

# Datenqualität: Anwendungsempfehlungen für Data Repositories

Thomas Schrader

Fachhochschule Brandenburg, Fachbereich Informatik und Medien

May 23, 2014



# Outline

- 1 Einleitung
  - Data Repositories
  - Daten, Datenqualität & eine Ontologie
- 2 TMF-Datenqualität
  - Ausgewählte Indikatoren
  - Untersuchungen zur Datenqualität
- 3 Diskussion

# Outline

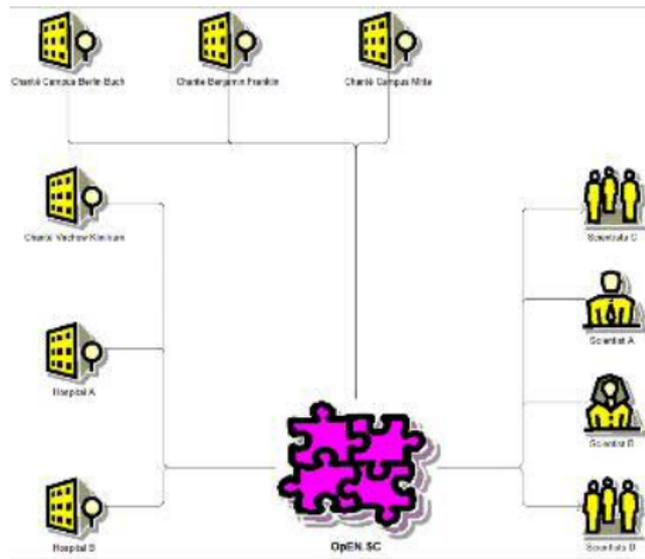
- 1 Einleitung
  - Data Repositories
  - Daten, Datenqualität & eine Ontologie
- 2 TMF-Datenqualität
  - Ausgewählte Indikatoren
  - Untersuchungen zur Datenqualität
- 3 Diskussion

# Data Repository

## Definition

Ein Dienst, der Daten aus unterschiedlichen Quellen für Forschungszwecke zugänglich macht.

[Nonnemacher et. al: Datenqualität in der medizinischen Forschung Leitlinie zum adaptiven Management von Datenqualität in Kohortenstudien und Registern. Version 2.0, TMF, 2014]



# Eigenschaften von Data Repositories

- Heterogenität bezüglich
  - Quellen
  - Datenformate
  - Datenstrukturen
  - Dateninhalte
- kein einheitlicher Datensatz aus allen Quellen
- Unabhängigkeit von der Forschungsfrage

# Definitionen von Datenqualität und Forschungsdaten

## Definition

Eigenschaften von Daten in Bezug auf ihre Eignung, festgelegte Anforderungen zu erfüllen.

[DIN EN ISO 14050 2010]

**ISO 9000** “Degree to which a set of inherent characteristics fulfills requirements.“

**Joseph M. Juran** “Fitness for use.“

**American Society for Quality** “A subjective term for which each person has his or her own definition.“

# Aufgaben des Data Quality Assessments in Data Repositories

## Definition

Summe der Eigenschaften eines Objektes oder Systems

## Definition

Güte der Eigenschaften eines Objektes oder Systems

- 1 Beschreibung der Eigenschaften des Daten Repositories
- 2 Erfüllen der Anforderungen an Datenqualität für die wissenschaftliche Verwendung

# Outline

- 1 Einleitung
  - Data Repositories
  - Daten, Datenqualität & eine Ontologie

- 2 TMF-Datenqualität
  - Ausgewählte Indikatoren
  - Untersuchungen zur Datenqualität

- 3 Diskussion

# Image & Data Quality Assurance Ontology

- Dimensionen der Datenqualität und Qualitätsparameter
- Verhältnis von Datenqualität, Lebenszyklus, Verantwortlichkeit und Anforderungen
- Messung von Datenqualität



