Klassifizierung von Datenqualitäts-Verbesserungsschritten

Katalog zu generischen Qualitätsverbesserungsmaßnahmen für Datenbestände mit teilweise unsicheren Daten (AP 2.8)

Entwickelt innerhalb des BMBF-geförderten Forschungsprojekt KONDA

KONDA - Kontinuierliches Qualitätsmanagement von dynamischen, zum Teil unsicheren Forschungsdaten

Autoren	Arno Kesper, Julia Rössel, Gabriele Taentzer
Datum	29.06.2023
Version	1.0
Status	published
Veröffentlichung	*29.06.2023*

Bundesministerium für Bildung und Forschung
Philipps Universität Marburg
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN
NIEDERSÄCHSISCHE STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN | SUB

Inhaltsverzeichnis

Einleitung

Am 3. und 4. März 2021 fand im Rahmen des KONDA Projekts der 2. Community Workshop statt. Im Rahmen dieses Workshops wurde mit über 30 Expertinnen und Experten ausführlich über Schritte der Qualitätsverbesserung von Forschungsdaten diskutiert. In Folge dessen kam diese Klassifizierung von Qualitätsverbesserungsschritten zustande.

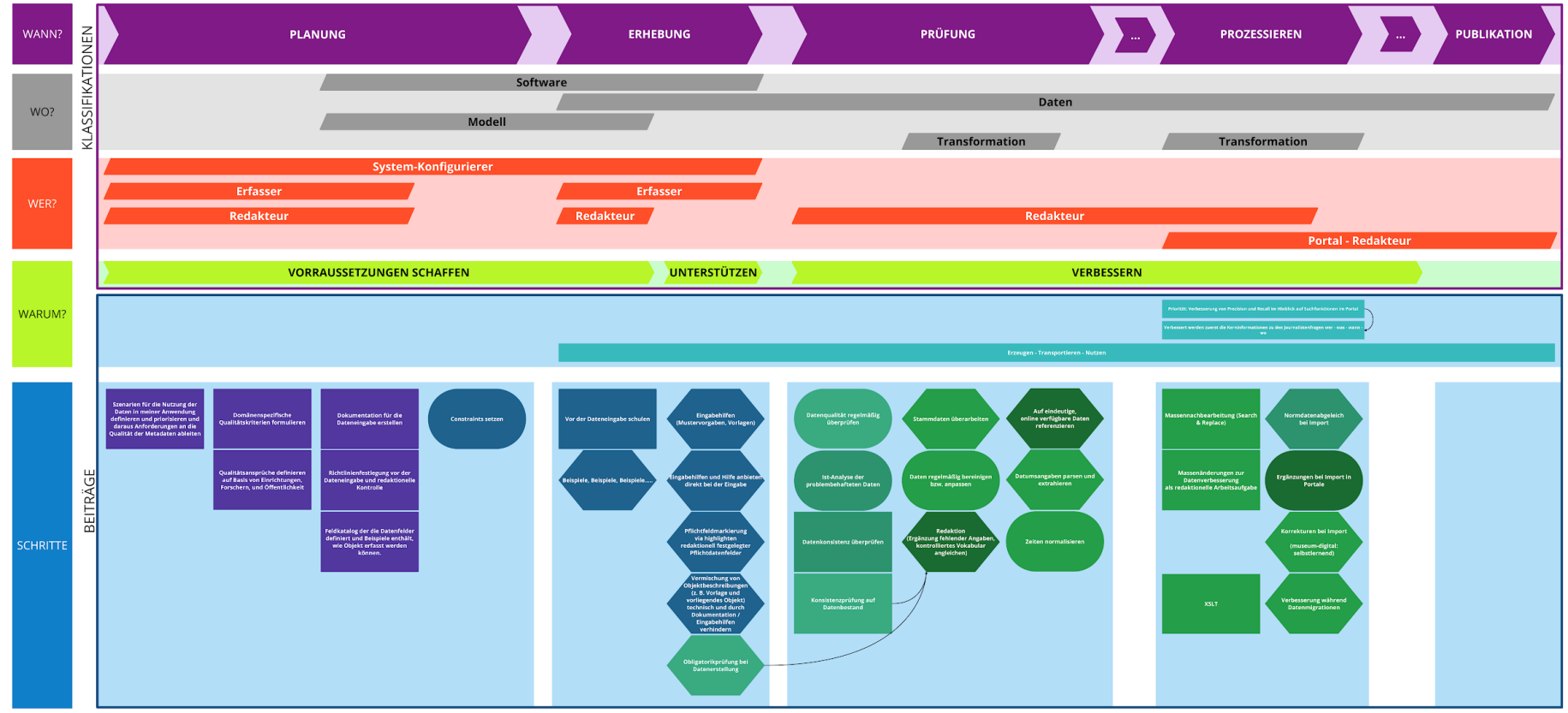
Die Diskussion wurde kollaborativ an einem Whiteboard-Tool geführt, an dem jeder Teilnehmer mitwirken konnte. Hierbei wurden zu Beginn die Ideen aller Teilnehmer gesammelt. Diese wurden dann als Diskussionsgrundlage genutzt, um eine Klassifizierung zu erstellen.

