

# A Register for Data Quality Measures

WorkShop Standards in Action  
Copenhagen, Denmark  
28 May 2008



DTU



**Dr. Gerhard Joos**

Associate Professor - Technical University of Denmark

[gj@space.dtu.dk](mailto:gj@space.dtu.dk)

<http://www.dotGIS.de>

# Motivation

- How is the data quality measure "number of faulty point-curve connections" defined?
  - ISO 19138 provides the definition and examples
  - Should the definition be repeated in each metadata record whenever the data quality measure is used?
  - If the definition gets changed – e.g. due to an error – should then all metadata records be corrected?



There is a need for a flexible administration of normative elements with history tracing

**Register services**

# Definitions from ISO 19135

ISO 19135 Geographic information - Procedures for registration of items of geographic information (ISO 19135:2005)

- **Register**
    - set of files containing **identifiers** assigned to items with descriptions of the associated items
  - **Registry**
    - information system on which a **register** is maintained
- 
- **Register service**
    - The register should be accessible to users through an internet web site or other electronically processable form

# Why register services?

- There are normative elements, that may change occasionally
  - New entries may come in, may have to be retired or changed
  - Printed standards are not flexible enough
    - From starting an amendment or a corrigendum to the finalization and publication several months pass by
  - Users will not dispence with the bindingness and the assured quality of standards coming from a trustworthy organization

# How is that related with Geoinformation?

- Several aspects of Geoinformation require registers:
  - CoordinateReferenceSystems
  - Symbols
  - Terms and definitions
  - Feature definitions (Concept dictionaries, catalogues)
  - Data quality measures
  - Code lists
  - ...

# Is Geoinformatics the only community with this requirement for registers?

- OASIS works since 1999 on a standard for electronic business processes
- Standards supporting electronic businesses are also known as **ebXML**
- The OASIS standards were accepted as ISO-standards
  - ISO 15000-1: ebXML Collaborative Partner Profile Agreement
  - ISO 15000-2: ebXML Messaging Service Specification
  - ISO 15000-3: ebXML Registry Information Model
  - ISO 15000-4: ebXML Registry Services Specification
  - ISO 15000-5: ebXML Core Components Technical Specification, Version 2.01.

# ISO 19135 Procedures for item registration

- ISO/TC 211 has a standard for registers: ISO 19135
- How is ebXML and ISO 19135 related?
  - ISO 19135 addresses
    - Enterprise viewpoint
    - Information viewpoint
  - ebXML addresses
    - Computational Viewpoint
    - Engineering Viewpoint