# Lebenszyklus von Daten

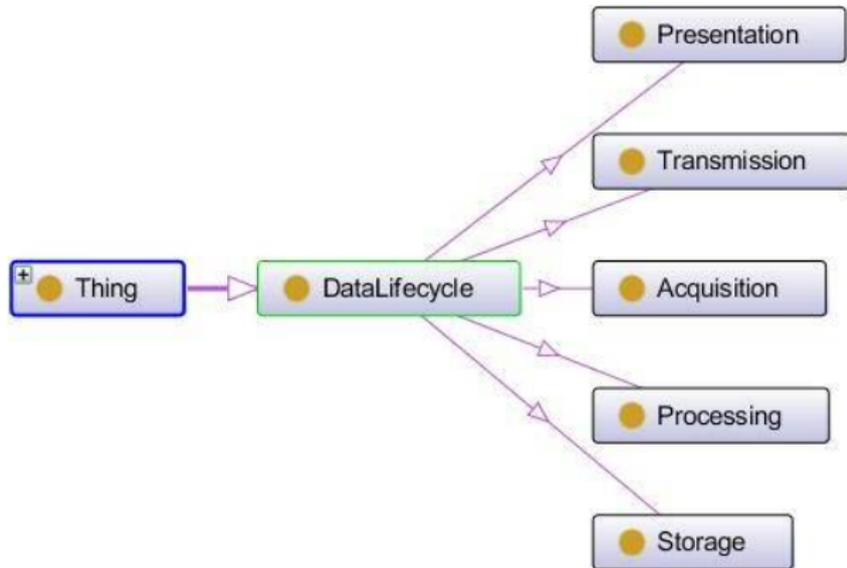
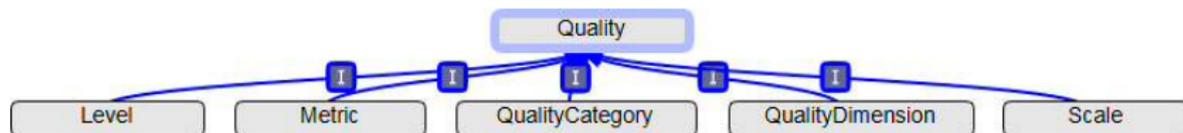
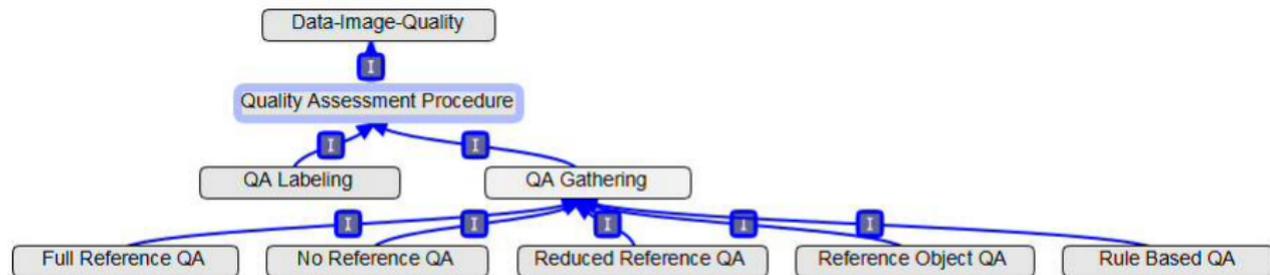


