LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

AtMoDat

TIB

Ein Vorschlag für die Sichtbarmachung der Qualität von Forschungsdaten am Beispiel des DataCite DOI

Anette Ganske[1], Daniel Heydebreck[2], Angelina Kraft[1]

1: Technische Informationsbibliothek (TIB)
2: Deutsches Klimarechenzentrum (DKRZ)

#vBiB2020 28.05.2020

Leibniz Gemeinschaft

Agenda

1. Motivation
2. Was ist AtMoDat?
3. Was versteht man unter der Qualität von Datensätzen?
4. Wie wird Qualität geprüft?
5. Wie kann man die Qualitätsprüfung sichtbar machen?
6. Fazit und Ausblick

1. Motivation

Anzahl der Datenpublikationen (incl. DataCite DOIs) wächst, aber….

Daten oft nicht wiederverwendbar:

- Unterschiedliche Datenformate
- Metadaten lückenhafte und/oder ohne disziplin-spezifische Standards
- Keine Information zur Qualität

Dies beeinflusst die gesamte Community:

- Datennutzer: Hilfreiche Daten sind schwer zu finden
- Datenproduzenten und Repositorien:
  Daten werden seltener benutzt, geringerer Bekanntheitsgrad
- Förderinstitutionen: Projektergebnisse nicht wiederverwendbar

2. Was ist AtMoDat?

BMBF-Projekt, Laufzeit 2019 – 2021

Ziel: Verbesserung der Wiederverwendbarkeit von **At**mosphärischen **Mo**dell **Dat**en

Grund: Atmosphärische Modelle produzieren meist sehr viele Daten, die für viele Projekte genutzt werden können

Basis: für CMIP (Controlled Model Intercomparison Project) wurden bereits Standards zur FAIRen Veröffentlichung von Daten entwickelt

→ **Standards auf andere Teildisziplinen der Meteorologie übertragen**

TIB

Universitäre Partner

Infrastruktur Partner

Kernstandard
für die verbesserte Wiederverwendbarkeit von atmosphärischen Modelldaten

Data Maturity Indicator
für DataCite DOIs

Disziplin-spezifische Standards für Modellergebnisse von Teildisziplinen:

- Stadtklimatologie
- Modellvergleichsprojekte (Wolkenphysik)

AtMoDat

3. Was versteht man unter der Qualität von Datensätzen?

Physikalische Konsistenz von Messdaten

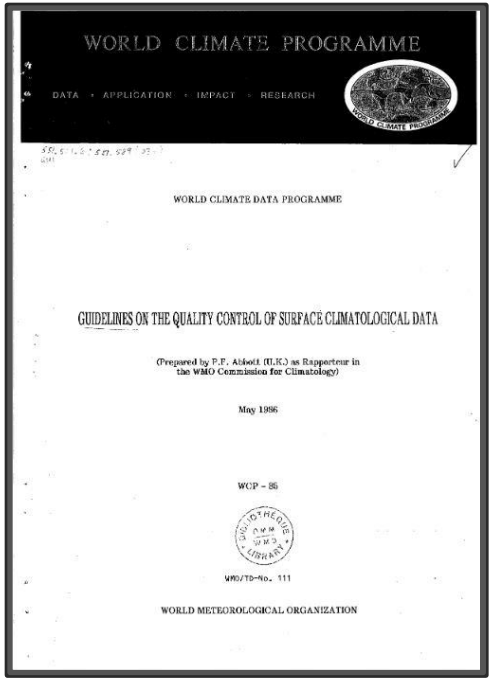
WORLD METEOROLOGICAL ORGANIZATION
WEATHER CLIMATE WATER

user name Password OK
Forgotten password?
Register

TIB

Guidelines on the quality control of surface climatological data
WMO/TD- No. 111; WCP- No. 85

This paper describes various methods which can be used for the quality control of meteorological data. A range of methods are covered, from very basic checks for correct coding, to time and areal consistency.


