OpEN.SC

Messung von Datenqualität mit Kennzahlen in Open.SC

Hanß Sabine[1], Niepage Sonja[2], Schrader Thomas[3]
1)Institut für Medizinische Informatik, Charité Universitätsmedizin Berlin
2)Institut für Pathologie, Charité Universitätsmedizin Berlin
3)Fachbereich Informatik und Medien, Fachhochschule Brandenburg

CHARITÉ

OpEN.SC

Agenda

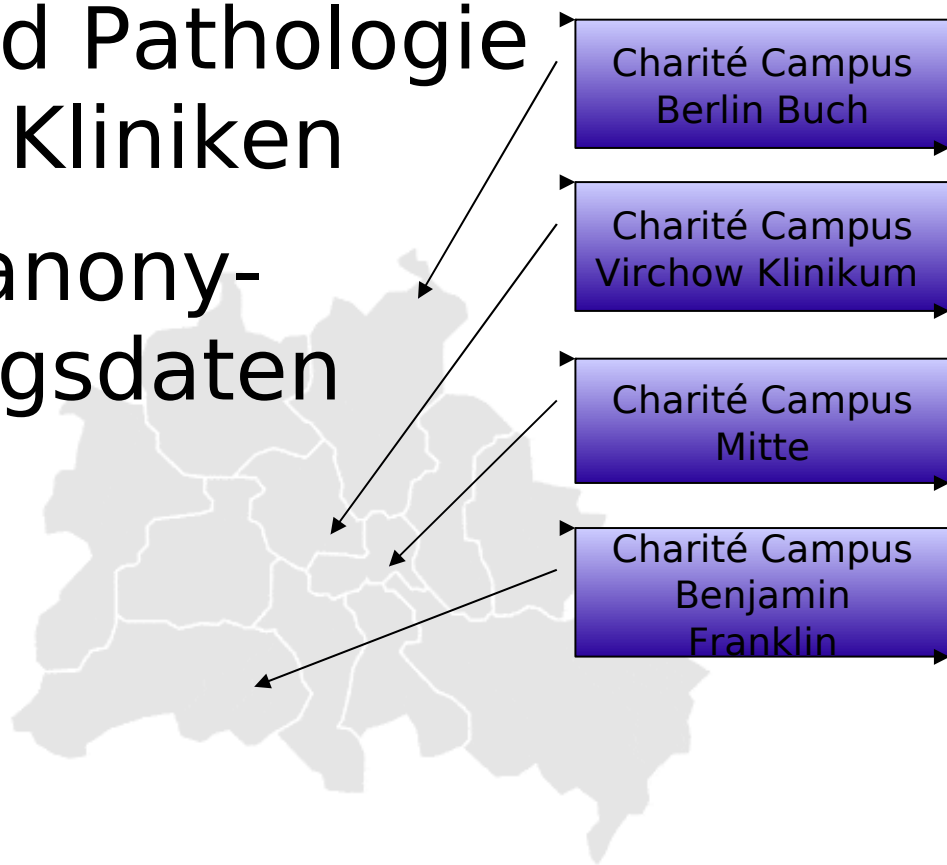
- Projekt OpEN.SC
- Ontologie für Datenqualität
- Framework zur Bestimmung der Anforderungen
- Zwei Beispiele
- Diskussion & Ausblick

CHARITÉ

OpEN.SC

Open European Nephrology Science Center

- Leistungszentrum für Forschungs-information
- Integration von Primärdaten der Nephrologie und Pathologie aus verschiedenen Kliniken
- Bereitstellung der anony-misierten Forschungsdaten über das Internet



CHARITÉ

OpEN.SC

Was ist Datenqualität?

