

# Using metadata quality to filter datasets in data portals

Susanne Feistel<sup>1</sup>, Romina Ihde<sup>4</sup>, Ulrike Kleeberg<sup>2</sup>, Jörn Kohlus<sup>3</sup>, Rainer Lehfeldt<sup>4</sup>, Carsten Schirnack<sup>5</sup>, Stefanie Schumacher<sup>6</sup>, Susanne Tamm<sup>7</sup>

## Good Practice

Quality information may be stored in a practical number of specified recommended fields of the metadata standard. Therefore data portals with specific filter options can increase dissemination and availability of data with documented quality information.

## Creating accessible quality information

### Document your workflow:

1. Purpose of your research
2. Subject of your research
3. Data sampling or sources
4. Geo-reference your data
5. Data processing steps
6. Quality Control
7. Availability of your data

### Put quality information in ISO metadata:

- ➔ MD\_Identification.purpose
- ➔ MD\_Identification.abstract
- ➔ LI\_Source, LI\_Lineage
- ➔ EX\_Extent, MD\_Identification.spatialResolution
- ➔ LI\_ProcessSteps
- ➔ DQ\_QualityElement, DQ\_EvaluationMethod
- ➔ MD\_Distribution, MD\_Constraints

## Creating a quality flag

### Create standardized string as quality flag:

- Batch xml analysis via software
- Manually via web application form
- Manually via web application xml upload

The result: **a comparable alphanumeric string combining quality scheme and flags, e.g. „SDN::1“** for „good quality“ as defined by SeaDataNet.

### Put quality flag in ISO metadata:

DQ\_StandaloneQualityReportInformation (o,n)

**Automatic quality assessment for ISO 191\*\***

Wie wurden die im Metadatensatz beschriebenen Daten erzeugt, erhoben,prozessiert und geprüft:

Datatype	Source			
Diese Auswahl würde die nachfolgenden Felder bestimmen: ISO: LI_Source	<input checked="" type="radio"/> Point measurement / Transect			
	<input type="radio"/> Calculation, Count			
	<input type="radio"/> Modelling data			
	<input type="radio"/> Produkt / Kombination / Schema (Netzanalyse)			
	<input type="radio"/> Foto / Video / Satellit			
	<input type="radio"/> Karte als Darstellung von insitu-Daten			
	<input type="radio"/> Karte als Produkt verschiedener Datensätze			
<input type="radio"/> Experimentelle Daten ohne räumlichen Bezug				
bei Feldern wie Zweck oder Verarbeitung soll eine Gegenprüfung zu Inhalten in ISO-Feldern stattfinden.				
Nach- / Nutzung	Purpose	Nutzungsbedingungen		
erschient immer	<input checked="" type="checkbox"/> Erhebungszweck angegeben <input type="checkbox"/> als Referenzdatensatz geeignet	<input checked="" type="checkbox"/> Lizenz / Nutzungsbedingungen		
ISO: purpose MD_Constraint MD_Distribution				
Extent	Ausdehnung	Auflösung		
erschient immer, beschreibt geogr. & zeitl. Ausdehnung	<input checked="" type="checkbox"/> Coordinates and reference system (geographic) angegeben. <input checked="" type="checkbox"/> Time and reference system (temporal) angegeben. <input checked="" type="checkbox"/> Depth and reference system (vertical) angegeben.	<input checked="" type="checkbox"/> Räumlich <input checked="" type="checkbox"/> Zeitlich		
ISO: EX_Extent				
Sampling	Sampling	Processing	Storage	Measurement
erschient bei Punktmessung, Zählung, Hochrechnung, Foto / Video / Satellit	<input checked="" type="checkbox"/> dokumentiertes Verfahren.	<input checked="" type="checkbox"/> dokumentiertes Verfahren.	<input type="checkbox"/> dokumentiertes Verfahren.	<input checked="" type="checkbox"/> dokumentiertes Verfahren.
ISO: LI_Lineage LI_ProcessStep				
Messdaten	Qualitätsinformationen	Evaluation		
erschient bei Punktmessung, Zählung, Hochrechnung	<input checked="" type="checkbox"/> angegeben. <input type="checkbox"/> Genauigkeit <input type="checkbox"/> Messunsicherheit <input type="checkbox"/> Fehlerangabe	<input checked="" type="checkbox"/> angegeben. <input type="checkbox"/> automatisch / Software <input type="checkbox"/> gegen Referenzdaten validiert <input type="checkbox"/> Expert Review		
ISO: DQ_DataQuality DQ_EvaluationMethod				

**Zusammenfassung**

- Datentyp: point
- Zweck: angegeben /
- Gültigkeit: Nutzungsbedingungen / räumliche Auflösung / zeitliche Auflösung
- Extent: geographische Koordinaten / Zeitpunkt / vertikale Koordinate
- Sampling Dokumentation: Probenahme / Verarbeitung / / Messung
- Verfahren Dokumentation: / /
- Prüfung: / /

