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

- **Foundation:** Framework and governance model
- **Ambition level 1:** Overview of existing data exists
- **Ambition level 2:** Data and APIs are available
- **Ambition level 3:** Reuse of data and APIs is the rule of thumb

Digital Agenda for Norway
Digitalization strategy
The digitalization circular
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
Describing ISO 19157 DQ_Result using DQV

1. \( dqv: \text{hasQualityMeasurement} \)  
   \( dqv: \text{QualityMeasurement} \)  
   \( dqv: \text{isMeasurementOf} \)  
   \( dqv: \text{Metric} \)  
   \( dqv: \text{inDimension} \)  
   \( dqv: \text{inDimension} \)  
   \( dqv: \text{Dimension} \)  

2. \( \text{dct:conformsTo} \)  
   \( \text{dct:Standard} \)  
   \( dqv: \text{inDimension} \)  
   \( dqv: \text{inDimension} \)  

3. \( dqv: \text{hasQualityAnnotation} \)  
   \( dqv: \text{QualityAnnotation} \)  
   \( dqv: \text{UserQualityFeedback} \)  

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 ;  
dqv:isMeasurementOfOf :Metric1 .
```

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 in 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 ;

! 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

ISO 19157 DQ_Completeness

- accuracy (dqv:Dimension)
- completeness (dqv:Dimension)
- currentness (dqv:Dimension)
- consistency (dqv:Dimension)

qualityDimension

- under-coverage (dqv:Dimension)
- over-coverage (dqv:Dimension)
- imputation (dqv:Dimension)

qualitySubDimension

ISO 19157 Omission
ISO 19157 Commission

SKOS: Simple Knowledge Organization System
Pre-defined metrics, subdimensions and dimensions

<table>
<thead>
<tr>
<th>Quality dimension</th>
<th>Quality subdimension</th>
<th>Quality metric (with data type)</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>completeness</td>
<td>under-coverage</td>
<td>missing objects (boolean)</td>
<td>&quot;false&quot; (all buildings present)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of missing objects (integer)</td>
<td>&quot;2&quot; (two buildings missing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of missing objects (percentage)</td>
<td>&quot;0.02%&quot; (0.02% of buildings missing)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of objects with missing value for a given property (integer)</td>
<td>&quot;2&quot; (two buildings with missing values for &quot;usable area&quot;)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of objects with missing value for a given property (percentage)</td>
<td>&quot;0.02%&quot; (0.02% of buildings with missing values for &quot;usable area&quot;)</td>
</tr>
<tr>
<td>over-coverage</td>
<td>excess objects (boolean)</td>
<td>&quot;true&quot; (some excess buildings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>number of excess objects (integer)</td>
<td>&quot;2&quot; (two excess buildings)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rate of excess objects (percentage)</td>
<td>&quot;0.02%&quot; (0.02% excess buildings)</td>
<td></td>
</tr>
<tr>
<td>imputation</td>
<td>number of objects with imputed value for a given property (integer)</td>
<td>&quot;2&quot; (two buildings with imputed values for &quot;year of construction&quot;)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rate of objects with imputed value for a given property (percentage)</td>
<td>&quot;0.02%&quot; (two buildings with imputed values for &quot;year of construction&quot;)</td>
<td></td>
</tr>
</tbody>
</table>

Please conf. the published paper for the definitions
<table>
<thead>
<tr>
<th>Quality dimension</th>
<th>Quality subdimension</th>
<th>Quality metric (with data type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>currentness</td>
<td>delay</td>
<td>overall time difference (xsd:duration)</td>
</tr>
<tr>
<td>consistency</td>
<td>consistency within the dataset</td>
<td>rate of objects with inconsistent properties (percentage)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of objects with inconsistency between given properties (percentage)</td>
</tr>
<tr>
<td>accuracy</td>
<td>identifier correctness</td>
<td>number of objects with incorrect identifiers (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of objects with incorrect identifiers (percentage)</td>
</tr>
<tr>
<td></td>
<td>classification correctness</td>
<td>number of incorrectly classified objects for a given property (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of incorrectly classified objects for a given property (percentage)</td>
</tr>
</tbody>
</table>

*Please conf. the published paper for the definitions*
## Mapping to ISO-standards

<table>
<thead>
<tr>
<th>Quality dimension</th>
<th>Quality subdimension</th>
<th>Quality metric (with data type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>completeness</td>
<td>under-coverage</td>
<td>missing objects (boolean)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of missing objects (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of missing objects (percentage)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of objects with missing value for a given property (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of objects with missing value for a given property (percentage)</td>
</tr>
<tr>
<td></td>
<td>over-coverage</td>
<td>excess objects (boolean)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>number of excess objects (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of excess objects (percentage)</td>
</tr>
<tr>
<td>imputation</td>
<td></td>
<td>number of objects with imputed value for a given property (integer)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rate of objects with imputed value for a given property (percentage)</td>
</tr>
</tbody>
</table>

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
## Mapping to ISO-standards (cont.)

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

Definitions from ISO 25012:2008

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

ISO 19157:2013 Geographic information — Data quality

Definitions based on ISO 19157:2013
Summary – standardized, machine-readable, unified

DCAT-AP-NO
- data catalog
  - contains
    - description of dataset
      - conforms to
        - has
          - quality annotation
            - has
              - quality measurement
                - measurement of
                  - quality metric
                    - in
                      - quality subdimension
                        - in
                          - quality dimension
                            - defined as
                              - user quality feedback
                                - defined as
                                  - standard/specification
                                    - in
                                      - coverage
                                        - defined as
                                          - under-coverage
                                            - over-coverage
                                              - defined as
                                                - whether... number of ...
                                                  - rate of...
                                            - in
                                              - consistency
                                                - defined as
                                                  - completeness
                                                    - accuracy
                                                      - defined as
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.borrebæk@kartverket.no