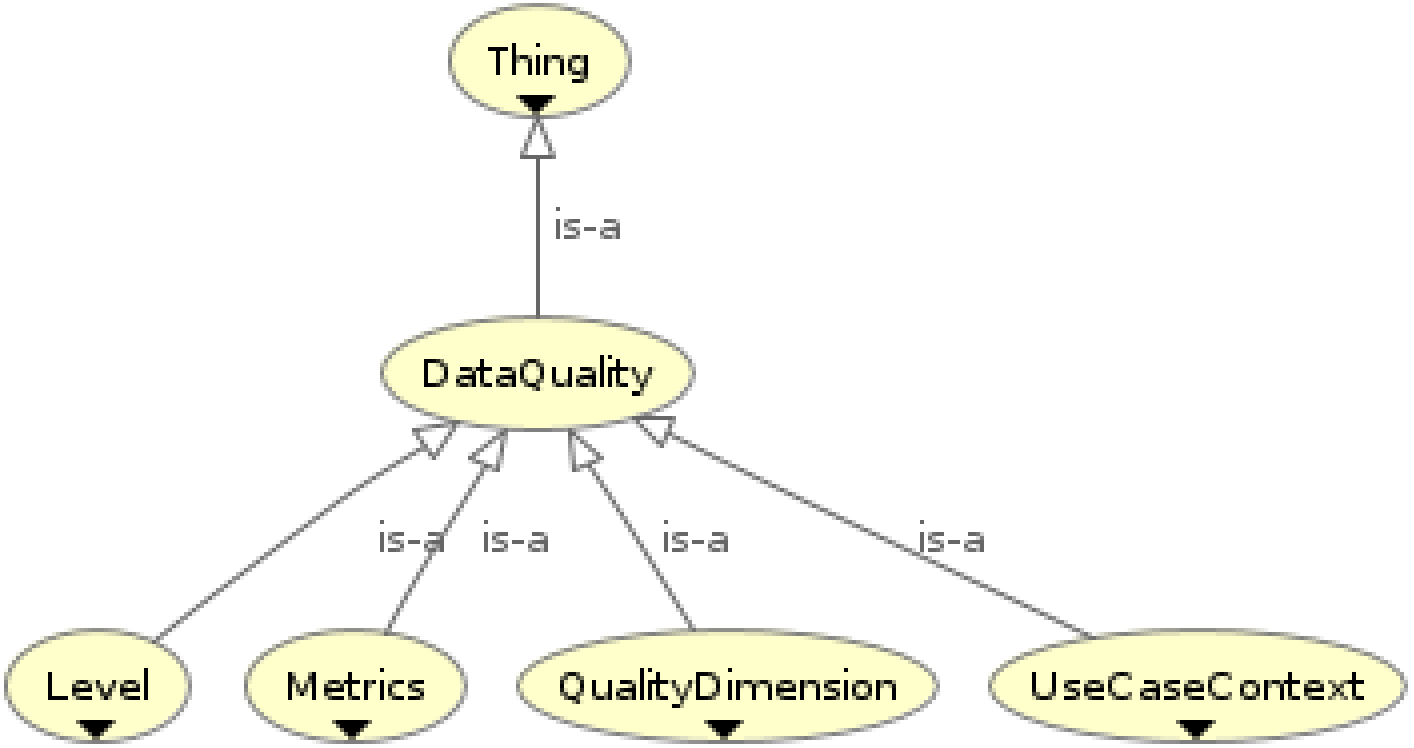
“We define data quality as data that are fit for use by data consumers“

[1]Wang, R Y; Strong, D M: *Beyond Accuracy: What data quality means to data consumers.* Journal of Management Information Systems, 12,4. 1996

Datenqualität in einer Ontologie

- Übersicht über die Aspekte der Datenqualität
- Charakterisierung der Datenqualitätsprozesse
- Integration der Qualitätsdimensionen von Wang

Ontologie[2]



[2]Schrader, T; Zhou, Y; Hahn, C; Niepage, S; Keune, D; Wetzel, T; Weckend, T; Schaaf, T; Hanß, S: *The problem of data quality in medical research science centers.* In CS&P 2008

