



Norwegian University of
Science and Technology

Experiences on using Data Quality Measures for automatic validation of geographical datasets

Knut Jetlund and Erling Onstein

2ND INTERNATIONAL WORKSHOP ON
SPATIAL DATA QUALITY

Dates: 6th to 7th February 2018

Venue: Old University Campus, Valletta, Malta



Sponsor: www.geoforum.no, funding the travel to Malta

Content

- Background
- ISO/DQ framework
- Implementation of framework in Norway
- Case 1: Transport network data
- Case 2: Areal plan and DQ
- Conclusions / Further work

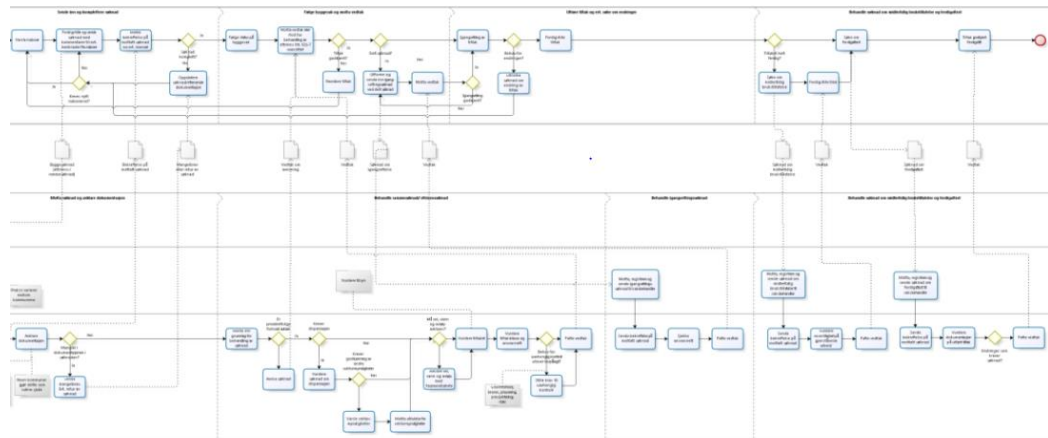
Background:

Digitizing of environment and processes

- Two parts:
 - Real world facts
 - Decisions based on facts
- We are clever at digitizing real world facts
 - Physical environment, political decisions
- We are just in the beginning of digitizing processes
 - Digital decisions based on digital real world facts
- Areal Plan example:
 - We can validate candidate plans, but not choose the best candidate

Digital decisions

- Must be based on reliable real world facts (validation)
- Digital decisions are carried out by robots, not people
 - The rules leading to good decisions must be computer-interpretable



ISO framework for data quality

- Dataset specifications – ISO 19131
- Data Quality – ISO 19157
 - Further developed from ISO 19113, 19114 and 19138
- Data Quality evaluation – ISO 19157
 - Important mechanism: Data Quality Measures
- Quality assurance of data supply – ISO/TS 19158
- Metadata – ISO 19115

Implementation in Norway

- ISO 19131 DPS and ISO 19157 DQ adopted as Norwegian standards
- Both also followed up with additional work:
 - SOSI Produktspesifikasjon (2014)
 - Geodatakvalitet (2015)



Norsk Standard
NS-EN ISO 19131:2008

ICS 35.240.70
Språk: Engelsk



Norsk Standard
NS-EN ISO 19157:2013

ICS 35.240.70
Språk: Engelsk

Are they then implemented?
...hardly.. They are translated

National SDI framework

- National regulations (from ministry)
 - Member organisations **shall** maintain and update **datasets** and corresponding **metadata**
- Criteria for DOK for national public organisations (from geodata coordinator)
 - Datasets shall have a valid product specification according to national DPS standard

