



Agency for
Data Supply and
Efficiency

New common method for declaring data quality in Denmark

Lars Erik Storgaard and Jan Hjelmager
Agency for Data Supply and Efficiency

Agenda

- Introduction
- Method
- Summery and conclusion

Introduction

Driver:

- eGovernment in Denmark
 - Danish Basic Data Program
 - Real property
 - Addresses
 - Person/individuals
 - Business
 - Geospatial data
 - Water and climate
 - Rules and governance
- EU/INSPIRE

Use of standards from W3C, OGC, ISO and CEN

Use of metadata and data quality standards

Introduction

Used standards:

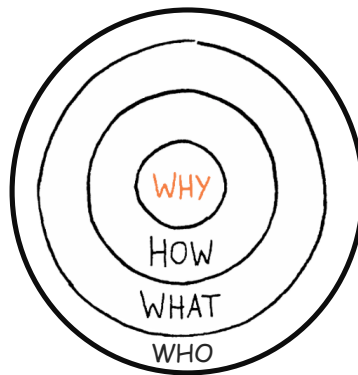
- ISO 19115/INSPIRE metadata guidance
- ISO 19157
- ISO 19158
- **ISO 25012**
- W3C Linked Data Quality Dimensions

Quality elements:

- Completeness
- Accuracy
- Currentness
- Reusability

Visualisation of quality element

- Quick
- Intuitive



Why?

Need for authorities to talk same language when assessing the quality of data from widely different domains. The goal is to make better use of data across the public sector.

How?

By using four quality elements, the new data quality declaration should enable a better and more secure way to have a dialogue about data quality across the public sector.

What?

A common public declaration template and undelaying vocabulary.

Who?

Public authorities and private companies are on voluntary basis welcome to use the declaration method.

Data on the Web Best Practices: Data Quality Vocabulary



W3C Working Group Note 15 December 2016

<https://www.w3.org/TR/vocab-dqv>



ISO 25012:2008

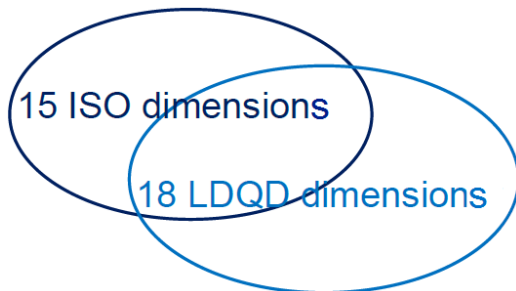
Software engineering –
Software product
Quality, Requirements
and Evaluation
(SQuaRE) – Data quality
model



15 ISO dimensions

18 LDQD dimensions

Linked Data
Quality
Dimensions



4 DQ Dimensions

The fundament:

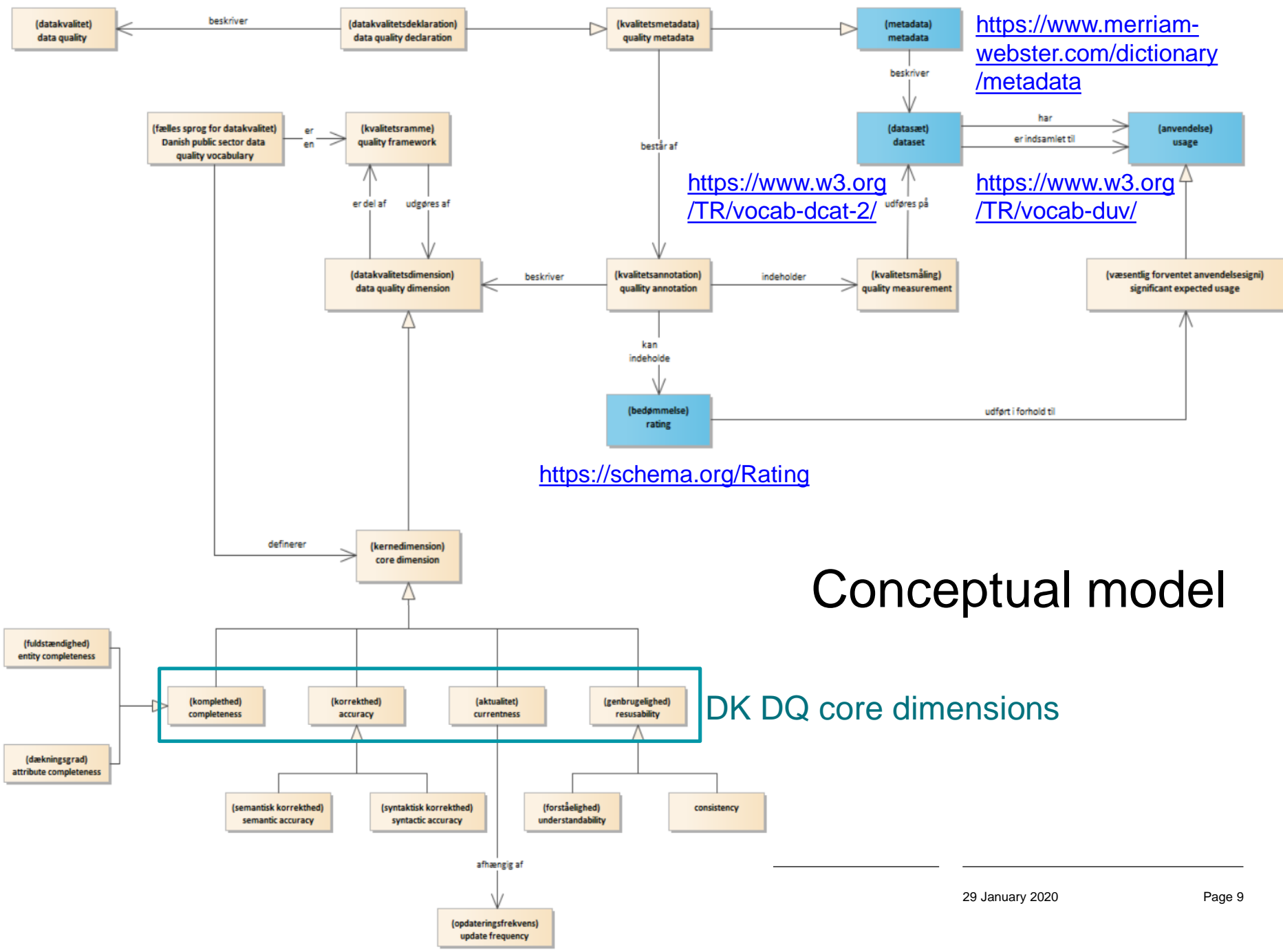
ISO 25012 og Linked Data Quality Dimensions – this is the two sets of DQ dimensions from W3Cs Data Quality Vocabulary

Four quality dimensions is chosen – named "core dimensions" – as the most important to evaluate the usability of a data set including new purposes

DQ dimension	Definition
completeness	the degree to which the data set contains the data elements that are expected based on the specification of the data set
accuracy	the degree to which data values corresponds to actual values
currentness	indicating to which degree data is current
reusability	the degree to which data is understandable and without difficulties can be used by others

Mapping to ISO 25012 and Linked DQ

"DK" DQ dimension	ISO 25012 dimension	Linked Data Quality dimension
completeness	skos:narrowMatch	skos:exactMatch
accuracy	skos:exactMatch	skos:narrowMatch
currentness	skos:exactMatch	skos:exactMatch
reusability	skos:narrowMatch	skos:narrowMatch



Conceptual model







The declaration template

The template has two fields for each quality dimension:

- 1) Measurements and descriptions and
- 2) Assessment.

Measurement and description is meant to provide objective information describing the data set in relation to the relevant dimension.

In **Assessment**, the data owner provides assessment of how good the quality of the data set is in relation to the relevant dimension when it is assessed in relation to the main use case (application).