Service Discovery

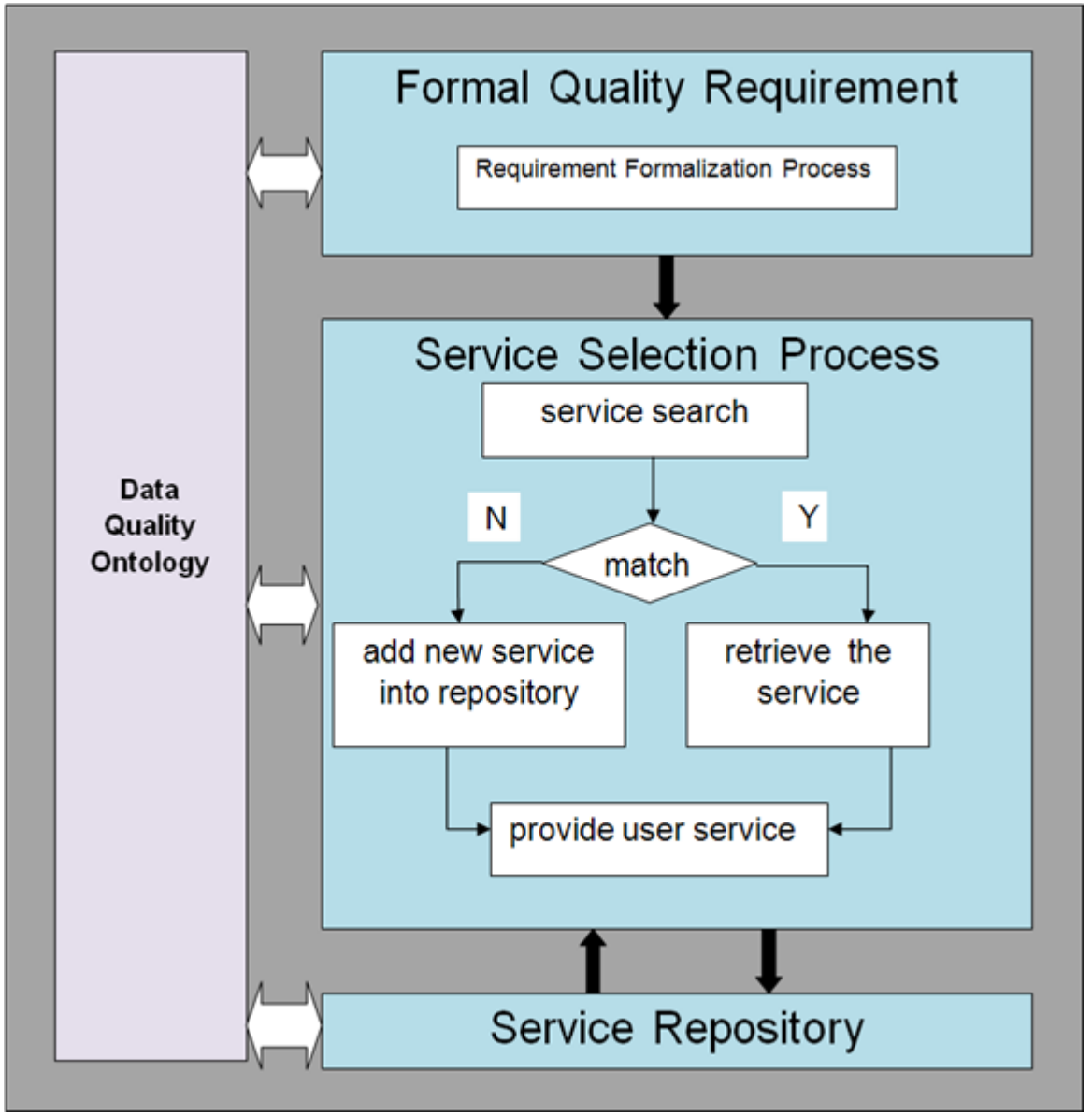
eine Eigenschaft von Services:

- Discoverable → UDDI

Semantic Web Services

- WSDL-S
- Semantische Beschreibung der Services mit Hilfe der Ontologie

Framework[3]



[3]Zhou, Y; Cornils, M; Hahn, C; Hanß, S; Niepage, S; Schrader, T: *A SOA-based Data Quality Assessment Framework in a Medical Science Center.* In: 14th ICIQ, 2009

Qualitätsmessung während der Datenintegration

Ontology class	
Quality dimension	Completeness/ FreeOfError
Use Case Context	Research
Metrics	Numeric
Level	Basic/Syntactic

Qualitätsmessung während der Datenintegration

<wssem:dataquality dimension="completeness" metric= "numeric" level="basic" usecasecontext="research" />