Example DQ requirement Dataset Roads

- Part 1 –
classification of
feature types to
DQ classes

Objekttype	Klasser stedfestingsnøyaktighet								Klasser fullstendighet	
	Grunnriss				Høyde				1	2
	1	2	3	4	1	2	3	4		
Vegskulderkant		X			X				X	
Vegdekkekant	X ¹				X ¹				X	
Kjørebane	X				X				X	
Trafikkø									X	
Trafikkøykant	X ¹				X ¹				X	
Fortauskant	X ¹				X ¹				X	
VegkantAvkjørsel		X			X				X	
VegkantAnnetVegareal		X			X				X	
AnnetVegarealAvgrensning			X		X					X
VegkantFiktiv										X
Veg									X	
VeggrøftÅpen			X			X				X
GangSykkelveg									X	
GangSykkelvegkant		X			X				X	
Gangvegkant		X			X					X
Parkeringsområde										X
FartsdemperAvgrensning		X			X					X
FeristAvgrensning		X			X					X
Trafikksignalpunkt		X				X				X
VegoppmerkingLangsgående	X				X				X	
Skiltportal		X					X		X	
GangfeltAvgrensning		X			X					X
Vegrekkverk		X				X				X
Vegsperring		X				X				X
Traktorveg										X
Traktorvegkant				X				X		X