Figure : Ontologieebene Lebenszyklus

# Eigenschaften von Qualitätskriterien



# Strategien des Qualitätsassessments



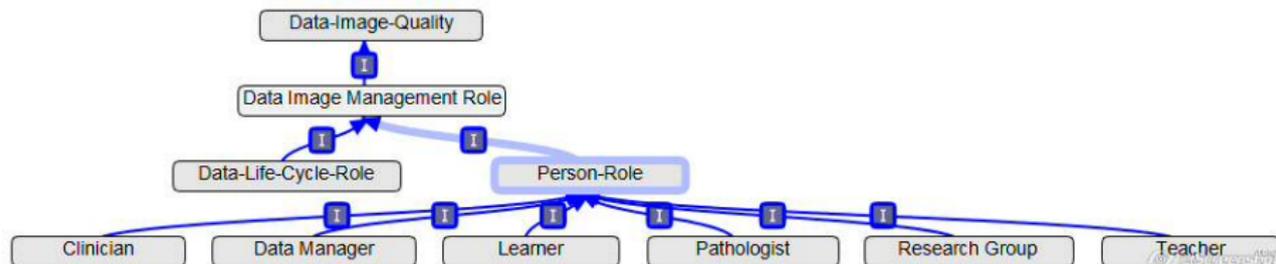
© Jelline

# Qualitätsdimensionen

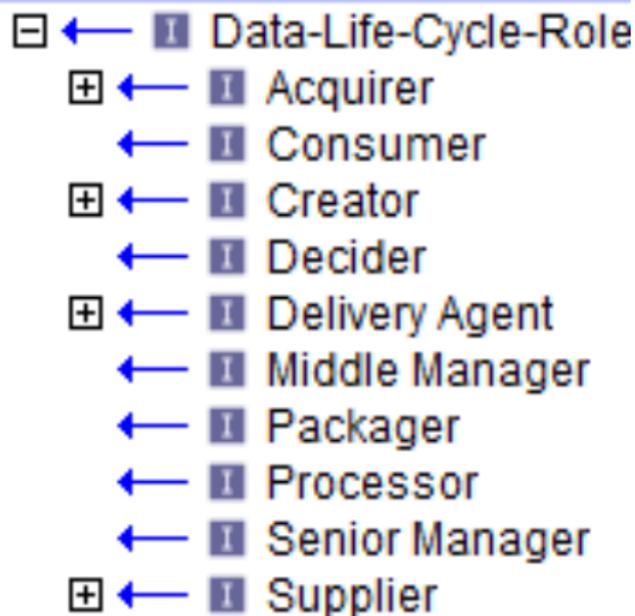
- ☐ ← **1** QualityDimension
  - ← **1** Accessibility
  - ☐ ← **1** Appropriate Amount of Information
  - ☐ ← **1** Believability
  - ☐ ← **1** Completeness
    - ← **1** Concise Representation
  - ☐ ← **1** Consistent Representation
    - ← **1** Ease of Manipulation
  - ☐ ← **1** Free of Error
  - ☐ ← **1** Interpretability
  - ☐ ← **1** Objectivity
    - ← **1** Relevancy
  - ☐ ← **1** Reputation
    - ← **1** Retrieval Quality
  - ☐ ← **1** Security
  - ☐ ← **1** Time Relation
  - ☐ ← **1** Understandability
  - ← **1** Value Added



# Rollen in der Datennutzung: Seite der NutzerIn



# Rollen im Datenmanagement: Repository-Seite



# Outline

- 1 Einleitung
  - Data Repositories
  - Daten, Datenqualität & eine Ontologie
- 2 TMF-Datenqualität
  - **Ausgewählte Indikatoren**
  - Untersuchungen zur Datenqualität
- 3 Diskussion

# TMF: Einordnung der Qualitätsindikatoren

- 1 Integrität - 30 Indikatoren
- 2 Organisation - 15 Indikatoren
- 3 Richtigkeit - 6 Indikatoren

# TMF-1020 Werte aus Standards

**TMF-Zuordnung** Integrität

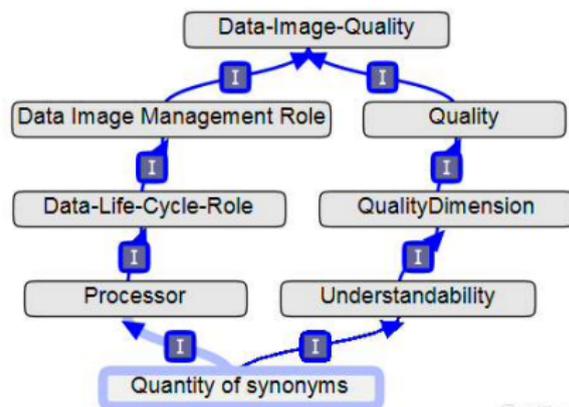
**IDQA-Ontology** Understandability

**Berechnung** Zähler: Anzahl von Werten mit Bezeichnungen aus kontrollierten Vokabularen; Nenner: Anzahl überprüfter Werte

**Probleme** geeignete Klassifikation bzw. Terminologie für die medizinische Domäne, richtiger Einsatz der Klassifikation, richtige Kodierung

# TMF-1036: Anzahl Synonyme

TMF Organisation  
 IDQA Understandability  
 Problem Datenverarbeitung



*© Microsoft*

# Problem: Vollständigkeit

- ☐ ← **1** Completeness
  - ☐ ← **1** Column Completeness
    - ☐ ← **1** Missing values in data elements
      - ← **1** Missing values in mandatory data elements
      - ← **1** Missing values in optional data elements
    - ← **1** Rate of refused data elements
  - ← **1** Metadata Completeness
  - ← **1** Missing module
  - ← **1** Object Completeness
- ☐ ← **1** Population Completeness
  - ← **1** Rate of death certificate only
  - ← **1** Rate of preterm retired observation units
  - ← **1** Rate of recruitment
  - ← **1** Rate of refused investigations
  - ← **1** Rate of refused modules
- ☐ ← **1** Resource completeness
  - ← **1** Count of data resources per observation unit
  - ← **1** Count of observation units with follow up
  - ← **1** Rejected reports
  - ← **1** Single report of a pathologist
  - ← **1** Single report per observation unit
- ← **1** Schema Completeness



