Datenqualität: Anwendungsempfehlungen für Data Repositories

Thomas Schrader

Fachhochschule Brandenburg, Fachbereich Informatik und Medien

May 23, 2014

Outline

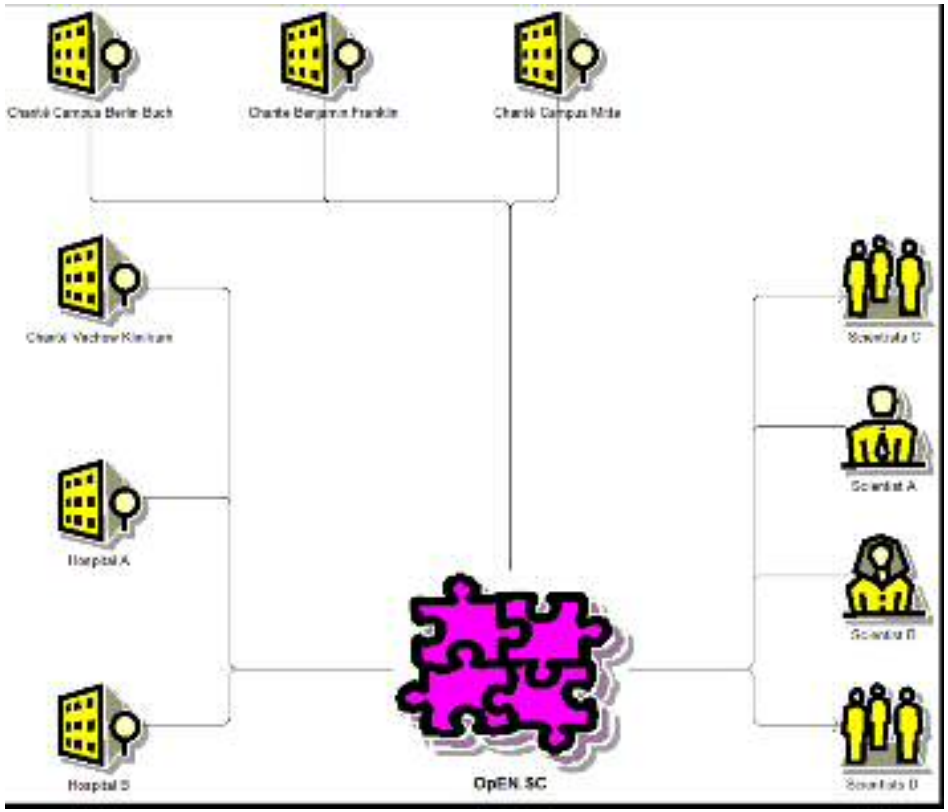
Outline

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 Respositories

- 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

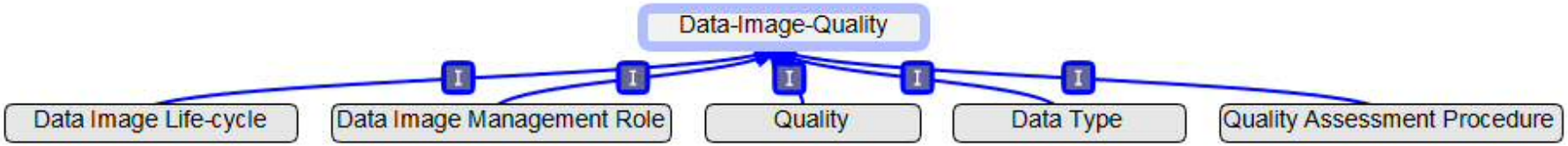
Güte der Eigenschaften eines Objektes oder Systems