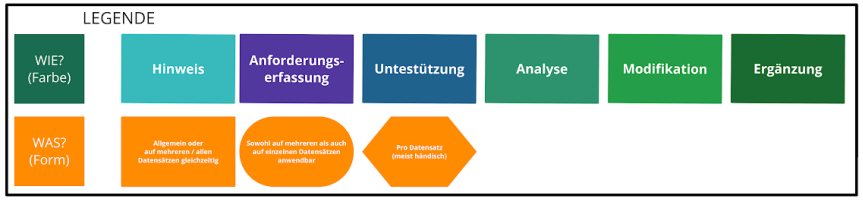
Während des Workshops wurde die folgende Darstellung entwickelt. Zu sehen ist eine Liste von gesammelten Verbesserungsschritten (Schritte, blau), basierend auf einem Brainstorming der Teilnehmer.

Nach dem Brainstorming wurden die genannten Schritte zuerst nach Ähnlichkeit graphisch aufgeteilt. Basierend darauf wurden mehrere, zueinander orthogonale Klassifizierungen erarbeitet. Diese entwickelten Klassifizierungen lassen sich anhand von Leitfragen (Wann, Wo, Wer, Warum, Wie, Was?) definieren. Die Leitfragen der Klassifizierungen sind auf der linken Seite zu sehen. Die Gruppen jeder Klassifizierung sind jeweils auf derselben Höhe in derselben Farbe zu finden. Im Folgenden werden die vorgeschlagenen Klassifizierungen im Detail vorgestellt. Die Beiträge der Workshop Teilnehmer waren Grundlage zur Erstellung der Kategorien und sind im Diagramm zu finden, werden aber im Text nicht mehr vollständig erwähnt.

Die folgende Darstellung kann zusätzlich unter folgendem Link gefunden werden:
https://doi.org/10.5281/zenodo.8096036

Konda Community Workshop 2 - Session 1: Klassifizierung von Daten Qualitätsverbesserung Schritten - Ergebnis





Aktivitäts-orientierte Klassifizierung (Wann?)

Qualitätsverbesserung kann an mehreren Punkten im Datenlebenszyklus ansetzen, welche zeitlich aufeinander folgen. Hier wird eine Klassifizierung der Qualitätsverbesserungsschritte anhand der Phasen des Datenlebenszykluses oder des Datenmanagementprozesses vorgestellt. Hierfür legen wir den Datenlebenszyklus des Rates für Informationsinfrastrukturen[1] zu Grunde.