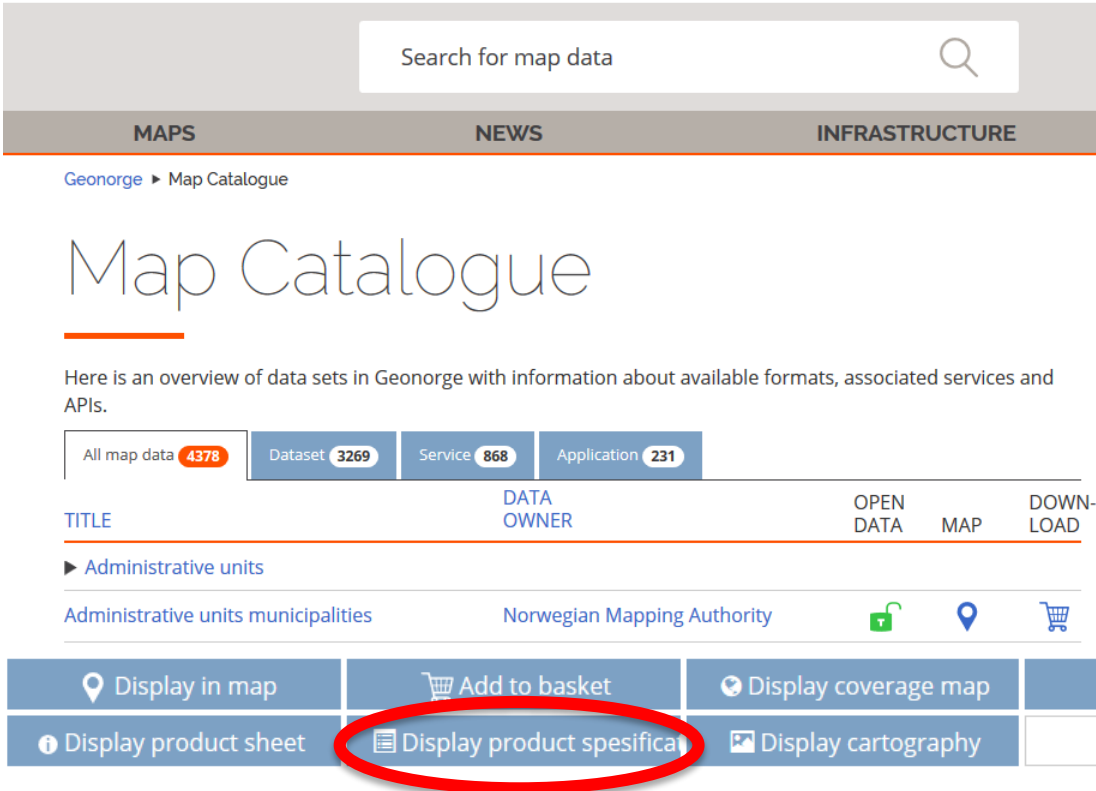
Example DQ Tolerances Dataset Roads

- Part 2 DQ quality requirements

Kvalitetskategori	Kvalitetselement	Kvalitetsmål	Klasse	FKB-standard			
				A	B	C	D
				Krav	Krav	Krav	Krav
Fullstendighet	manglende data	andel manglende enheter	1	0.5 %	0.5 %	0.5 %	0.5 %
Fullstendighet	manglende data	andel manglende enheter	2	2 %	2 %	2 %	2 %
Fullstendighet	overskytende data	andel overskytende enheter	1	0.5 %	0.5 %	0.5 %	0.5 %
Fullstendighet	overskytende data	andel overskytende enheter	2	2 %	2 %	2 %	2 %
Stedfestingsnøyaktighet	absolutt stedfestingsnøyaktighet	stedfesting - Prosentandel grove feil		1 %	1 %	1 %	1 %
Stedfestingsnøyaktighet	absolutt grunnrissnøyaktighet	stedfesting - Standardavvik	1	0.10 m	0.15 m	0.48 m	0.48 m
Stedfestingsnøyaktighet	absolutt grunnrissnøyaktighet	stedfesting - Standardavvik	2	0.15 m	0.20 m	0.55 m	0.55 m
Stedfestingsnøyaktighet	absolutt grunnrissnøyaktighet	stedfesting - Standardavvik	3	0.35 m	0.35 m	0.70 m	0.70 m
Stedfestingsnøyaktighet	absolutt grunnrissnøyaktighet	stedfesting - Standardavvik	4	0.55 m	0.55 m	1.00 m	1.00 m
Stedfestingsnøyaktighet	absolutt høydenøyaktighet	stedfesting - Standardavvik	1	0.10 m	0.15 m	0.48 m	0.48 m
Stedfestingsnøyaktighet	absolutt høydenøyaktighet	stedfesting - Standardavvik	2	0.15 m	0.20 m	0.70 m	0.70 m
Stedfestingsnøyaktighet	absolutt høydenøyaktighet	stedfesting - Standardavvik	3	0.25 m	0.35 m	0.90 m	0.90 m
Stedfestingsnøyaktighet	absolutt høydenøyaktighet	stedfesting - Standardavvik	4	0.40 m	0.50 m	1.50 m	1.50 m
Egenskapskvalitet	klassifikasjonsnøyaktighet	feilklassifikasjons andel		0.5 %	0.5 %	0.5 %	0.5 %
Logisk konsistens	formatkonsistens	formatkonsistens		0	0	0	0
Logisk konsistens	konseptuell konsistens	antall enheter der regler i konseptuelt skjema ikke er oppfylt		0	0	0	0
Logisk konsistens	topologisk konsistens	antall ulovlige småpolygoner		0	0	0	0
Logisk konsistens	topologisk konsistens	antall ulovlige egenkryssinger		0	0	0	0
Logisk konsistens	topologisk konsistens	antall ulovlige egenoverlappinger		0	0	0	0
Logisk konsistens	topologisk konsistens	antall ulovlige løse ender		0	0	0	0
Logisk konsistens	topologisk konsistens	antall ulovlige lenkekryssing		0	0	0	0

Use of the standards in NSDI

- Every dataset is expected to have a product specification according to the national DPS standard
- Implies inclusion of data quality requirements



Search for map data




MAPS NEWS INFRASTRUCTURE






Geonorge ► Map Catalogue






Map Catalogue

Here is an overview of data sets in Geonorge with information about available formats, associated services and APIs.