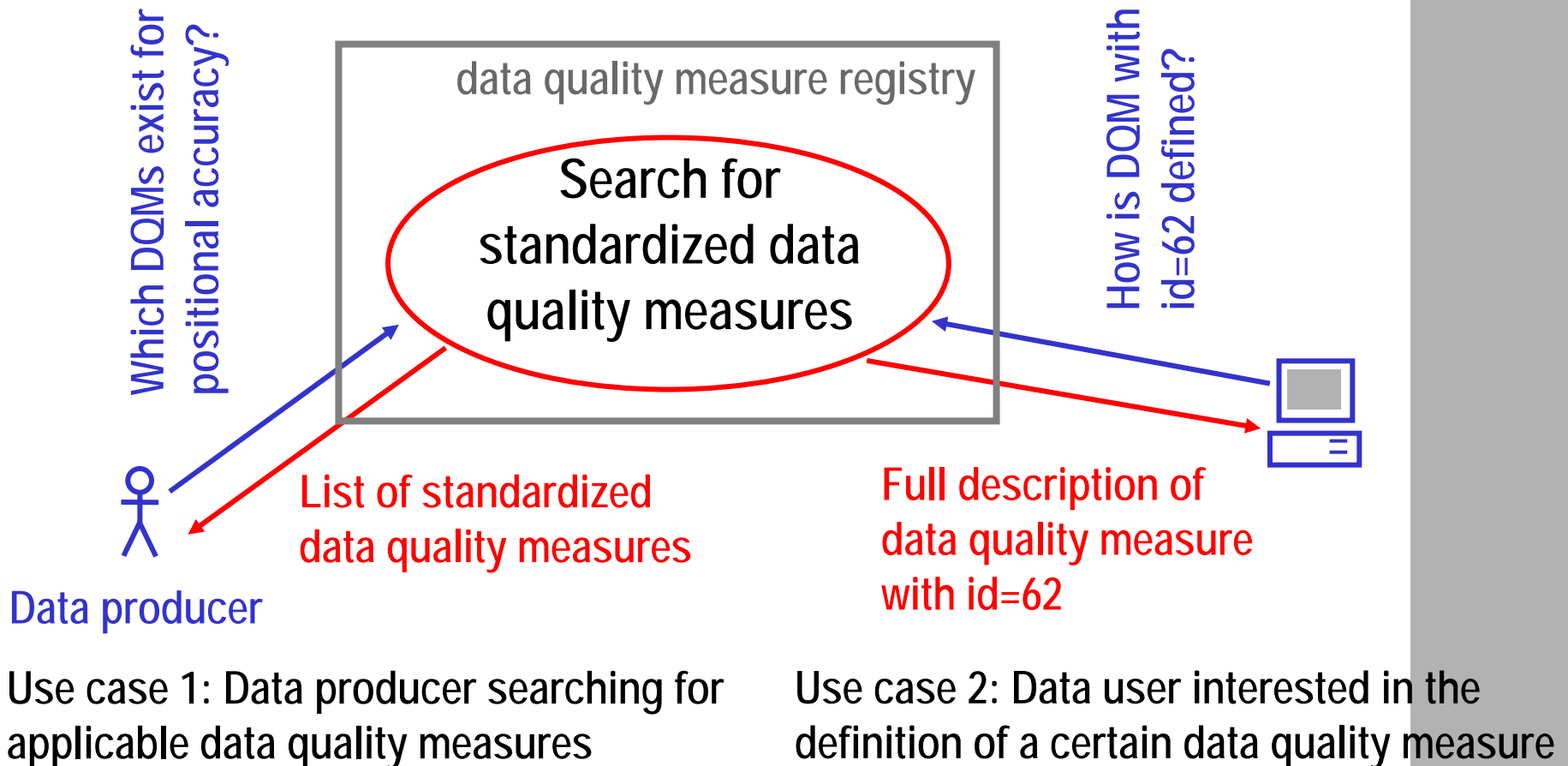
# Register administration

- All entries of a register get administrative register metadata
  - About the submitter for addition, change or retirement
  - Time stamp
  - Entry type
  - Additional administrative metadata elements
  - Metadata about the register specific elements have to be defined (ebXML-RIM provides slots for that)

# Advantages of Registers

- Online access to normative elements
  - Data quality measures (ISO 19138),
  - Coordinate Reference Systems (ISO 19111),
  - Symbols and portrayal rules (ISO 19117 Portrayal),
- Flexibility
- Administration of register metadata (e.g. originator, registration date, ...)
- Only approved elements are registered
- Short time from application to acceptance and online availability