Planung

Der erste Schritt zu einer guten Datenqualität fängt bereits bei der Planung der Daten an. Dabei soll festgelegt werden, welche Daten überhaupt erfasst werden sollen und in welcher Struktur das passieren soll. Ziel ist es, passende Voraussetzungen für die Datenerhebung zu schaffen, sodass die Daten mit höchstmöglicher Qualität erstellt werden können.

Die Planung sollte damit anfangen, unterschiedliche Szenarien für die Nutzung der Daten zu definieren, welche in der Anwendung möglich sein sollen. Daraufhin sollten diese priorisiert werden und Anforderungen abgeleitet werden, was die Qualität ausmacht. Anhand dessen sollten Qualitätskriterien definiert werden. Diese sind üblicherweise domänenspezifisch und hängen stark von der Einrichtung, den Datenerfassern, Datennutzern (z.B. Forschern) und der Öffentlichkeit ab. Eine zentrale Rolle sollte der Redakteur einnehmen, welcher diese Aufgabe aber in enger Absprache mit den Datenerfassern und den System-Konfiguratoren erfüllen sollte.

Wenn ein Konzept für die Datenerfassung steht, sollten die System-Konfigurierer der Erfassungssoftware (in Zusammenarbeit mit den Erfassern und Redakteuren) das Datenmodell definieren und die Erfassungssoftware dementsprechend konfigurieren.

Wenn dann das Datenmodell definiert und die Software konfiguriert wird, soll auch eine Dokumentation für die Dateneingabe erstellt werden. Diese wird meistens in Form eines Feldkatalogs für die Datenfelder definiert und mit Beispielen hinterlegt. Dies geschieht wieder hauptsächlich durch die Redakteure.

Des Weiteren sollte die Dateneingabe möglichst stark unterstützt werden, indem Regeln und Richtlinien als Constraints in die Erfassungssoftware implementiert werden.

Erhebung

Nachdem die Daten dann im Zuge der Erhebung strukturiert in ein Datenformat übernommen werden, müssen die Datenerfasser bestmöglich unterstützt werden, um Daten von höchstmöglicher Qualität erfassen zu können. Hierfür sollten die Erfasser durch System-Konfigurierer und Redakteure in den Funktionalitäten und der Nutzung der Software geschult werden, sowie in den Feinheiten der domänenspezifischen Datenerfassung. Diese Schulung sollte möglichst stark von Beispielen unterstützt werden.

Während der Datenerfassung selbst sollten die Datenerfasser von implementierten Eingabehilfen begleitet und unterstützt werden. Beispiele sind Mustervorgaben oder Vorlagen oder Pflichtfeldmarkierungen. Die definierten Constraints (z.B. Obligatorikprüfungen) sollten schon direkt während der Dateneingabe evaluiert werden und den Erfassern deutlich angezeigt werden.

Prüfung

Sobald eine Menge an Daten erfasst wurde, geht es im nächsten Schritt darum, die Datenqualität im gesamten Datensatz zu evaluieren und zu verbessern. Dies ist Aufgabe der Redaktion. Dazu muss zuerst die Datenqualität der existierenden Daten überprüft werden. Bei längeren Erfassungszeiträumen ist es dringend zu empfehlen, diese Qualitätsanalysen regelmäßig

[1] RfII – Rat für Informationsinfrastrukturen: Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, zweite Auflage, Göttingen 2019, 172 S, https://rfii.de/?p=4043.

durchzuführen. Hierbei geht es darum, problembehaftete Daten zu erkennen, sowie die Konsistenz des gesamten Datensatzes zu evaluieren.

Wenn eine solche Analyse abgeschlossen ist, kann der Redakteur die problematischen Daten verbessern, ergänzen und bereinigen. Dies geschieht zu großen Teilen manuell, bei häufigen Fehlern können jedoch auch Datentransformationen genutzt werden. Hierbei ist es die Aufgabe des Redakteurs, fehlende Angaben zu korrigieren und die kontrollierten Vokabulare ggf. anzugleichen. Des Weiteren sollten Referenzen auf online verfügbare Daten eingefügt werden und die Daten (beispielsweise Datumsangaben) normalisiert werden.