All map data **4378** Dataset **3269** Service **868** Application **231**

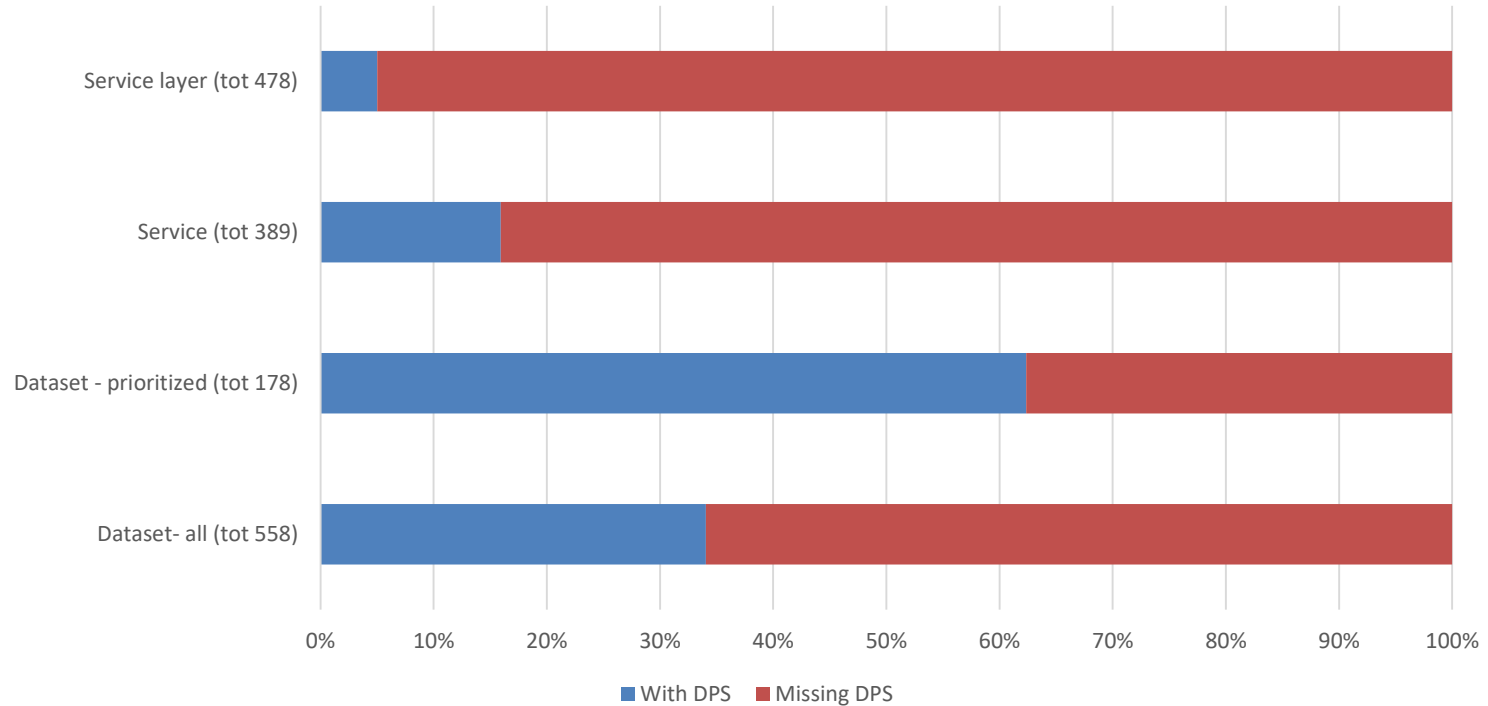
TITLE	DATA OWNER	OPEN DATA	MAP	DOWN-LOAD
► Administrative units				
Administrative units municipalities	Norwegian Mapping Authority			

 Display in map
  Add to basket
  Display coverage map
  Help
  Contact dataowner

 Display product sheet
  **Display product specification**
 Display cartography
  Webpage
  Display productpage

Statistics www.geonorge.no

Use of Data Product Specifications



Possible reasons for low score?

- In spite of being digital information, common use include human interpretation
 - Information in the datasets not used for digital decisions
 - Digital quality information not really needed ??
- Are there any differences in DQ requirements for “professional data maintenance” compared to “end user use”?
- Main source for DPS is documenting existing data. Another situation when producing new data?

DQ status in two selected cases

- Transport network data
- Areal plans
- Method (for both):
 - Reading specifications searching data quality statements and requirements and hopefully data quality measures.

Transport network data

- High demand for data for vehicle/driver support
- Human drivers are being replaced by robots
 - Digital decisions needed
- Need for re-thinking the need for digital validation and evaluation of data?



INSPIRE
Infrastructure for Spatial Information in Europe

D2.8.1.7 Data Specification on Transport Networks – Technical Guidelines

Section	Data quality element	Data quality sub-element	Definition	Evaluation Scope	Quality purpose
7.1.1	Completeness	Commission	excess data present in the dataset, as described by the scope	dataset	evaluation
7.1.2	Completeness	Omission	data absent from the dataset, as described by the scope	dataset	evaluation
7.1.3	Logical consistency	Conceptual consistency	adherence to rules of the conceptual schema	spatial object type; spatial object	evaluation
7.1.4	Logical consistency	Domain consistency	adherence of values to the value domains	spatial object type; spatial object	evaluation
7.1.5	Logical consistency	Format consistency	degree to which data is stored in accordance with the physical structure of the dataset, as described by the scope	dataset	evaluation
7.1.6	Logical consistency	Topological consistency	correctness of the explicitly encoded topological characteristics of the dataset, as described by the scope	dataset	network
7.1.7	Positional accuracy	Absolute or external accuracy	closeness of reported coordinate values to values accepted as or being true	dataset	evaluation
7.1.8	Thematic accuracy	Classification correctness	comparison of the classes assigned to features or their attributes to a universe of discourse	dataset	evaluation

Recommendation 18 Where it is impossible to express the evaluation of a data quality element in a quantitative way, the evaluation of the element should be expressed with a textual statement as a data quality descriptive result.