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

Outline

Image & Data Quality Assurance Ontology

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



Lebenszyklus von Daten

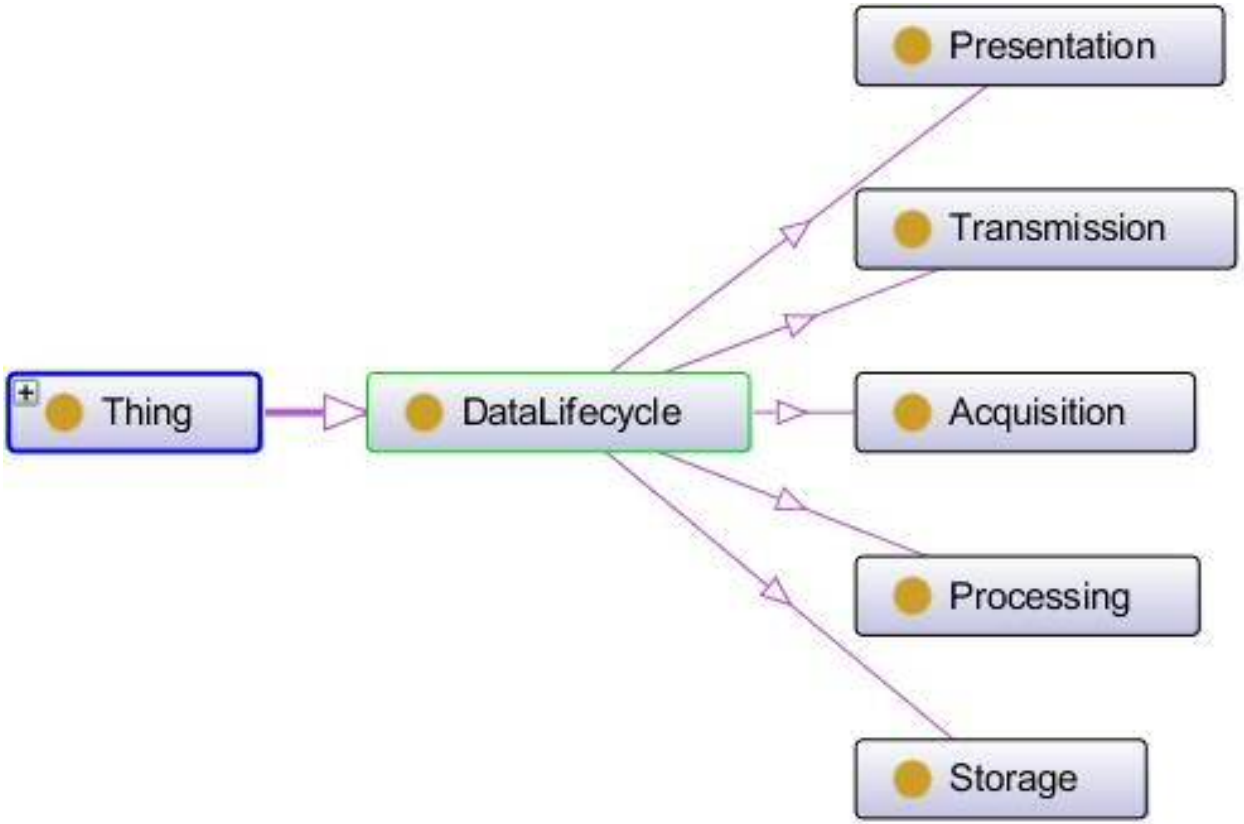
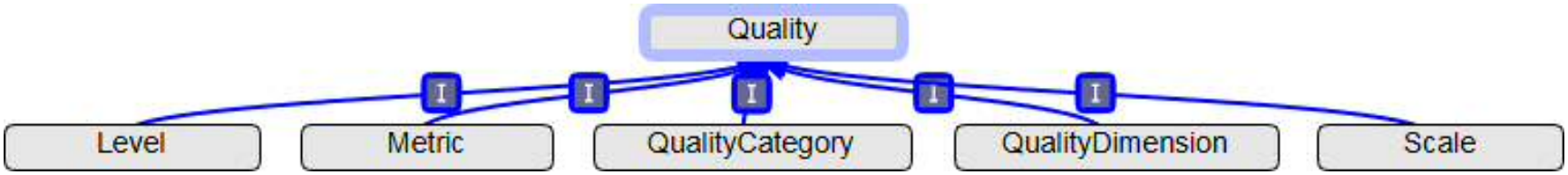
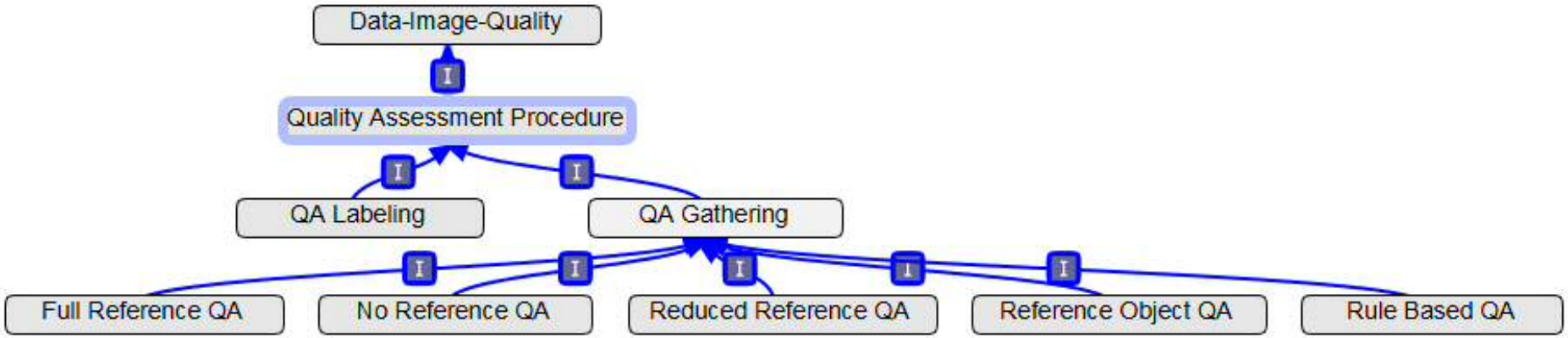