Bei den Bereinigungen sollten jeweils auch die Stammdaten (bereits lange bestehende Daten) miteinbezogen werden, um diese langsam zu optimieren oder an neue Gegebenheiten und Erkenntnisse anzupassen.

Prozessieren

Wenn Daten im großen Stil erfasst wurden, werden diese üblicherweise in irgendeiner Weise auf Portalen der Öffentlichkeit zugänglich gemacht. Jedoch müssen die Daten oftmals nochmal mittels Transformationen umstrukturiert werden, um sie in ein für das Portal passendes Format zu transferieren. Dies wird oftmals von den Redakteuren der Erfassung-Organisation und des Portals umgesetzt. Ein wichtiges Ziel hierfür ist es, die Precision und Recall im Hinblick auf die Suchfunktion des Portals zu optimieren. Dabei sollte eine Priorität auf die Fragen “Wer?, Was?, Wann? Wo? Warum?” gelegt werden.

Hierbei gibt es oftmals noch Massenänderungen zur Datenverbesserung. Wenn zum Austausch das Format XML genutzt wird, werden diese Transformationen üblicherweise per XSLT realisiert.

Insbesondere, während des Imports der Daten in die Datenbank des Portals, sollten jeweils noch einige Qualitätsanalysen durchgeführt werden. Beispielsweise sollten die Normdatenverlinkungen direkt abgeglichen werden. Gefundene Probleme sollten dann möglichst direkt während dieser Migration korrigiert werden.

Eine Datenmigration in Portale sollte möglichst regelmäßig erfolgen. Dabei sollte auch stets darauf geachtet werden, dass die bereits veröffentlichten Daten aktualisiert werden.

Publikation

Nach der Migration der Daten in die Umgebung des Portals folgt die Publikation im Portal. Damit werden die erfassten Daten mit größtmöglicher Qualität der Öffentlichkeit zugänglich gemacht.

Artefakt-orientierte Klassifizierung (Wo?)

Die Klassifizierung von Qualitätsverbesserungsschritten kann auch anhand des Artefakts erfolgen, welches Basis oder Gegenstand von Analysen und Verbesserungsmaßnahmen ist. Dieses kann das Datenmanagementsystem (Software), das Datenmodell, die Daten selbst oder eine Datentransformation sein.

Software

Einen wichtigen Teil des Datenqualitätsprozesses stellt das Datenmanagementsystem (DMS). Es dient dazu, die Daten zu verwalten und die Datenerfassung zu unterstützen. Das geeignete DMS kann einen großen Teil dazu beitragen, Daten in hoher Qualität zu erfassen, die Qualität weiter zu verbessern und diese Qualität zu bewahren.

Hier ist bereits die Auswahl wichtig, um beispielsweise eine möglichst flexible Implementierung von Constraints für die Datenbank zu unterstützen und die geeigneten Hilfsmittel für die Datenerfassung zu bieten. Dies umfasst die Optimierung der UI, der Dokumentation und der Performanz, sowie das Ergänzen von benötigten Features und das Einarbeiten von Nutzer-Feedback.

Modell

Die Klassifizierungsgruppe zum Modell enthält die Verbesserungsschritte, die die Entwicklung des Datenmodells sowie dessen Dokumentation und Feldkataloge betrifft.

Das Datenmodell ist wichtig, da es die Grundstruktur und das Format der Daten definiert. Des Weiteren bildet es auch die Grundlage für das Formular, welches die Datenerfasser erhalten.

