

QUALITÄTSSIEGEL FÜR DAS (SELF)PUBLISHING VON DATEN

Dr. Katrin Moeller
Historisches Datenzentrum
Sachsen-Anhalt

GLIEDERUNG

- Datenmanagement aus Perspektive der Wissenschaft (nicht für die Wissenschaft): Organisatorisches Zusammenwachsen
- Vergleich der Vor- und Nachteile bei Qualitäts- und Rohdaten in ihrem Mehrwert für die Forschung: a) Datenqualität b) Nachvollziehbarkeit c) Dokumentation
- Erfordernis neuer Datenkulturen: Akzeptanz von fehlerhaften Daten
- Abgestufte Qualitätssiegel zur Kennzeichnung

VORBEMERKUNG

Hier geht es nicht um die essentiellen Metadaten zu Datenpublikationen, die Voraussetzung für jede Datenpublikation sind!

Hier geht es um die inhaltliche Datenqualität, Dokumentation und Transparenz von Datensätzen!

DATENQUALITÄT: DATENSTRATEGIE 1

- Ziel des bisherigen Datenpublishings:
- hoch qualitätsgesicherte Daten – Double Keying oder nachträgliche Bereinigung und Korrektur maschinell erhobener Daten
- gilt auch für Digitalisate: individuelle Digitalisate werden von Bibliotheken für Publishing als zu schlecht beurteilt – für Forschungsprojekte mitunter aber schwierig erreichbar (Problem der mittelgeförderten und nichtgeförderten Projekte)
- Gilt auch für Norm- und Metadaten: Auszeichnung wird allmählich zur Pflicht (z.B. TEI) (Normdaten und Standards)
- Ziel: Standard einer Edition oder Buchqualität mit Ziel 100% Korrektheit, Qualität und Maschinenlesbarkeit

DATENQUALITÄT: DATENSTRATEGIE 2

- Problem: „Publikationsreife von Daten“ ist methodisch, fachlich und institutionell in Geisteswissenschaften bisher wenig ausgebildet bzw. institutionell untersetzt und kostet auf Seiten von Wissenschaft und Gedächtnisinstitutionen sehr viel Zeit
- führt zu wenigen, aber hochwertigen Publikationen – nicht aber unbedingt besserer Nutzbarkeit (Normierung von Texten in Geschichtswissenschaft)
- Verlust von vielen wertvollen, aber eben nicht dem Standard entsprechenden Daten
- Notwendig: Strategie der Massendigitalisierung und Texterkennung mit fehlerhaften Daten
 - Fehlerhafte Daten von Wissenschaftlern (Transkripte, Exzerpte, Gedächtnisprotokolle etc.)
 - Rohdaten, Rohdaten mit erster Bereinigung, Arbeitsdaten
 - Kooperative Daten mit unsicherer Datenqualität
- Angst vor „Verunreinigung“ von hochwertigen Daten

PROBLEM UND DISKUSSION

Hochqualitative Daten

Besonders wertvolle Einzeldaten

Publikationsreife

hoher Zeit- und Geldeinsatz

Reputation und wissenschaftlich
anerkannt

Fehlerbehaftete Daten

Massenhaft vorhanden

Arbeitsdaten von Wissenschaftlern

effektives Handling

Problem der Reputation, keine
Akzeptanz in vielen
Geisteswissenschaften (Wikipedia)

Problem: Wie bekommen wir ein gutes Handling nicht nur für hochqualitative Daten sondern auch fehlerhafte Daten hin?
Wie erreichen wir eine Trennung zwischen wichtigen fehlerhaften Daten und Müll?