Recommendation 20 Omission should be evaluated and documented using *Rate of missing items* as specified in the table below.

Recommendation 21 For the tests on conceptual consistency, it is recommended to use the *Logical consistency – Conceptual consistency* data quality sub-element and the measure *Number of items not compliant with the rules of the conceptual schema* as specified in the table below.

Recommendation 24 Topological consistency should be evaluated and documented using *Number of invalid overlaps of surfaces, Number of missing connections due to undershoots, Number of missing connections due to overshoots, Number of invalid slivers, Number of invalid self-intersect errors, Number of invalid self-overlap errors* as specified in the tables below.

INSPIRE	Reference: D2.8.1		
TWG-TN	Data Specification on <i>Transport Networks</i>	2014-04-17	Page 30

Recommendation 6 The objects in the Transport Networks theme should be positionally consistent with spatial objects from other themes (e.g. with buildings and rivers, forestry extents)

COMMISSION REGULATION (EU) No 1089/2010

of 23 November 2010

implementing Directive 2007/2/EC of the European Parliament and of the Council as regards interoperability of spatial data sets and services

(OJ L 323, 8.12.2010, p. 11)

7.9. Theme-specific Requirements

7.9.1. Consistency between spatial data sets

1. Transport Networks centreline representations and nodes shall always be located within the extent of the area representation of the same object.

INSPIRE	Reference: D2.8.1.7_v3.2		
TWG-TN	Data Specification on <i>Transport Networks</i>	2014-04-17	Page 128

7 Data quality

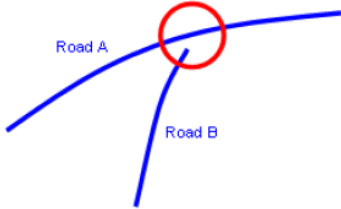
INSPIRE	Reference: D2.8.1.7_v3.2		
TWG-TN	Data Specification on <i>Transport Networks</i>	2014-04-17	Page 128

Section	Data quality element	Data quality sub-element	Definition	Evaluation Scope	Quality purpose
7.1.1	Completeness	Commission	excess data present in the dataset, as described by the scope	dataset	evaluation
7.1.2	Completeness	Omission	data absent from the dataset, as described by the scope	dataset	evaluation
7.1.3	Logical consistency	Conceptual consistency	adherence to rules of the conceptual schema	spatial object type; spatial object	evaluation
7.1.4	Logical consistency	Domain consistency	adherence of values to the value domains	spatial object type; spatial object	evaluation
7.1.5	Logical consistency	Format consistency	degree to which data is stored in accordance with the physical structure of the dataset, as described by the scope	dataset	evaluation

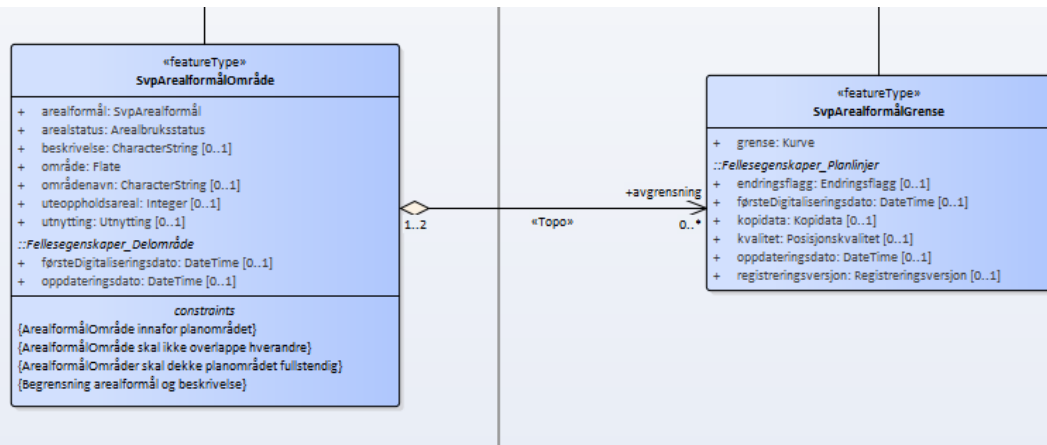
7.1.6 Logical Consistency – Topological consistency

Recommendation 24 Topological consistency should be evaluated and documented using *Number of invalid overlaps of surfaces, Number of missing connections due to undershoots, Number of missing connections due to overshoots, Number of invalid slivers, Number of invalid self-intersect errors, Number of invalid self-overlap errors* as specified in the tables below.