Diese Struktur muss so gestaltet und in das DMS umgesetzt werden, dass sie alle gewünschten Daten so präzise wie möglich fassen kann. Dafür muss das Datenmodell die Daten ausreichend kleinteilig strukturiert werden, sodass die Daten in einer wiederverwendbaren, möglichst maschienenlesbaren Art und Weise erfasst werden können. Andererseits muss das Modell auch flexibel genug sein, um besondere, insbesondere granularere Daten beinhalten zu können.

Jegliche Mängel des Datenmodells müssen behoben werden. Hierbei kann das Modell mittels Refactorings verbessert werden, beispielsweise mittels Design-Patterns. Wichtig ist hier auch das Hinzufügen von fehlenden Constraints in das Datenmodell. Des Weiteren sollte das Modell mit einer geeigneten Dokumentation und einem Feldkatalog ergänzt werden, um die Datenerfassung zu unterstützen.

Daten

Das Haupt- und Endprodukt des Qualitätsmanagementprozess sind die Daten selbst. Diese können von den Datenerfassern maschinell oder (beispielsweise besonders häufig im Kulturerbebereich) händisch erfasst werden und von den Redakteuren ergänzt und korrigiert. Schließlich werden sie an die Portale zur Publikation weitergegeben. Die Daten sind das Hauptaugenmerk von Analysen und Verbesserungstechniken in diesem Prozess.

Aufbauend auf unterschiedlichen Datenanalysen, sind hier die Verbesserungsschritte der Datenbereinigungen und Nachbearbeitungen angesiedelt.

Transformation

Datenmodifizierungen werden oftmals mittels Transformationen realisiert. Im Laufe eines Datenlebenszyklus kann es vorkommen, dass die Daten mehrfach durch unterschiedliche Transformationen modifiziert werden, insbesondere um Qualitätsverbesserungen umzusetzen, jedoch auch um Strukturänderungen vorzunehmen. Es ist wichtig darauf zu achten, dass die Transformation eine hohe Qualität aufweist, sodass die gewünschte Modifikation korrekt umgesetzt wird und nirgends neue Qualitätsmängel an den Daten entstehen.

Es muss also z.B. sichergestellt werden, dass die Daten während des Transformationsvorgangs in die richtigen Felder gelangen und nicht unbrauchbar werden. Insbesondere muss darauf geachtet werden, wenn die Datenbanken die Inkremente der Informationen unterschiedlich aufspalten (z.B. Vor- und Nachname in einem oder mehreren Feldern).
Zusammengefasst gehören in diese Kategorie die Verbesserungsschritte, welche sich um die Entwicklung und Qualitätssicherung von Datentransformationen drehen.

Rollen-orientierte Klassifikation (Wer?)

Eine weitere Kategorisierung der Datenqualitäts-Verbesserungsschritte kann anhand der Rollen der Personen getroffen werden, welche an der Datenerstellung beteiligt sind. Die Verbesserungsschritte sind jeweils der Rolle zuzuordnen, welche die Haupttätigkeit ausführt.

System-Konfigurator

Der System-Konfigurator ist dafür verantwortlich, das Datenmodell zu entwickeln und in ein DMS zu übertragen. Dabei muss er die domänenspezifischen Anforderungen berücksichtigen. Im Laufe des Datenlebenszyklus können neue Anforderungen dazu kommen, sich Anforderungen ändern oder Mängel an der bisherigen Definition und Konfiguration des Modells bekannt werden. Der System-Konfigurator ist dafür verantwortlich, diese Anforderungen einzupflegen oder zu korrigieren.

Erfasser

Die Hauptaufgabe eines Datenerfassers ist es, neue Daten anhand der Eingabemaske des Erfassungssystems zu erstellen. Teilweise, insbesondere im Kulturbereich, ist das Erfassen keine triviale Aufgabe, da hier eine Transferleistung passieren muss, für die die Erfassenden bestimmte Vorkenntnisse zum Gegenstand der Erfassung benötigen, um diesen adäquat im Kontext der Datenbank zu erschließen. Dabei muss er sich an die Vorgaben halten, die durch das Datenmodell, das DMS, sowie Feldkataloge definiert wurden. Wenn neue Forschungsergebnisse bekannt werden, beispielsweise neue Erkenntnisse oder Widerlegungen von alten Annahmen, müssen diese in die Daten eingepflegt werden. Für solche semantischen Verbesserungen der Daten sind die Datenerfasser zuständig.