PROBLEM 2: NACHVOLLZIEHBARKEIT (STANDARDS, NORMDATEN, DOKUMENTATION)

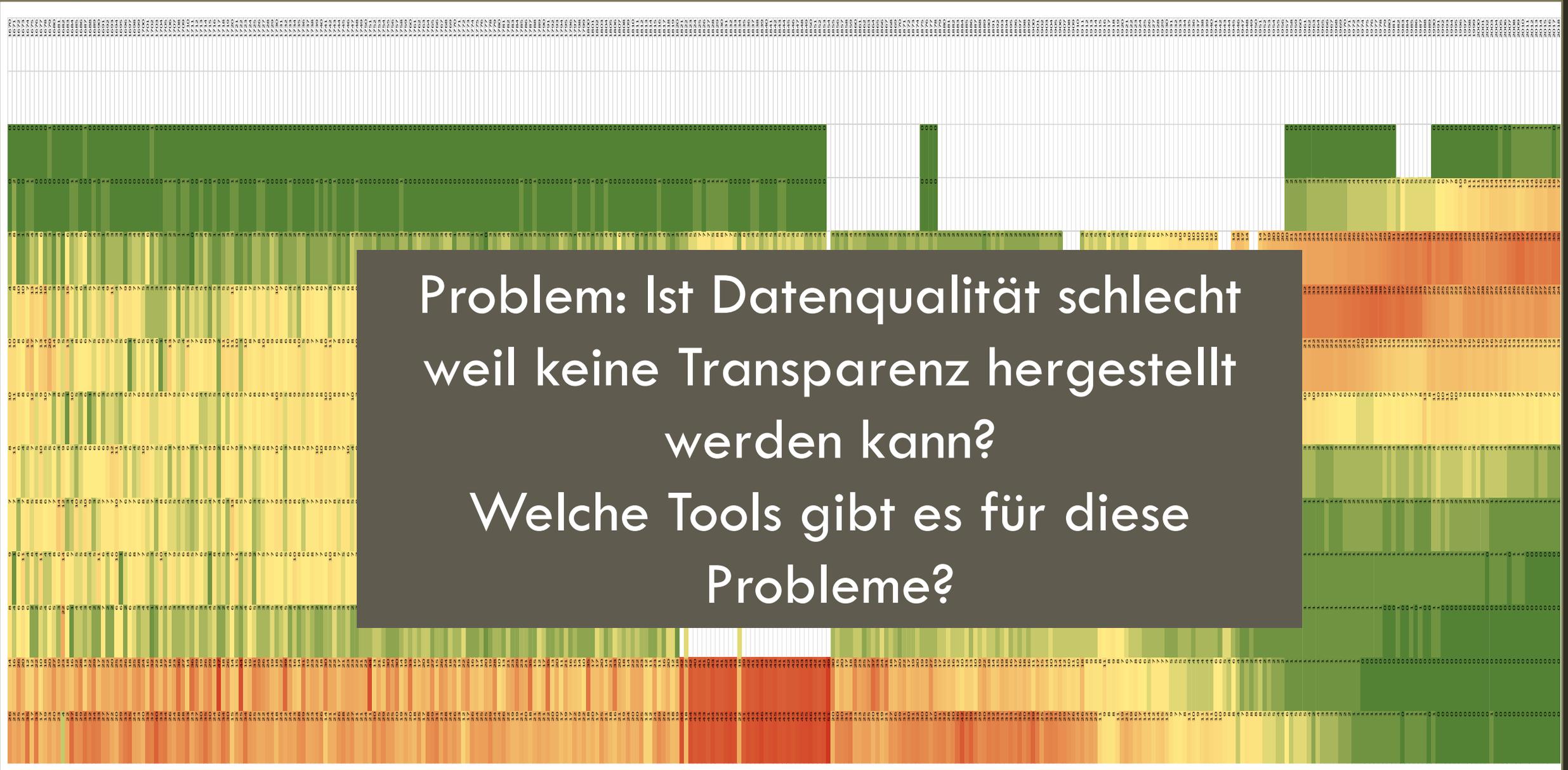
- Forderung FAIRE-Daten: Norm- und Metadaten, Quellen- und Literaturnachweis, Lizenzen, Dokumentation, Daten – Nachvollziehbarkeit
- Etablierung von Normdaten und Standards – etwa DFG Standards für Editionen und Basisstandard von DTA – nur sprachwissenschaftliche Standards – kein direkter Nutzen für andere Disziplinen
- Auch hier wieder Problem von Aufwand und Nutzen – Zusätzlich aber auch Problem der Darstellung und deren technischen Unterstützung
- Selfpublishing als besondere Herausforderung des Datenmanagements
- Standards holen WissenschaftlerInnen häufig nicht bei den eigentlichen Problemen ab bzw. schaffen eben hohe Hürden (z.B. DDI)