Xml-Schema (xsd):

```
<element name="dataquality" maxOccurs="1">
  <complexType>
   <complexContent>
      <extension base="wsdl:documented">
        <attribute name="dimension" type="String" use="required"/>
        <attribute name="metric" type="String" use="required"/>
        <attribute name="usecasecontext" type="String" use="required"/>
        <attribute name="level" type="String" use="required"/>
      </extension>
   </complexContent>
  </complexType>
 </element>
```

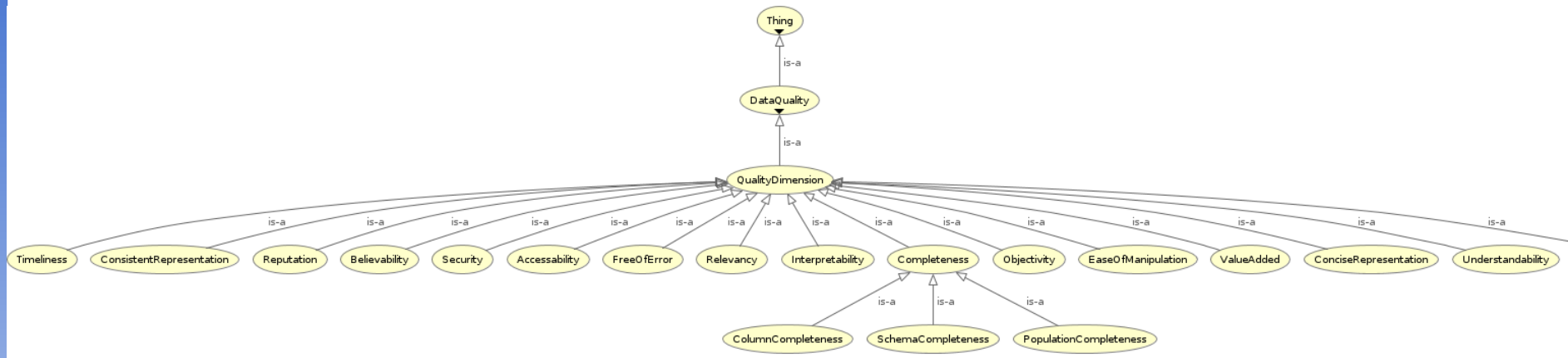
Qualitätsmessung während der Datenintegration

Entsprechend der Beschreibung der Datenfelder

- Completeness:
  - Felder Geschlecht und Geburtsdatum müssen für jeden Patienten vorhanden sein.
- Free of Error:
  - Feld CodeICD10 muss mindestens 3 Zeichen (265 von 38224 =0,7% <1%) und höchstens 5 (86 von 38224 =0,2% <1%) enthalten, Qualitätscheck gegen ICD10-Katalog

Fazit

- Keine Änderung der ursprünglichen Daten
- Rückmeldung an die Quelle
- Präsentation für den User



„118 data quality attributes collected from data consumers are consolidated into twenty dimensions“[1]

[1]Wang, Richard Y; Strong, Diane M: *Beyond Accuracy: What data quality means to data consumers.* Journal of Management Information Systems, 12,4. 1996

OpEN.SC

Vielen Dank für Ihre Aufmerksamkeit!

Kontakt:

sabine.hanss@charite.de

sonja.niepage@charite.de

open.sc-core@charite.de

CHARITÉ

Veröffentlichungen

- Schrader, Thomas; Zhou, Yao; Hahn, Claudia Niepage, Sonja; Keune, Dietmar; Wetzel, Thomas; Weckend, Thomas; Schaaf, Thorsten; Hanß, Sabine: *The problem of data quality in medical research science centers*. In CS&P 2008
- Zhou, Yao; Cornils, Malte; Hahn, Claudia; Hanß, Sabine; Niepage, Sonja; Schrader, Thomas: *A SOA-based Data Quality Assessment Framework in a Medical Science Center*. In: 14th ICIQ, 2009

Bsp2: Telekonsultation

Ontology class	Keywords
Quality dimension	Believability
Use Case Context	Different opinions, Data customer
Metrics	String
Level	Semantic