Figure : Ontologieebene Lebenszyklus

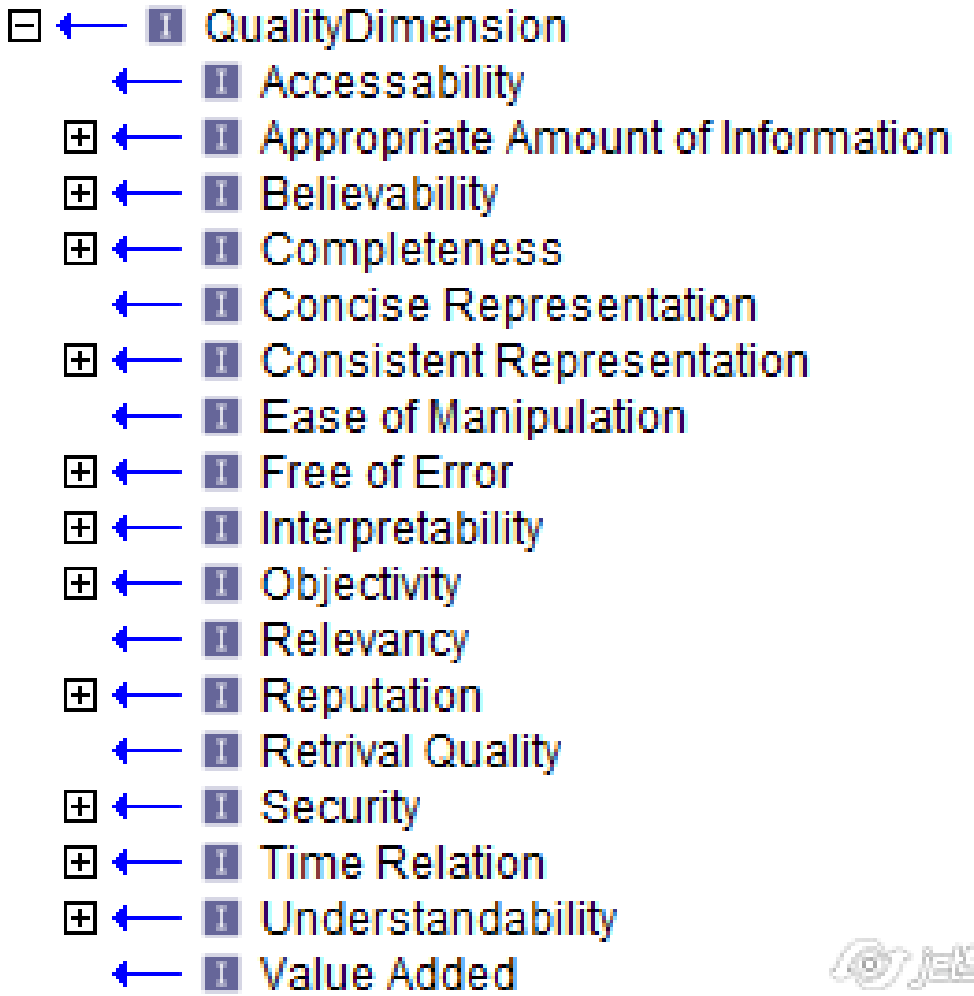
Eigenschaften von Qualitätskriterien



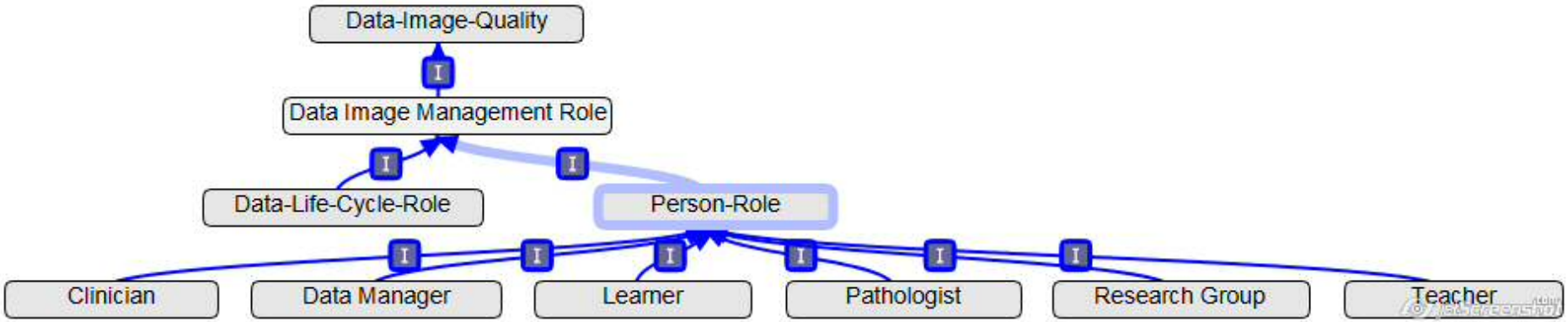
Strategien des Qualitätsassessments



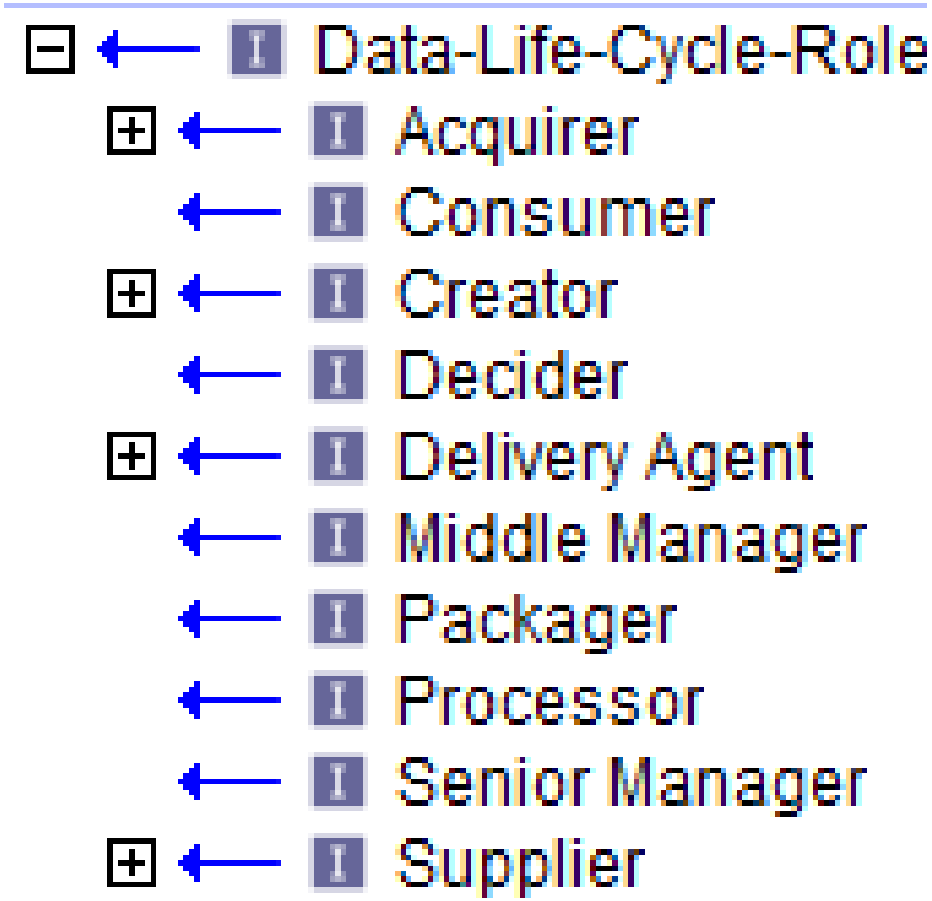
Qualitätsdimensionen



Rollen in der Datennutzung: Seite der NutzerIn