PROBLEM DER NACHVOLLZIEHBARKEIT

- ✓ Metadaten
- ✓ Lizenzen
- ✓ Versionen
- ? Dokumentation (Studiendesign und Methoden z.B. DDI)
- ? Quellenangaben – welche der 100 Einzelquellen repräsentiert welche Zelle, wie wurde der Datensatz aufgebaut?
- ? Nachvollziehbarkeit einzelner Datenbestandteile: Fehlende Werte
- ? Nachvollziehbarkeit bei parallelen Veröffentlichungen von Daten (unterschiedliche Quellen)
- ? Nachvollziehbarkeit bei Extrapolierung

Problem der Nachvollziehbarkeit von Daten

Sterbealter Halle in zehnjährigen Altersgruppen 1670-2018 (4524 Zellen)

The background of the slide is a large heatmap representing death rates in a hall over time. The x-axis represents time from 1670 to 2018, and the y-axis represents age groups in 10-year intervals. The color scale ranges from green (low death rate) to red (high death rate). A central grey box contains the following text:

Problem: Ist Datenqualität schlecht
weil keine Transparenz hergestellt
werden kann?
Welche Tools gibt es für diese
Probleme?

DATENQUALITÄT NACH WANG & STRONG

Merkmalsklasse	Qualitätsmerkmal		
	Intrinsische Datenqualität	Glaubhaftigkeit	Repräsentationelle Datenqualität
		Genauigkeit	
		Objektivität	
		Reputation	
	Kontextuelle Datenqualität	Mehrwert	Zugriffsqualität
		Relevanz	
		Zeitnähe	
		Vollständigkeit	
		Datenmenge	
			Interpretierbarkeit
			Verständlichkeit
			Konsistenz der Darstellung
			Knappheit der Darstellung
			Verfügbarkeit
			Zugriffssicherheit

GEWÜNSCHTER ANSATZ

- fehlerbehaftete Daten (Rohdaten) verstärkt publizieren
- Aufwand der Korrektur rechtfertigt Nutzen nur bei spezifischen fachwissenschaftlichen Nutzungsszenarien – die aber eher Problem der Nachnutzung darstellen
- Nachvollziehbarkeit ist in der Geisteswissenschaft (und sicherlich auch anderen Disziplinen) nicht grundsätzlich immer herstellbar
- Mehrwert der Nachnutzung auch bei fehlerhaften Daten hoch – Fehler können schnell erkannt werden
- Reputation

Absicherung, dass Information über Fehlerhaftigkeit erhalten bleibt!

Datenkulturen verändern

WELCHE INFORMATIONEN SOLLEN SIEGEL ÜBERMITTELN? WIEVIELE INFORMATIONSSCHICHTEN WERDEN BENÖTIGT UND SIND SINNVOLL?

Schicht 1: Qualitätslevel

Fehlerhafte Daten

Rohdaten / Aufbereitete
Rohdaten

Analysedaten

Publikationsdaten

Schicht 2: Datenerhebung

Maschinelle Daten

Kooperative Daten

Dynamische Daten

Projektdaten

Schicht 3: Transparenz und Dokumentation

Intransparente Daten

halbtransparente Daten

volltransparente Daten

DATENKULTUREN: ROH- UND QUALITÄTSDATEN BESSER KENNZEICHNEN

Qualitätssiegel die auf einen Blick über Datenqualität Auskunft geben

Aufgabe:

- Verständigung auf den Gebrauch dieser durch Datenzentren der Community
 - Wovon hängt die mögliche Nutzung in einzelnen Datenzentren ab? Was würde Nutzung begünstigen? Brauchen wir das?
- Bestimmen und Definition relevanter Informationen bzw. Qualitätsstufen in Abhängigkeit mit Datentypen
 - Welche Informationsschichten und welche Qualitätstypen sind in unterschiedlichen Forschungszusammenhängen bekannt und sollten ausgewiesen werden?
- Entwickeln von Siegeln
- Gebrauch durch Datenzentren