# Use cases

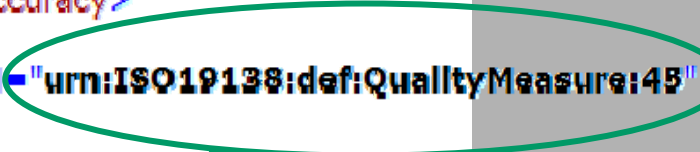


# Comparison

- Quality evaluation results without reference to a DQM register
- Quality evaluation results with reference to a DQM register

```
- <coordinateOperationAccuracy>  
- <DQ_RelativeInternalPositionalAccuracy>  
  - <result>  
    - <DQ_QuantitativeResult>  
      - <circularError>  
        - <significanceLevel>  
          <valueUnit>%</valueUnit>  
          <Decimal>95.0</Decimal>  
        </significanceLevel>  
        <valueUnit>m</valueUnit>  
      - <value>  
        - <Record>  
          <Decimal>3.0</Decimal>  
        </Record>  
      </value>  
    </circularError>  
  </DQ_QuantitativeResult>  
</result>  
</DQ_RelativeInternalPositionalAccuracy>  
</coordinateOperationAccuracy>
```

```
- <coordinateOperationAccuracy>  
- <DQ_RelativeInternalPositionalAccuracy>  
  - <result>  
    - <DQ_QuantitativeResult DQM="urn:ISO19138:def:QualityMeasure:45">  
      - <value>  
        - <Record>  
          <Decimal uom="urn:x-ogp:def:uom:EPSG:9001">3.0</Decimal>  
        </Record>  
      </value>  
    </DQ_QuantitativeResult>  
  </result>  
</DQ_RelativeInternalPositionalAccuracy>  
</coordinateOperationAccuracy>
```



Registry

# Register entries for data quality measures

Line	Component	
1	Name	circular error at 95
2	Alias	Navigation accuracy
3	Data quality element	positional accuracy
4	Data quality subelement	absolute or external
5	Data quality basic measure	CE95
6	Definition	Radius describing probability of 95%
7	Description	See C.3.3
8	Parameter	-
9	Data quality value type	measure
10	Data quality value structure	-
11	Source reference	-
12	Example	
13	Identifier	45

## C.3.3 Two-dimensional random variable $X$ and $Y$

The case of the one-dimensional random variable  $Z$  can be expanded to two dimensions where the measurand is always observed by two values. The measurand is given by the tuple  $X, Y$ . They underlay the same assumptions as in the case of the one-dimensional random variable.

The observations are  $x_{mi}$  and  $y_{mi}$ . The equivalence of the confidence interval in one dimension is the confidence area, which is usually described as a circle around the best estimation for the true value. The probability for the true value to lie in this area is calculated by area integration over the two-dimensional density function of the normal distribution. A circular area is characterised by its radius. This radius is used as measure for the accuracy of two-dimensional random variables.

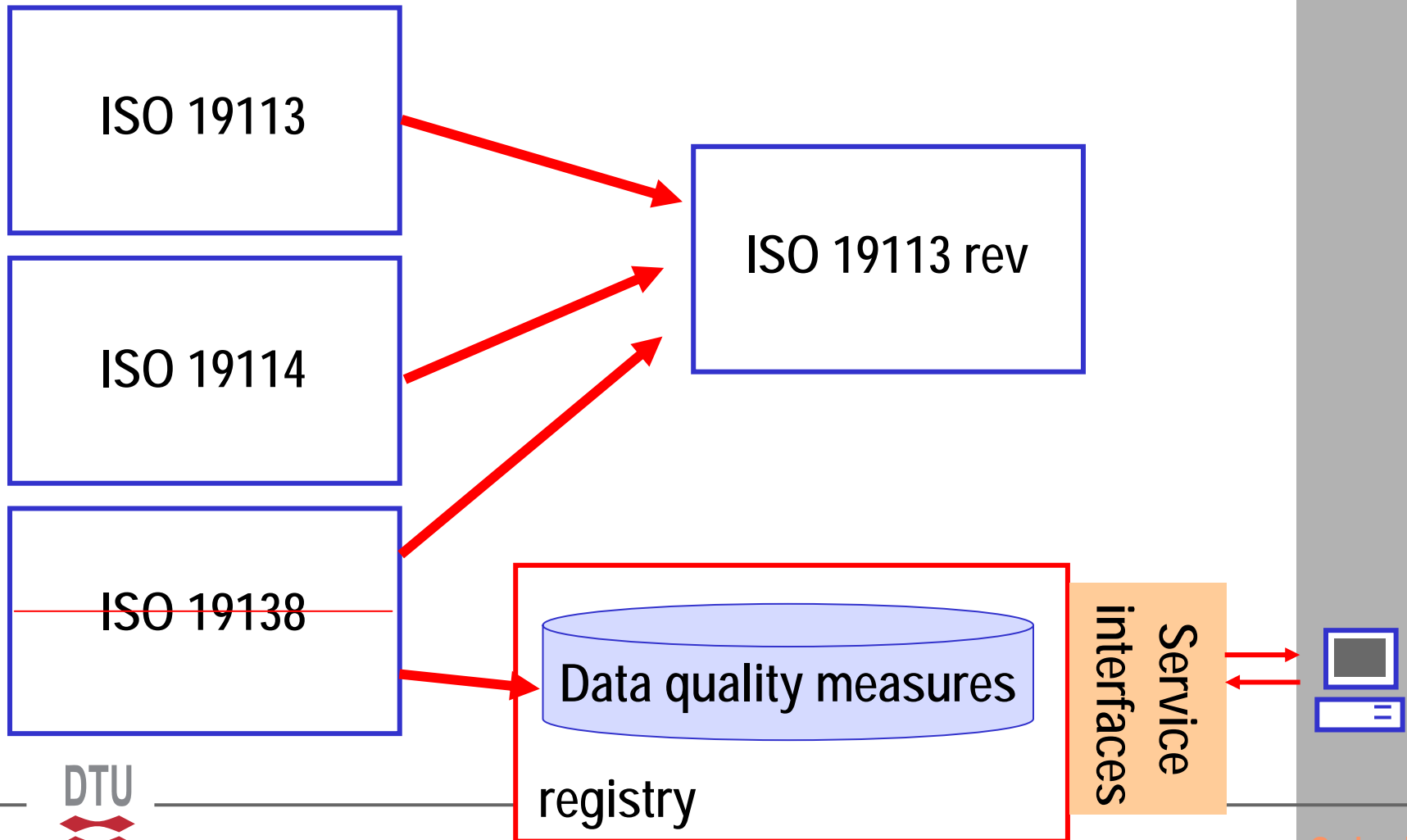
$$P(radius, \sigma_X, \sigma_Y) = \frac{1}{2\pi\sigma_X\sigma_Y} \iint_{(x-x_i)^2 + (y-y_i)^2 = radius^2} e^{-\frac{1}{2} \left( \frac{(x-x_i)^2}{\sigma_X^2} + \frac{(y-y_i)^2}{\sigma_Y^2} \right)} dx dy$$