**Resultat Qualitätsflag**

- Summary: Metadaten geprüft. Angaben zu Zweck, Extent, Auflösung (zeitlich & räumlich), Methoden, Qualität und Nutzungsbedingungen vorhanden. Resultat: "Probably Good"
- ODV::0
- SDN::1

## Discussion

### Pro:

- applicable for all data types
- minimal quality information
- motivation to document
- self assessment of quality
- objective assessment
- easy implementation

### Contra:

- very generic
- no content analysis possible

### Next:

- data type specific quality information

Documenting data with standardized metadata for quality.

Prototype of the **web application** to assess what information is necessary to provide „good“ metadata for a dataset. Depending on the type of data the web form compiles the given information and displays the probable quality flag. (Example: point data)

**XML-Checker**  
Beta-Version

Drag & Drop Datei hierher, oder Doppel-Klick um Datei auszuwählen

In-situ Sampling

Ergebnisse leeren

Die Qualität der Dokumentation von CW\_SWATH.xml ist unbekannt. ⚠

Um die Qualität zu verbessern geben Sie bitte folgende Felder an: DQ\_EvaluationMethod, MD\_Identification.spatialResolution.

Die Qualität der Dokumentation von insituSampling\_IOW\_20191023.xml ist wahrscheinlich gut. ✓

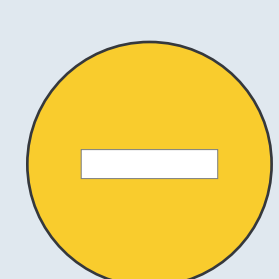
Die Qualität der Dokumentation von Produkte\_HZG\_20191021.xml ist wahrscheinlich gut. ✓

Beta version of the **web application** to check metadata for documentation quality. A drag & drop field allows the **direct upload** of a file in XML-format (ISO 19139). According to the selected datatype, the metadata is then checked for the necessary information (described in the web form to the right and in the table below). A **quality rating** and possibly information about missing elements are the results. (Result examples from NCEI, IOW and HZG)

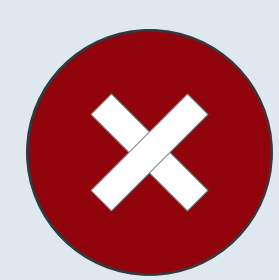
## Outlook: Rating in data portals



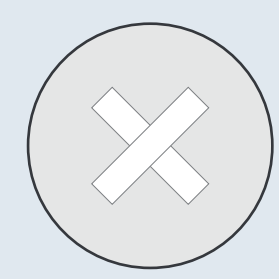
Quality checked: „(Probably) Good“  
„SDN::1“ || „ODV::0“ || „QUARTOD::3“



Quality checked: „Questionable“  
„SDN::3“ || „ODV::4“ || „QUARTOD::2“



Quality checked: „Unknown“  
„SDN::0“ || „ODV::1“ || „QUARTOD::0“



Quality unchecked

ABOUT THE DATA	Good Practice Documentation				ISO 191**	
	Modeling combining multiple source data & run algorithms	Products combining multiple source data & create visualization, maps, etc.	Media photo, audio, video	In-situ Sampling in the broadest sense	obligation	
WHY 1. Purpose of your research	X	X	X	X	1	o MD_Identification.purpose
WHAT 2. Subject of your research 3. Data sources	X	X			n	o LI_Source
WHERE & WHEN 4. Geo-reference your data	X	X	X	X	1	m MD_Identification.abstract
HOW 3. Data sampling 5. Data processing steps 6. Quality Control	X	X	X	X	n	o MD_Identification.descriptiveKeywords
WHO	X	X	X	X	n	o MD_Identification.spatialResolution
AVAILABILITY 7. Availability of your data	X	X	X	X	n	c EX_Extent
					n	o LI_Lineage
					n	c LI_ProcessStep
					n	m DQ_Element
					1	o DQ_EvaluationMethod
					n	o CI_Contact
					n	o MD_Distribution
					n	c MD_Constraint.accessConstraint

<sup>1</sup> susanne.feistel@io-warnemuende.de    <sup>4</sup> rainer.lehfeldt@baw.de  
<sup>2</sup> ulrike.kleeberg@hzg.de    <sup>4</sup> romina.ihde@baw.de  
<sup>3</sup> joern.kohlus@lkn.landsh.de    <sup>5</sup> cschirnack@geomar.de  
<sup>6</sup> stefanie.schumacher@awi.de    <sup>7</sup> susanne.tamm@bsh.de

\*\* ISO 19107, ISO 19111, ISO 19115 (2014), ISO 19139, ISO 19156, ISO 19157