	Rated unsuitable for the given application
	Is judged to be poorly used
	Can be used subject to a greater number of errors / unintended results
	Can be used subject to errors / unintended results
	Can be used subject to individual errors / unintended results
	Can be used immediately and unconditionally

The declaration template

DQ Dimension	Measurement and description <i>Describe the quality of the dataset in relation to each dimension.</i>	Assessment in relation to use <i>Fill in the appropriate number of stars.</i>
completeness		☆☆☆☆☆
accuracy		☆☆☆☆☆
currentness		☆☆☆☆☆
reusability		☆☆☆☆☆

Empirical method – Data set "Basemap DK-Building"

Dimension	Measures and descriptions <i>Describe the quality of the dataset in relation to each dimension.</i>	Assessment in relation to use <i>Fill in the appropriate number of stars</i>
Completeness	<p><i>What proportion of the individuals' dataset describing according to its specification is currently represented in the data set? - Is this an estimate?</i></p> <p>The completeness check is carried out by sampling in different numbers of samples and the number of grids. The inspection is performed as a visual inspection of the use of spring orthophoto from the year in which the data is updated (support data for centerline of rivers and lake from two additional years of spring orthophoto). SDFE, based on DS / EN ISO 19157: 2014, designates objects for the samples.</p> <p>99 percent of buildings visible in spring orthophoto from 2017 are represented in the topographic basic data set building theme. The percentage is calculated from random checks carried out in August 2018 according to DS / EN ISO 19157: 2014. The GeoDanmark specification allows for the completeness of 99 percent of buildings visible in Orthophoto.</p>	★★★★☆

Evaluation by sampling

Based on ISO 19157:2013 GeoDanmark data is delivered to a contractor for the sampling. The sample is based on 461 grids covering 10.586 km² containing 19.658 objects.

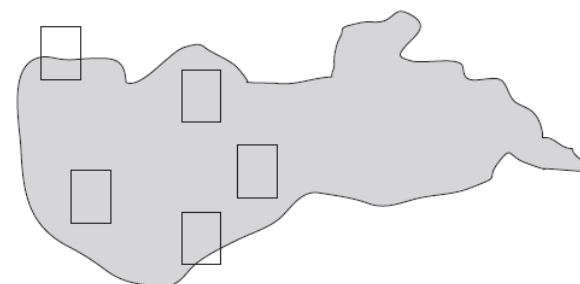
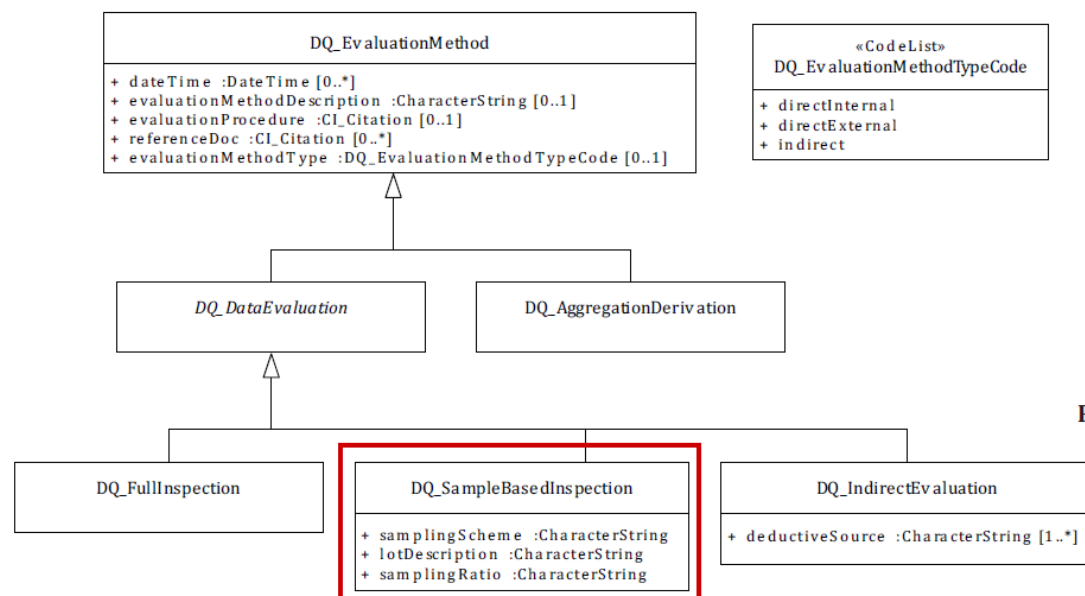


Figure F.3 — Example of area-guided random sampling

Figure 13 — Data quality evaluation methods

Accuracy

What proportion of the data values in the dataset can be expected to represent the actual value? And how did you come up with that figure?



