdirektoratet
Norwegian Digitalisation Agency

Future work

- Standardized machine-readable descriptions
  - DCAT-AP-NO will be revised, to incorporate DQV and to align with DCAT-AP v.2.0
- Pre-definitions
  - To be published bilingually and machine-readably
  - When needed, more common metrics/dimensions will be pre-defined
  - When needed, solutions for publishing (thus reusing) machine-readable sector/domain specific definitions

Thank you for your attention!

jim.yang@digdir.no;
anne.karete.hvidsten@digdir.no;
morten.borrebaek@kartverket.no



Kartverket

Digitaliseringsdirektoratet
Norwegian Digitalisation Agency