Redakteur

Der Redakteur nimmt die Rolle des Hauptverantwortlichen für die Daten der Erfassungsinstitution ein. Er definiert die Qualitätskriterien für die Daten, schult und unterstützt die Datenerfasser bei der Erfassung, und ist für die Nachbearbeitung der Daten verantwortlich. Dabei führt er Qualitätsanalysen durch und versucht, durch Ergänzungen und Korrekturen die Qualität der Daten nach und nach zu optimieren. Die Analysen und damit einhergehende Verbesserungsmaßnahmen des Redakteurs beziehen sich hauptsächlich auf syntaktische Aspekte der Daten, sowie die Konsistenz über die gesamte Datenbank hinweg.

Portal-Redakteur

Die Aufgabe, die Daten zu veröffentlichen, übernimmt die Rolle des Portal-Redakteurs. Er definiert und vermittelt das portalspezifische Datenmodell, sowie die Qualitätskriterien des Portals. Er bekommt die Daten von einer oder mehreren Organisationen gestellt und muss diese dann aufbereiten, sodass sie mit guter Qualität auf seinem Portal veröffentlicht werden können. Hierbei verbessert er Probleme, welche durch Daten-Transformationen entstehen und korrigiert Konsistenz-Mängel. Insbesondere konzentriert er sich aber darauf, die Daten so zu verbessern, dass die Anzeige im Portal valide ist und die Auffindbarkeit (Retrieval) gewährleistet ist.

Plan-orientierte Klassifikation (Warum?)

Die Qualitätsverbesserungsschritte können auch anhand ihres übergeordneten Zieles klassifiziert werden: Voraussetzungen schaffen für eine qualitativ hochwertige Erfassung, das Unterstützen der Erfasser und das Verbessern von bestehenden Daten.

Voraussetzungen schaffen (Datenmodell, Software)

Ein großes Ziel ist es, möglichst gute Voraussetzungen zu schaffen, damit neue Daten von höchstmöglicher Qualität erfasst werden können. Hier geht es darum, ein passendes Datenmodell auszuwählen oder zu erstellen, die Erfassungssoftware vernünftig zu konfigurieren und die Datenerfasser zu schulen. Dies ist ein fortlaufender Prozess, da anhand von neuen Erkenntnissen immer wieder Verbesserungen (Nachjustierungen) an der Konfiguration oder Neuschulungen stattfinden können. Insbesondere beim Anlernen der Datenerfasser sollten mit ihnen regelmäßig die Erfassungspraxis und gefundene Qualitätsprobleme besprochen werden.

Erfassung unterstützen (Arbeitskraft)

Während der Datenerfassung sollte ein Augenmerk darauf gerichtet sein, die Datenerfassung bestmöglich zu unterstützen. Hierbei ist insbesondere darauf zu achten, dass den Datenerfassern Eingabehilfen zur Verfügung stehen. Da es normal ist, dass bei der Datenerfassung einige seltsame Fälle auftreten können. Somit ist es wichtig, dass sie bei jeglichen Fragen einen Ansprechpartner finden.

Vorhandene Daten verbessern (Altdaten)

Beim Erfassen von Forschungsdaten ist darauf zu achten, dass auch die bestehenden Daten nie perfekt sind. Sobald Daten vorhanden sind, geht es also darum, diese fortwährend zu ergänzen und zu überarbeiten, sowie um neue Forschungsergebnisse zu erweitern. Auch diese Verbesserung bereits vorhandener Daten (Altdaten) ist ein fortlaufender Prozess.