Rollen im Datenmanagement: Repository-Seite



Outline

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

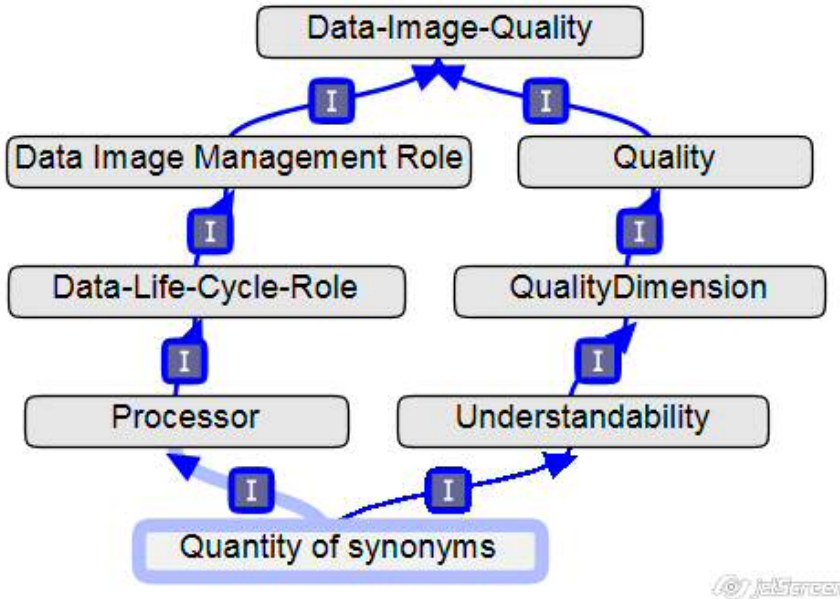
Probleme geeignete Klassifikation bzw. Terminologie für die medizinische Domäne, richtiger Einsatz der Klassifikation, richtige Kodierung