The aforementioned sampling from August 2018, included checking for thematic errors and defined by the data attribute is not right for this particular concrete object evaluated by the specification and the situation in reality, as seen on the orthophoto. The check showed that no mis-registrations of buildings with the "Houseboats" type of building are present in the data.

The geometric accuracy understood as the coordinate accuracy at the individual points in the data set. This applies to both the accuracy of the plane and the altitude. The geometric accuracy of an object is an important factor for its quality. Accuracy in urban areas is 3-10 cm. in the plane and 15 cm. in height. In rural areas, the accuracy is 75 cm. in the plane and 75 cm. in height.

Validated data (current or prior to insertion in a database)? Is there a bug fix?

Have relevant syntactic validation, including spell checking, been performed?

GeoDanmark data is continuously checked in two processes:

- 1) Logical control of data in relation to structural requirements in the specification. Control is done in 1Integrate validation (a rule-based data validation tool that can be configured to quality control data sets) as well as in FME scripts (ETL tool). For example, a stream that has the attribute "stream through lake" even though it is not near a lake.
- 2) Control of undesirable changes in data (done in FME scripts ArcGIS Pro), e.g. that someone accidentally deleted all forests in a municipality.

Data validation using 1integrate

Use of 1integrate allows

- Scheduled validation of authoritative geospatial data stored within SDFE
- Validation of updated data from municipalities before importing it
- Consistent and higher quality data
- Most validations are scheduled to run on a weekly basis.
- Weekly reports are issued on the number and types of objects found
- Approx. 15 categories of rules are developed replicated to > 50 feature types.

Report for SAMSOE2017_1

Validation performed by Annette Stammerjohan on data extracted 2017-10-27. Check performed 2017-11-06.

Statistics

Statistical analysis between original dataset and updated dataset divided on object type.

Featuretype	Number - difference		Area - difference		Length - difference	
	%	Antal	%	Areal	%	Længde
Road	2,7	117			5,2	1.532
Forest	-1,5	-3	-25,2	-1.250.756		
Building	0,0	0	0,03	345		

If there is no difference between object types in original and updated dataset, they are not displayed in the table.

Negative numbers indicate that the updated dataset has a smaller value than the original dataset.

Business markers

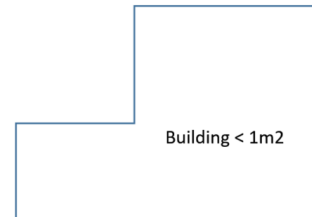
List of types and number of business markers identified in the updated dataset.

Code	Business marker	Number
n1	Building overlap	3
n2	Moor with backloop or spike < 15 grader	1
n3	Low settlement overlapping med technical area	1
n4	Forest with area > 2500 m ² and underMinimum = True	2
n5	Roads which do not snap correctly	3