# Glaubwürdigkeit

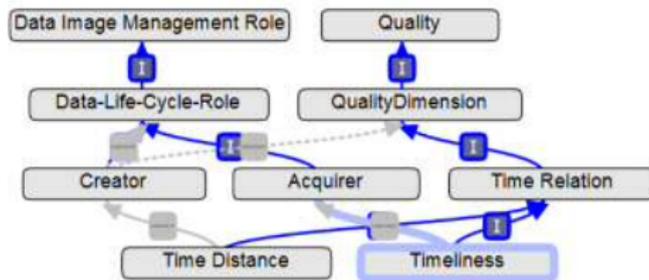
TMF-1007 Bevorzugung bestimmter Endziffern

TMF-1016 Anteil von Datenelementen mit dem Wert unbekannt o.ä.

TMF-1025 Datenelemente mit unspezifischen Werten

# TMF-1028: Aktualität der gespeicherten Daten

TMF Integrität  
 IDQA Timerelation  
 → Timelines



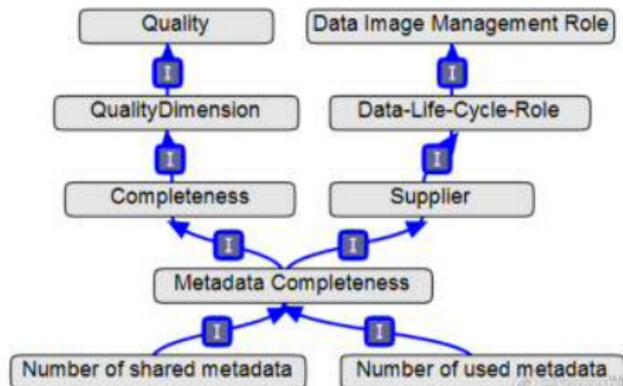
© paltrow/ISTE

# Outline

- 1 Einleitung
  - Data Repositories
  - Daten, Datenqualität & eine Ontologie
- 2 **TMF-Datenqualität**
  - Ausgewählte Indikatoren
  - **Untersuchungen zur Datenqualität**
- 3 Diskussion