TMF-1036: Anzahl Synonyme

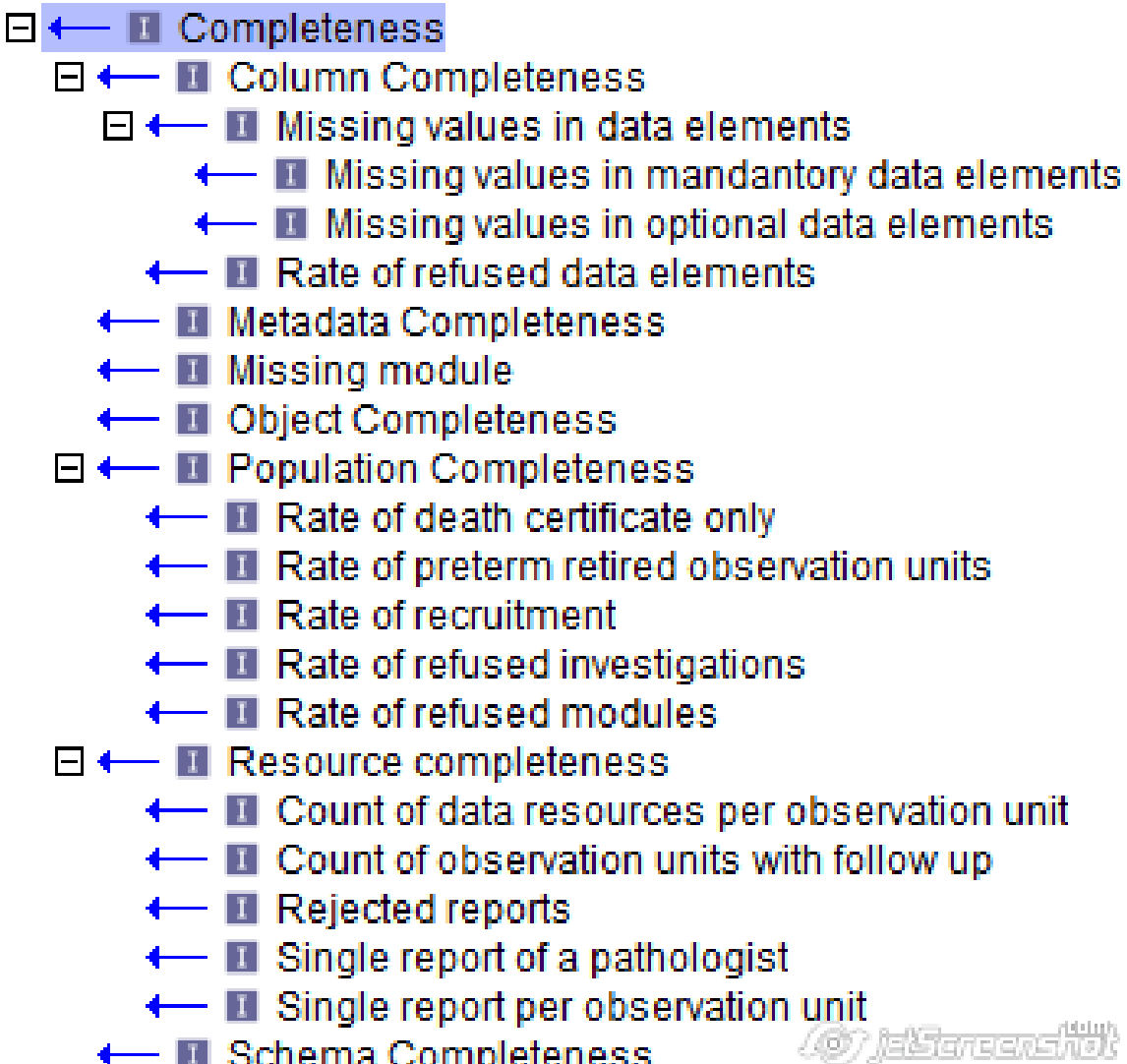
TMF Organisation

IDQA Understandability

Problem Datenverarbeitung



Problem: Vollständigkeit



Glaubwürdigkeit

TMF-1007 Bevorzugung bestimmter Endziffern

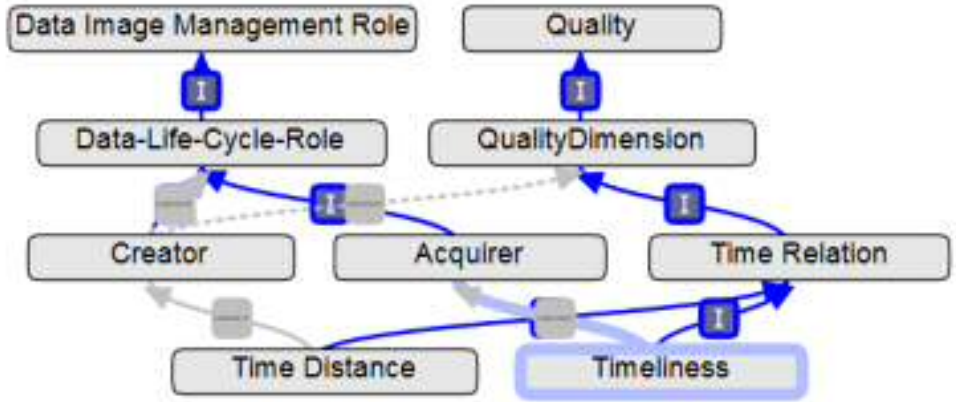
TMF-1016 Anteil von Datenelementen mit dem Wert unbekannt o.ä.