WORLD CLIMATE PROGRAMME

DATA • APPLICATION • IMPACT • RESEARCH

WORLD CLIMATE DATA PROGRAMME

GUIDELINES ON THE QUALITY CONTROL OF SURFACE CLIMATOLOGICAL DATA

(Prepared by P.F. Abbott (U.K.) as Rapporteur to the WMO Commission for Climatology)

May 1996

WCP – 85

WMO/TD-No. 111

WORLD METEOROLOGICAL ORGANIZATION

https://library.wmo.int/index.php?lvl=notice_display&id=11700#.Xrz7m0fgqpo

3. Was versteht man unter der Qualität von Datensätzen?

Physikalische Konsistenz von Messdaten

FAIR

FAIR: M. D. Wilkinson et al., 2016.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

https://doi.org/10.1038/sdata.2016.18

3. Was versteht man unter der Qualität von Datensätzen?

Physikalische Konsistenz von Messdaten

FAIR

Data Maturity

NOAA: Data Maturity Matrix

Maturity Scale / Key Component	Level 1 - Ad Hoc Not Managed	Level 2 - Minimal Managed Limited	Level 3 - Intermediate Managed Defined, Partially Implemented	Level 4 - Advanced Managed Well-Defined, Fully Implemented	Level 5 - Optimal Level 4 + Measured , Controlled , Audit
Preservability	*The state of dataset being preservable*				
Accessibility	*The state of dataset being publicly searchable and accessible*				
Usability	*The state of data product being easy to understand and use*				
Production Sustainability	*The state of data production being sustainable and extendable*				
Data Quality Assurance	*The state of data product quality being assured/screened*				
Data Quality Control /Monitoring	*The state of data product quality being controlled and monitored*				
Data Quality Assessment	*The state of data product quality being assessed*				
Transparency /Traceability	*The state of data product being transparent, trackable, and traceable*				
Data Integrity	*The state of data integrity being verifiable*				

Aus: N. A. Ritchey: NOAA/NCEI‘s Challenges in Meeting New Open Data.
https://presentations.copernicus.org/EGU2020/EGU2020-12419_presentation.pdf

3. Was versteht man unter der Qualität von Datensätzen?

Physikalische Konsistenz von Messdaten

FAIR

Data Maturity

.....

→ Beurteilung der Qualität hängt von der Anwendung ab

AtMoDat zielt auf FAIRness

Standardisierung der Dateien, basierend auf **CMIP6 Standard**

selbstbeschreibendes Dateiformat netCDF

Konventionen für Metadaten: CF Conventions

Anforderungen an Metadaten auf den menschen- und maschinenlesbaren **Landing Pages** der Datensätze

Projekt Bericht
https://www.atmodat.de/p/kernstandard

Veröffentlichung (Ganske, in prep.)

4. Wie wird Qualität geprüft?

Bsp.: FAIR Assessment tools

1	ANDS-NECTAR-RDS-FAIR data assessment tool	ARDC
2	DANS-Fairdat	DANS
3	DANS-Fair enough?	DANS
4	The CSIRO 5-star Data Rating tool	CSIRO
5	FAIR Metrics Questionnaire	The FAIR Metrics Group
6	Stewardship Maturity Mix	NOAA's CICS-NC, NOAA's NCDC
7	FAIR Evaluator	GO FAIR, LUMC CBGP, IDS, RDA FAIRsharing, IQSS
8	Data Stewardship Wizard	ELIXIR NL/CZ
9	Checklist for Evaluation of Dataset Fitness for Use	Assessment of Data Fitness for Use WG (WDS/RDA)
10	RDA-SHARC Evaluation	SHARC IG (RDA)
11	WMO-Wide Stewardship Maturity Matrix for Climate Data	The SMM-CD WG
12	Data Use and Services Maturity Matrix	The MM-Serv WG

https://www.rd-alliance.org/system/files/Results%20Analysis%20of%20FAIR%20assessment%20tools%20v3.pdf

FAIR self-assessment tool

siehe https://www.ands-nectar-rds.org.au/fair-tool

Findable

Does the dataset have any identifiers assigned? Globally Unique, citable and persistent (e.g. DOI, PURL, ARK o

Is the dataset identifier included in all metadata records/files describing the data? No

How is the data described with metadata? The data is not described

What type of repository or registry is the metadata record in? The data is not described in any repository