# TMF-1050: Anteil der von Untersuchungen übermittelten Metadaten

TMF Integrität  
 IDQA Completeness,  
 Interpretability



# Metadaten: Number of used metadata

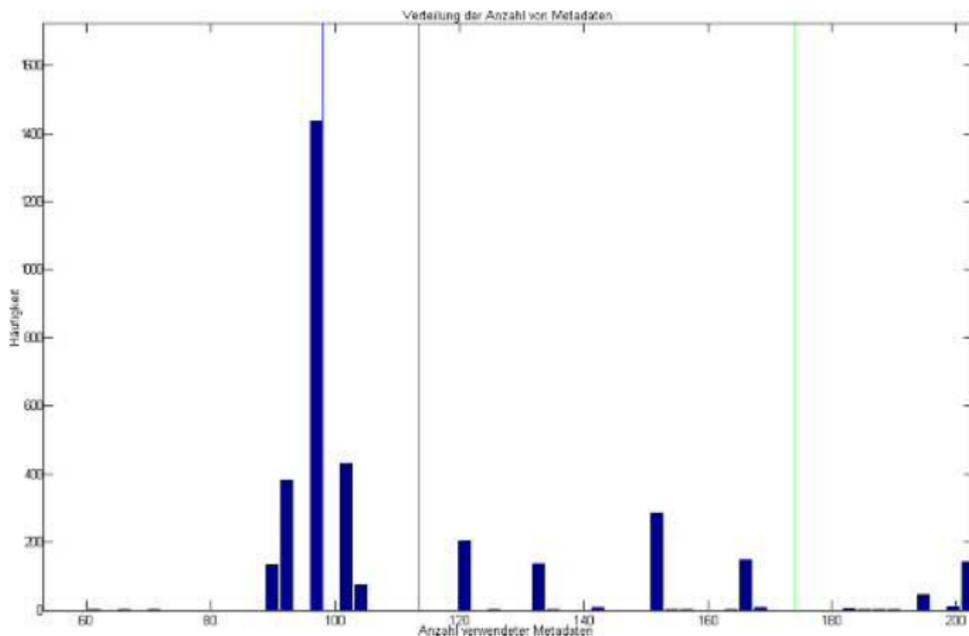


Figure : Anzahl der Metadaten in 3461 DICOM-Bildern

# Metadaten: Number of shared metadata

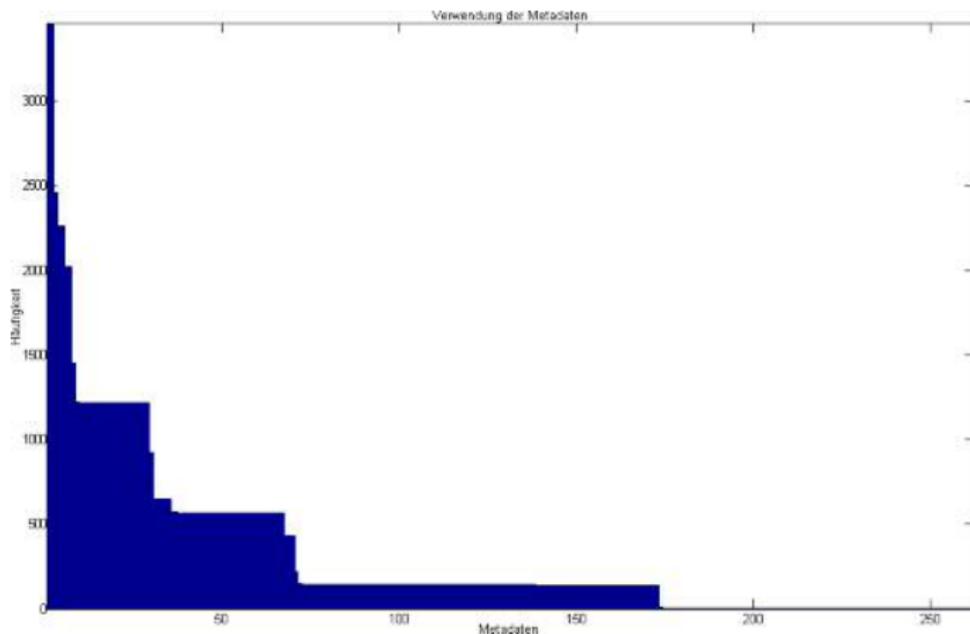


Figure : 5 von allen genutzte Metadaten

# Datenqualität

- TMF-Leitlinie - wichtiger Schritt in der Bewertung der Datenqualität von Registern
- Kontinuierlicher QM Prozess einerseits - Kontinuierliche wissenschaftliche Weiterentwicklung von Datenqualität andererseits
- Problem: Big Data
- Verantwortung für **alle** im Datenmanagement beteiligten Personen

# Image and Data Quality Assessment Ontology

BioPortal [Browse](#) [Search](#) [Mappings](#) [Recommender](#) [Annotator](#) [Resource Index](#) [Projects](#)

## Image and Data Quality Assessment Ontology

[Summary](#) [Classes](#) [Notes](#) [Mappings](#) [Widgets](#) [Get ontology information](#) [Add submission](#)

Jump To:

- Data-Image-Quality
  - Data-Image-Life-cycle
    - Acquisition
    - Processing
    - Storage
    - Transmission
  - Data-Image-Management-Role
    - Data-Life-Cycle-Role
    - Person-Role
  - Data-Type
    - Clinical-Data
    - Image-Data
  - Quality
    - Level
    - Matrix
    - Quality-Category
    - Quality-Dimension
      - Scale
    - Quality-Assessment-Procedure
      - QA-Labeling
      - QA-Labeling

Property	Value
Preferred Name	Data-Image-Quality
Definition	Recreated by Thomas Schrader - 2013-08-24. Ontology for representation of data and image quality properties and assessment methods.
ID	<a href="http://purl.bioontology.org/obo/ID_Q000000">http://purl.bioontology.org/obo/ID_Q000000</a>
comment	Recreated by Thomas Schrader - 2013-08-24.
creation_date	2013-08-24T11:06:56Z
id	Ontology for representation of data and image quality properties and assessment methods.
label	Data-Image-Quality
inversion	ID-Q000000
prefixLabel	Data-Image-Quality
subClassOf	<a href="http://www.e3.ubc.ca/2002/07/owl#Thing">http://www.e3.ubc.ca/2002/07/owl#Thing</a>

The National Center for Biomedical Ontology is one of the National Centers for Biomedical Computing supported by the NIH/NIH, the NIH/NIH, and the NIH Common Fund under grant USA-MH000034. Copyright © 2005-2014. The Board of Trustees of Iceland State University. All rights reserved.

[NCBO Website](#) [Release Notes](#) [Terms of Use](#) [Privacy Policy](#) [How to Use](#)



Figure : <http://bioportal.bioontology.org/ontologies/IDQA>

Vielen Dank für die Aufmerksamkeit!

# Haben Sie Fragen?

Kontakt

Thomas.Schrader@computer.org