TMF-1025 Datenelemente mit unspezifischen Werten

TMF-1028: Aktualität der gespeicherten Daten

TMF Integrität

IDQA Timerelation
→ Timelines

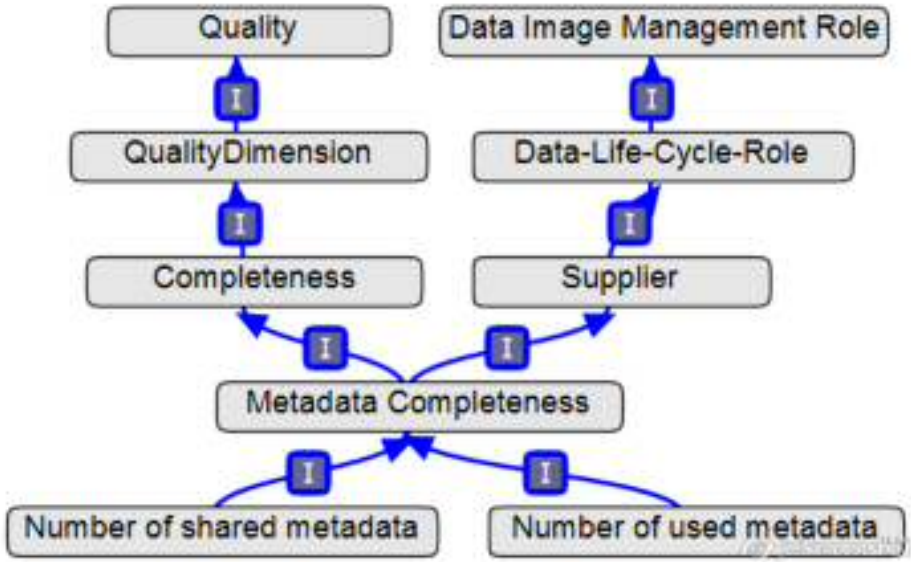


Outline

TMF-1050: Anteil der von Untersuchungen übermittelten Metadaten

TMF Integrität