For some particular probabilities  $P$  the radius can be calculated dependent on the standard deviations  $\sigma_X$  and  $\sigma_Y$ .

Table C.5 — Relationship between the probability  $P$  and the corresponding radius of the circular area

Probability $P$	Data quality basic measure	Name	Data quality value type
$P = 39,4\%$	$\frac{1}{\sqrt{2}} \sqrt{\sigma_X^2 + \sigma_Y^2}$	CE39.4	measure
$P = 50\%$	$\frac{1,1774}{\sqrt{2}} \sqrt{\sigma_X^2 + \sigma_Y^2}$	CE50	measure
$P = 90\%$	$\frac{2,146}{\sqrt{2}} \sqrt{\sigma_X^2 + \sigma_Y^2}$	CE90	measure
$P = 95\%$	$\frac{2,4477}{\sqrt{2}} \sqrt{\sigma_X^2 + \sigma_Y^2}$	CE95	measure
$P = 99,8\%$	$\frac{3,5}{\sqrt{2}} \sqrt{\sigma_X^2 + \sigma_Y^2}$	CE99.8	measure

# Implications for quality in ISO/TC 211



# Conclusions

- Register services make normative elements online accessible
- That produces synergy between standards and services
  - Avoids unnecessary duplication of definitions, descriptions or examples for reporting quality using data quality measures
  - The amount of overhead for quality related metadata would be reduced
- A registry for data quality measures would help users to choose the suitable data quality measure
- Registries shall have online interfaces for humans (HTML) and for GIS-Software (XML)
- Quality evaluation results can be used for evaluation of analysis results (e.g. in decision making)
- Can be embedded in business processes via ebXML

# Questions?

Dr. Gerhard Joos  
Associate professor

Technical University of Denmark  
National Space Institute

Presently located at:  
Informatics and Mathematical Modelling  
Building 321, office 219  
DK-2800 Kongens Lyngby, Denmark

Group            Geoinformatics  
Phone            +45 4525 5984

Email            [gj@space.dtu.dk](mailto:gj@space.dtu.dk)