Accessible

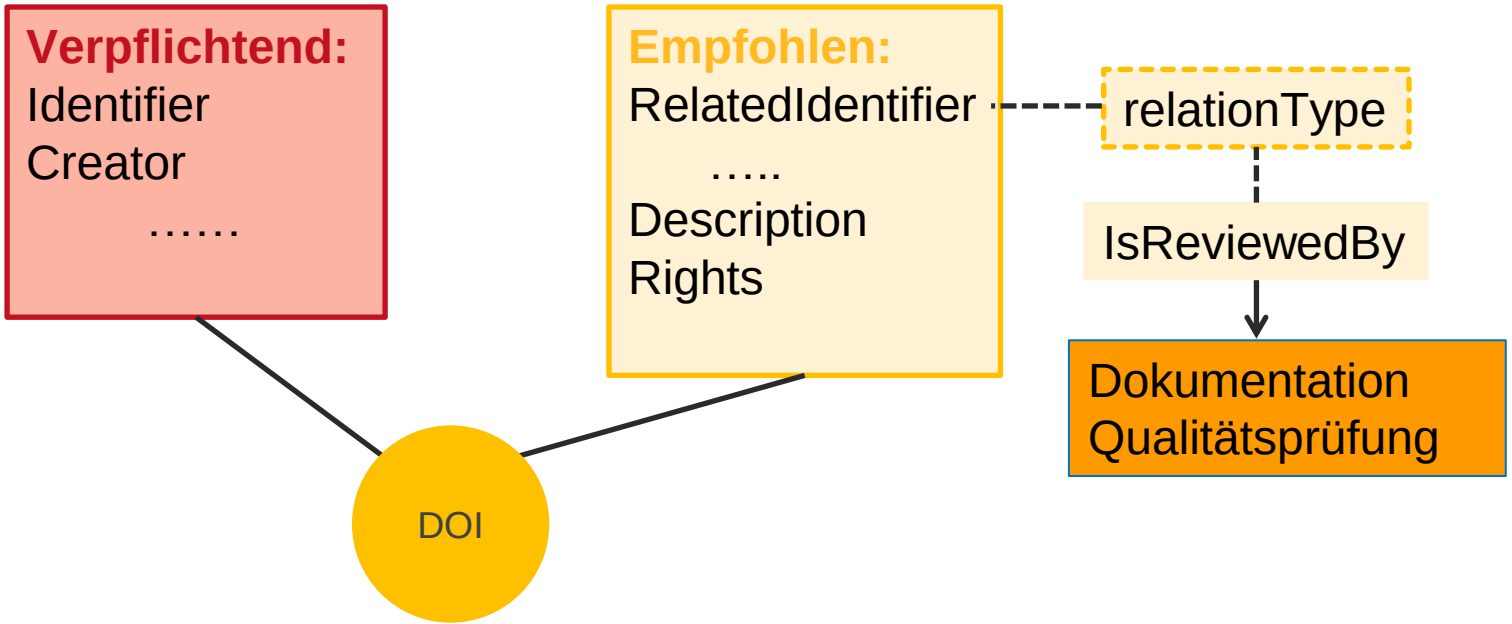
Interoperable

Reusable

Total across F.A.I.R

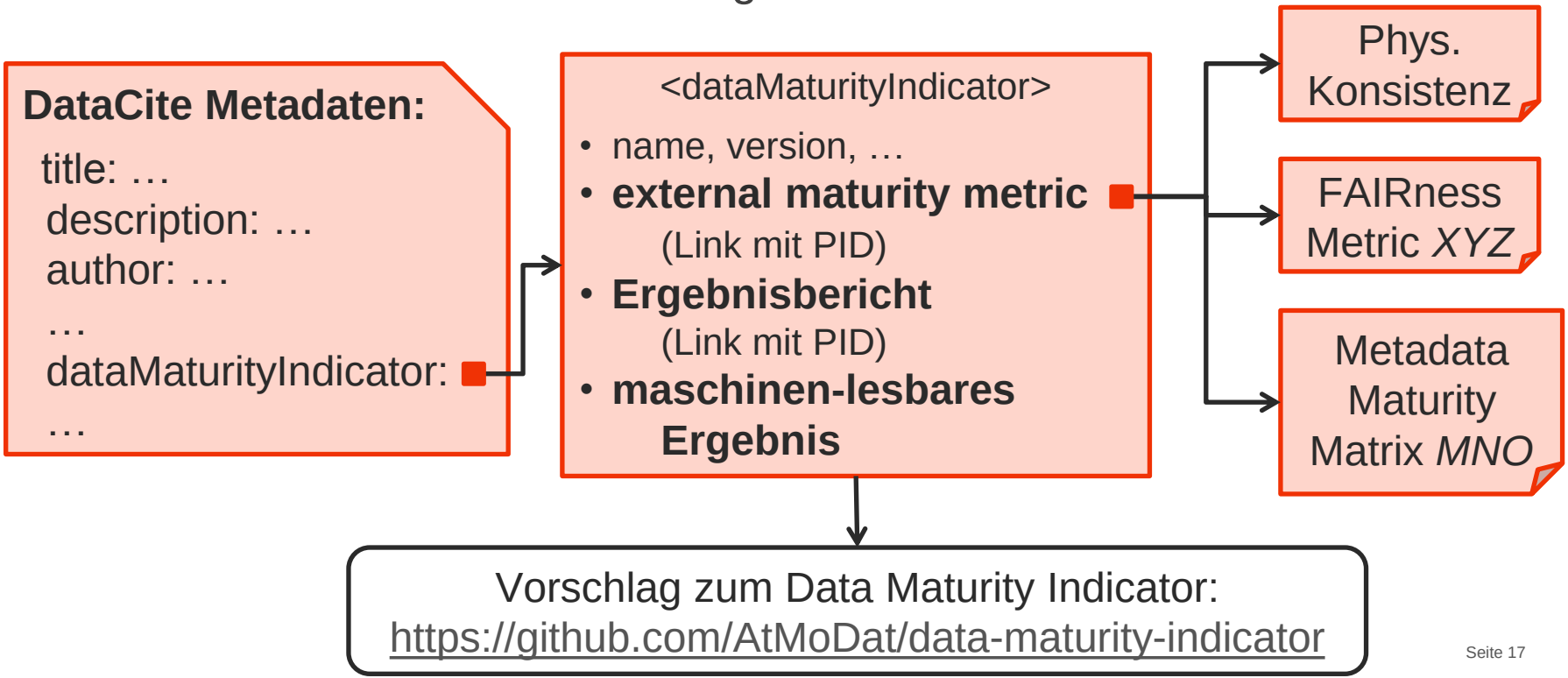
4. Wie kann man die Qualitätsprüfung sichtbar machen?

a.) heutiges DataCite Metadaten Schema



4b.) Vorschlag: Data Maturity Indicator

Idee: zum **DataCite Metadaten Schema** eine Property mit Informationen zur Datenreife hinzufügen



6. Fazit und Ausblick

- Datenqualität ist für die Wiederverwendbarkeit von Daten wichtig
  → sie sollte einfach aus den Metadaten abzulesen sein
- Allerdings: Datenqualität hat viele Facetten
  → viele Möglichkeiten, Qualität zu definieren
- Datenqualität sollte möglichst automatisiert überprüfbar sein
  → wie z.B. bei FAIR Analysetools
- Datenqualität sollte in den Metadaten des DOI sichtbar sein
  → **Data Maturity Indicator**

6. Fazit und Ausblick

Diskussion mit DataCite Community über Data Maturity Indicator:

- Auf Tagungen
- Bei einem Webinar (Termin steht noch nicht fest)
- Mit Ihnen: sagen Sie uns, was sie davon halten –

  jetzt in der Diskussionsrunde oder

  schreiben Sie uns!

LEIBNIZ-INFORMATIONSZENTRUM
TECHNIK UND NATURWISSENSCHAFTEN
UNIVERSITÄTSBIBLIOTHEK

TIB

Vielen Dank!

https://www.atmodat.de/

Anette.Ganske@tib.eu
https://orcid.org/0000-0003-1043-4964
Daniel.Heydebreck@dkrz.de
https://orcid.org/0000-0001-8574-9093
Angelina.Kraft@tib.eu
https://orcid.org/0000-0002-6454-335X

Leibniz Gemeinschaft