Name	Number of missing connections due to undershoots
Alternative name	Undershoots
Data quality element	Logical consistency
Data quality subelement	Topological consistency
Data quality basic measure	Error count
Definition	Count of items in the dataset that are mismatched due to undershoots, given the parameter <i>Connectivity tolerance</i> .
Description	Lacks of connectivity exceeding the <i>Connectivity tolerance</i> are considered as errors if the real features are connected in the transport network.
Evaluation scope	data set
Reporting scope	data set; spatial object type
Parameter	<ul style="list-style-type: none"> Name: <i>Connectivity tolerance</i> Definition: Search distance from the end of a dangling line.

Data quality value type	Integer
Data quality value structure	-
Source reference	-
Example	 <p>Key 1 Connectivity tolerance = 3 m</p>
Measure identifier	23

Areal plan and constraints



Spatial requirement	ISO19125:1 method	GEOS/PostGIS method	Used for
Inside	Covered By	ST_CoveredBy	Test for inside condition
No Overlap	Overlaps	ST_Overlaps	Test for overlap
	Intersect	ST_Intersect	Identify and visualize overlapping areas
Complete tessellation	Union	ST_Union	Merge zone divisions inside regulatory plan areas
	Difference	ST_Difference	Create polygons of areas without zone divisions

Figure 7 Relevant ISO 19125-1 Methods and PostGIS equivalent terms

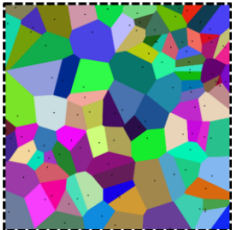
Line	Component	
1	Name	CompleteTessellation
2	Alias	Tessellation complete
3	Element name	Conceptual consistency
4	Basic measure	Either Error indicator, Error count or Error rate
5	Definition	Total number of erroneous polygons within the data
6	Description	All child polygons must be located inside the parent polygon. Reports the number of child polygons not inside the parent polygon. A set of feature instances (set of childs) which completely covers the tested feature (the parent)
7	Parameter	Parameter: Identification of child polygons
8	Value type	Either Boolean, Integer or Percentage
9	Value structure	
10	Source reference	
11	Example	 <p>Parent polygon (dashed quadrat) filled correctly with child polygons (coloured). Errors when gaps between childs, and also error when overlap between childs.</p> <p>Reporting by parent polygons.</p>
12	Identifier	NTNU/ConceptualConsistency03

Figure 6 DQM Complete Tessellation

Source fig 6 and fig 7:

Onstein, Erling; Stikbakke, Sverre. (2017) [Exploring subset profile and validation procedures of geographical markup language \(GML\) for 3D areal plan information. International Multidisciplinary Scientific GeoConference SGEM vol. 17 \(21\).](#)

Further work: User needs investigation

- What kind of data quality information are needed?
- Need for levels of DQ:
 - For professionals / data owners
 - For others
 - For automated use, e.g. in driver supporting systems