IDQA Completeness, Interpretability



Metadaten: Number of used metadata

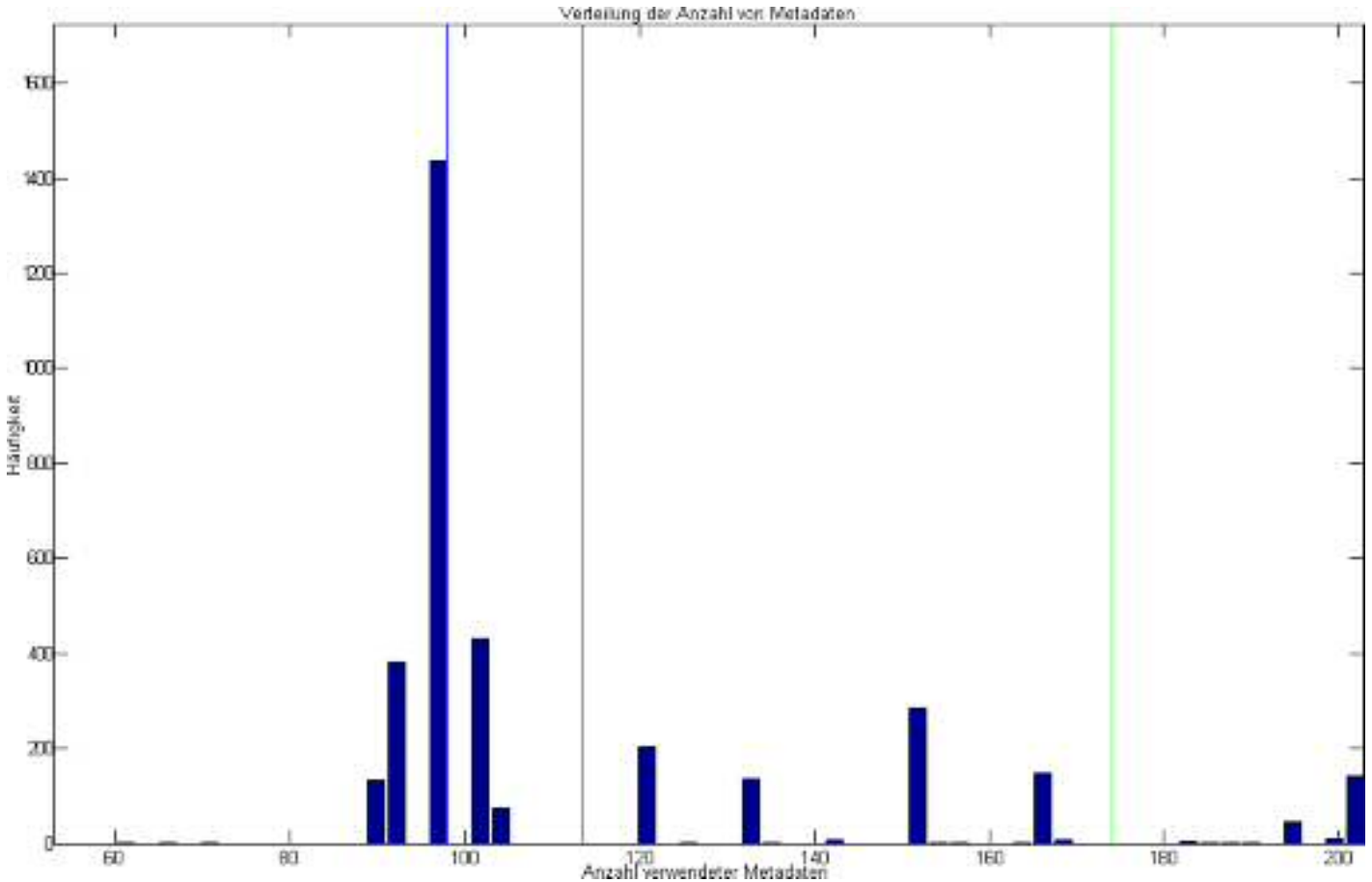


Figure : Anzahl der Metadaten in 3461 DICOM-Bildern

Metadaten: Number of shared metadata

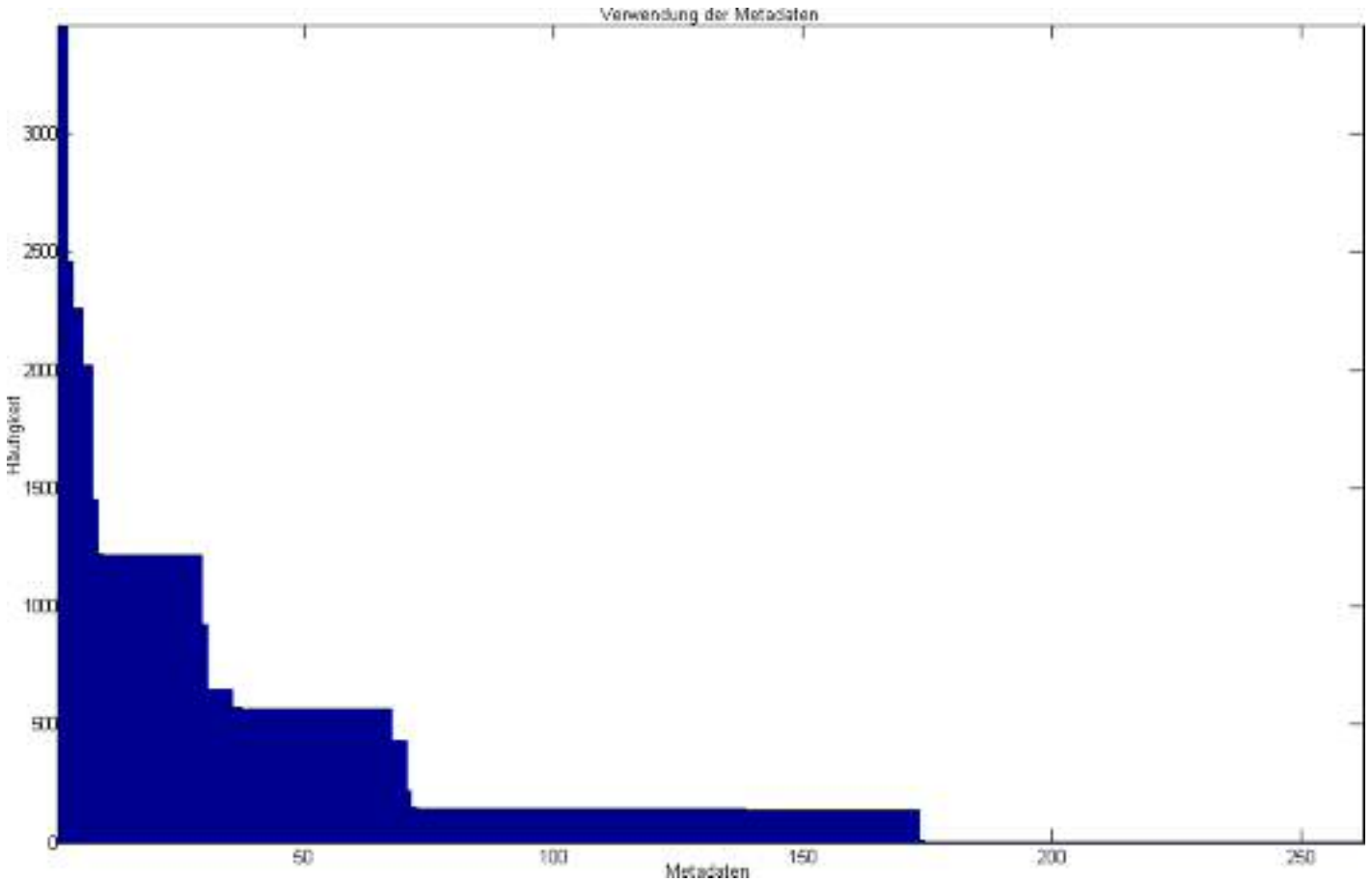


Figure : 5 von allen genutzte Metadaten