Examples of rules

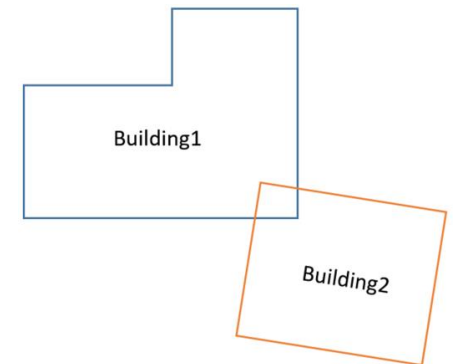
Building < 1m²

Buildings with an area of less than 1m²



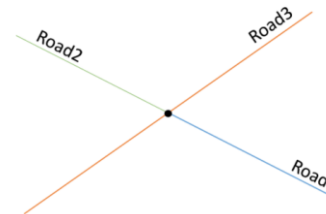
BuildingOverlap > 10m²

Buildings which overlap with more than 10m²



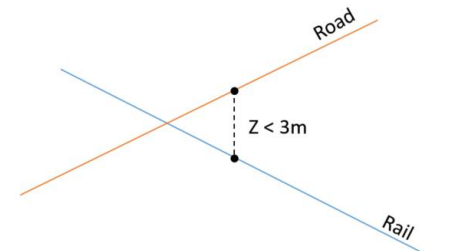
RoadSplitCross

Roads which are not split correctly where they cross other roads



roadXrail

Road crossing a railway with less than 3 m vertical distance

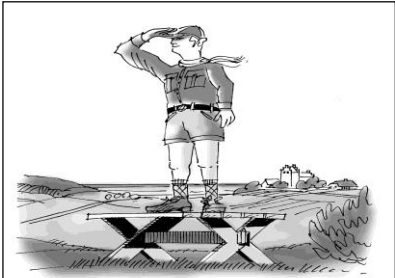


Currentness	<p><i>Is the dataset updated periodically or as needed?</i></p> <p>GeoDanmark data is updated both after a fixed cycle and as needed.</p> <p>Periodically updated: GeoDanmark data updated nationwide in a fixed cycle using photogrammetry. Photogrammetry is a method of locating objects by measuring in images taken from aircraft. An annual photogrammetric total update of GeoDanmark data is carried out in one of Denmark's five regions: the Region of Northern Jutland, the Region of Central Jutland, the Region of Southern Denmark, the Region of Zealand or the Capital Region of Denmark. In addition, photogrammetric updating of the other regions based on change designations that the municipalities and third parties have registered or the objects that have been updated by administrative updating. Thus, all of Denmark is fully updated over a period of 5 years while the most central objects in GeoDanmark data are updated annually.</p> <p>Update as needed: Administrative updating means data is updated independently of the annual joint update.</p> <p>Administrative updating is carried out eg. by a building case manager who gives a permit, and immediately the new building draws in data. Administrative updating can also be done as part of a major improvement project, where all or part of the nationwide data set is improved, e.g. to obtain uniform data across the country.</p>	<p>★★★★☆</p>
-------------	---	--------------

<p>Reusability</p>	<p><i>Does the dataset comply with relevant laws and standards?</i> Yes, section 4 of http://www.retsinformation.dk/eli/Ita/2017/380</p> <p><i>Does the dataset use relevant common classifications?</i> Yes, from the Danish Basic Data Program.</p> <p><i>Is business rules observed? And documented?</i> Yes, the rules are documented in the GeoDanmark Specification.</p> <p><i>Is the management of the data set documented?</i> Yes, in SDFEs Quality Manual (documentation of business services).</p> <p><i>Is there an accessible data model?</i> Yes, see: http://data.gov.dk/model/xmi/Domaenemodeller/GeoDanmark/</p> <p><i>Is the data model machine-readable?</i> Yes, see: http://data.gov.dk/model/xmi/Domaenemodeller/GeoDanmark/2.0.0_GeoDanmark.xml</p> <p><i>Does the data set comply with the data model?</i> Yes, the database schema is generated from the data model and data schema validated in order to be loaded into the database.</p> <p><i>Are data types, incl. outcomes, units and precision where relevant, documented?</i> Yes, data types are documented in the data model.</p> <p><i>Is the dataset provided with relevant metadata?</i> Yes, see: https://www.geodata-info.dk/srv/dan/catalog.search#/metadata/57155d38-aa5d-63c9-467e-870558753748</p> <p><i>Are the data type specifications met?</i> Yes, see. the Modeling Rules of the Danish Basic Data Program</p> <p><i>Are data of the same type, e.g. dates set uniform?</i> Yes, see. the Modeling Rules of the Danish Basic Data Program</p> <p><i>Are data values used unambiguous and understandable?</i> Yes, see. the Modeling Rules of the Danish Basic Data Program</p> <p><i>Does the dataset contain information about the lineage of the data elements?</i> Yes, see. the Modeling Rules of the Danish Basic Data Program</p> <p><i>Is the data set free of conflicting information?</i> Yes</p> <p><i>Refers solely to information actually exist?</i> Yes</p>	<p>★★★★★</p>
--------------------	---	--------------

Summery and conclusion

- Terminology issues between the standards
 - Mapping was required
- Visualisation and publication
- Future work
 - More detailed metadata and quality description
 - Visualisation and rating for more datailed descriptions



- **Discovery quality**

What are the overall quality of the data sets I am interested in? This enable organisations to publicise the quality of what data holdings they have.

- **Exploration quality**

Do the identified data sets contain sufficient quality information to enable a sensible analysis to be made for my purposes? This is documentation to be provided with the data to ensure that others use the data correctly and wisely