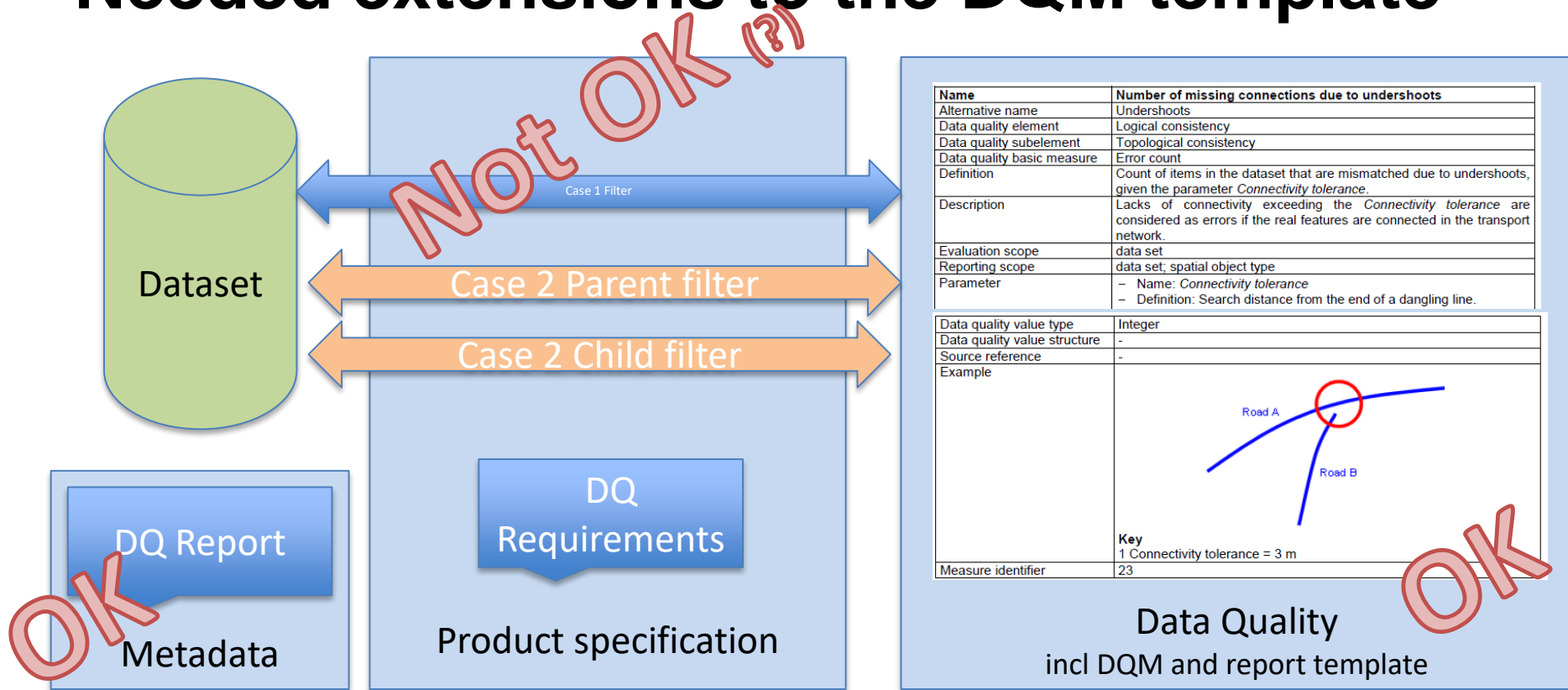
Further work: Support for digital decisions

- Model-driven architecture imply conceptual UML-models can be transformed to executable platform (e.g. XML/GML)
 - Information structure validation solved
- Also needed:
 - Geometry validation (closed solids, closed rings, connected curves,...)
 - Logical consistency validation needed (inside, not overlap,...)
 - Product spec info, e.g.
 - Spatial Reference Systems, DQ requirement/tolerances
- The needed solution for validation: All the above
 - implemented in one single tool,
 - with an understandable/useable output
 - output both for humans and robots
- Further needs:
 - Digitalization of processes and process requirements

Further work: Extended DQ framework

- Connection between conceptual schema and DQMeasure especially for Logical Consistency
 - A conceptual schema constraint for e.g. inside will be absolute
 - A DQMeasure requiring “inside” will open for conformance levels/tolerances, e.g. 5% in error
- Conceptual schema constraints (CSC) must be possible to validate
 - DQM include computational procedures
 - Using references to DQM when defining CSC will help this

Needed extensions to the DQM template



Summing up

1. Lots of energy have been used defining data quality framework
2. Not fully implemented, and possibly not really needed for everybody?
3. Automated use of data will require automated validation
 - Then the established DQ framework effort will be needed
4. Need for further development of
 - The DQ framework, included user needs for DQ
 - Automated data validation
 - Support for digital decisions
5. Short term: Need for authoritative GML Validation tool!!
 - *Anybody volunteering for participation? For funding?*

Thank you!

- Associate Professor Erling Onstein
 - erling.onstein@ntnu.no
- PhD-candidate Knut Jetlund
 - knut.jetlund@stud.ntnu.no