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 Ontolgy

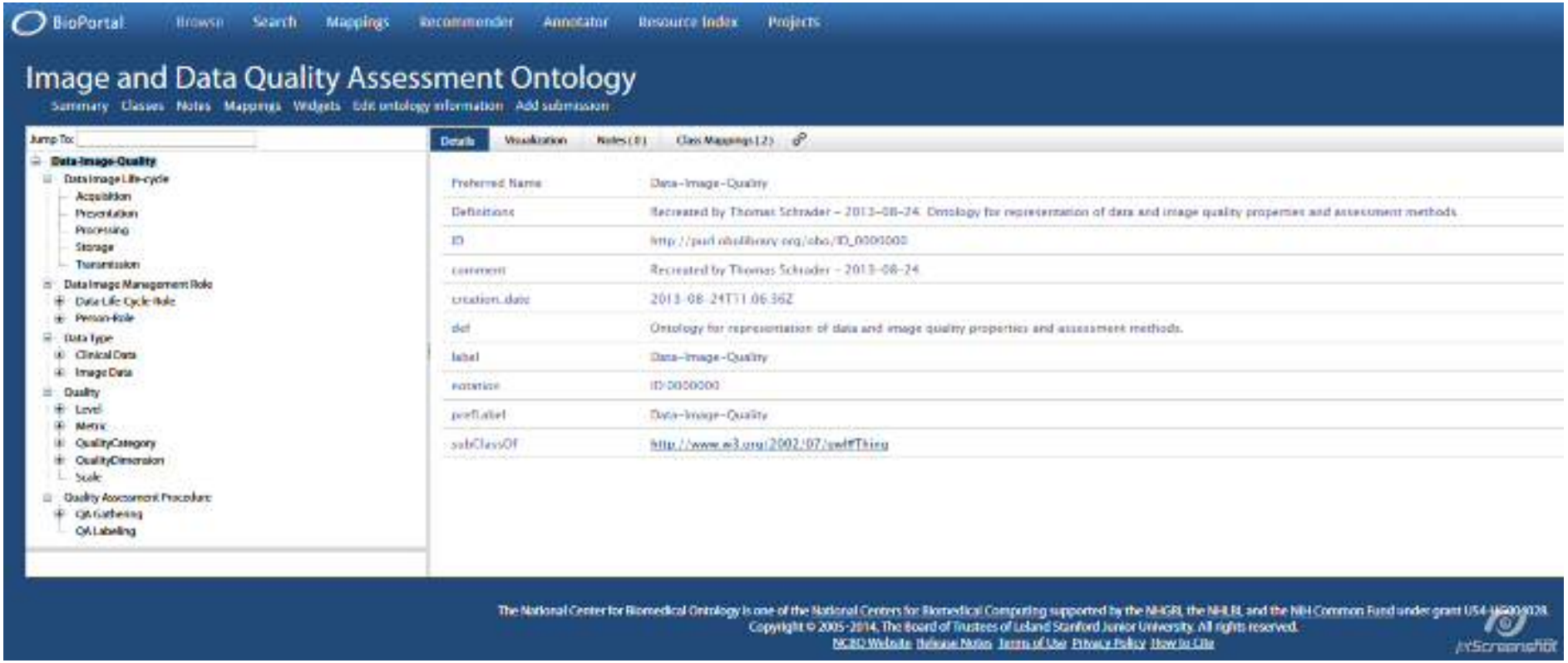


Figure : http://bioportal.bioontology.org/ontologies/IDQA

Vielen Dank für die Aufmerksamkeit!

Haben Sie Fragen?

Kontakt

Thomas.Schrader@computer.org