Data Quality für das IT-System PIZO des Zoos Pirmasens

Inhaltsverzeichnis

Zielsetzung und Rahmenbedingungen....1
Zielsetzung des Konzepts....1
Rahmenbedingungen und Empfehlung zum Vorgehen....1
Elemente der Datenarchitektur und Datenstruktur....2
Vollständige und korrekte Altdatenübernahme....3
Sicherstellung der Datenqualität im laufenden Betrieb....5
Anhang....7
Qualitätskriterien....7
Fehlerarten und Vorgehen zur Fehlerbereinigung....8
Tools, Programme, ….... 11

Zielsetzung und Rahmenbedingungen

Zielsetzung des Konzepts

Das Konzept soll eine Grundlage für die fortlaufende Bearbeitung des Datenbestandes bieten, sodass die Datenqualität auf ein Level über 97 % kommt und der Qualitätslevel gehalten werden kann.

Rahmenbedingungen und Empfehlung zum Vorgehen

Für die Gestaltung des neuen IT-Systems wurde eine Datenarchitektur definiert.

Die Daten sind vollständig und korrekt zu erfassen. Geplant ist eine manuelle Aufnahme durch Datentypisten. Essenziell für eine sichere und fundierte Nutzung des neuen IT-Systems ist, vor der erstmaligen Aufnahme ins IT-System, die Datenqualität zu verbessern.

Die Daten sind neben dem laufenden Betrieb auch für die Erfüllung der gesetzlichen Dokumentationsanforderungen (ZooG: zB. Artbestand, Tierbestand, Veterinärmedizinischische Daten, Fütterungspläne, Bildungs- und Informationsmaßnahmen) notwendig.

Ein Data Owner ist Teil einer Data Organisation und idealerweise einer Data Governance Struktur. Für die Sicherstellung der Datenqualität im laufenden Betrieb sollen „Spielregeln“ definiert werden und verantwortliche Data Owner benannt werden.

Elemente der Datenarchitektur und Datenstruktur

Das Datenbankdesign und die Generierung des logischen Datenmodells sehen folgende Spezifikationen vor:

ERM-Modell (Entity-Relationship-Modell):

Das spezifisches Modell innerhalb der Datenmodellierung, das Entitäten und deren Beziehungen zueinander beschreibt.

Das ERM Modell liegt unter Projektarbeit Gruppe 03 Ordner 02 ERM als **PIZO_ERM_final.png.**

Relationale Datenbank inklusive DB-Schema:

Die Struktur der Datenbank ist hier definiert.

Die DB mit der Struktur, Tabellen, Beziehungen und.. liegt unter Projektarbeit Gruppe 03 Ordner 03 als Operative Datenbank **PIZO_DB_final.db**

Datenmodell Data Warehouse:

Das Modell beschreibt, wie Daten gespeichert und miteinander in Beziehung gesetzt werden und liegt unter Projektarbeit Gruppe 03 Ordner 04 Datawarehouse: DataVault_als **PIZO_DataVault_DataModell_DWH_final.png**

Relationale Datenbank auf Basis eines Star Schemas (Zusatzaufgabe):

PIZO_Star_Schema_Lieferposition_DWH_final.png

Data Dictionary:

Ein Repository, das die Bedeutung, Struktur und andere Attribute der Daten in einem Informationssystem beschreibt.

Das Data Dictionary liegt unter Projektarbeit Gruppe 03 Ordner 05 Data Quality als **PIZO_Data_Dictionary_final.xls**

Data Quality Konzept:

beschreibt den Umgang mit Daten hinsichtlich Qualität und Zuständigkeiten.

Das Data Quality Konzept liegt unter Projektarbeit Gruppe 03 Ordner 05 Data Quality als **PIZO_Data_Quality_Konzept_final.pdf**

Es ist sehr empfehlenswert, dass alle Beteiligte, auch die Anwender, die inhaltliche Beschreibung zu lesen, um ein einheitliches Verständnis zu haben. Insbesondere das Data Dictionary hilft bei der Datenbereinigung.

Vollständige und korrekte Altdatenübernahme

Warum ?

Die Daten werden erstmalig aus Nicht-IT Systemen verschiedener Datenquellen in das neue IT-System PIZO übernommen.

Wer?

Die Altdatenübernahme soll manuell durch Datentypisten vorgenommen werden.

Die geplanten zwei internen IT-Mitarbeiter sind Projektleiter für die Altdatenübernahme, leiten die Datentypisten an, koordinieren die Reihenfolge der Datenimporte und erstellen und verwalten die Berechtigungen.

Data Owner stehen für Fragen zur Verfügung.

Wie?

Die Datenquellen sollen vollständig erfasst und ins neue IT-System integriert werden. Die Datentypisten sollen für die Erfassung in Hinblick auf die Qualitätskriterien und Fehlerarten sowie Fehlerbehebung geschult werden, eine Auflistung mit Beispielen zu den Qualitätskriterien findet sich in diesem Dokument. Das Data Dictionary gibt zu jedem Datenobjekt einen Hinweis zur inhaltlichen Interpretation und in der Spalte Datenqualität – Aspekte sind hilfreiche Ansätze. Mittels dieses Data Profilings werden inhaltliche Probleme, Redundanzen und Inkonsistenzen erkannt; diese können in zukünftigen Prozessen vermieden werden.

Vor dem Import soll eine Bereinigung, Korrektur und Abgleich der Daten (Data Cleansing) manuell erfolgen. Es gibt auch viele technische Lösungen, die hierbei automatisiert helfen können.

Bei der Datenbereinigung kann man als Werkzeug Regex (Reguläre Ausdrücke) benutzen, die in vielen Tools oder Programmiersprachen genutzt werden können. Ein populäres

Beispiel ist die Suchen und Ersetzen Funktion in MS Word; es gibt eine Vielzahl von spezifischen Tools und ist in Programmiersprachen integriert.

Wertebereiche:

Es wurden für folgene Parameter Wertebereiche definiert :

- Anrede (für Mitarbeiter, Ärzte, Lieferanten):Herr,Frau,,,
- Einheiten für die Futter- / Lagermengen (des eingekauften bzw. des eingelagerten Futters):Ballen,Rollen,Stück,Sack,kg, g
- Einheiten Gewicht (der Tiere):t,kg,g,,
- Einheiten Groesse (der Tiere):m,cm,mm,,
- Einheiten Groesse (des Geheges):m2,m3,,,
- Geschlecht 1 (Mitarbeiter):männlich,weiblich,,,
- Geschlecht 2 (Tiere):männlich,weiblich,zwitter,,
- ja / nein (für alle ja / nein Felder: im Zoo geboren, etc.)):ja,nein,,,
- Land: (i.S.v. Nation: Ärzte, Lieferanten): Afghanistan bis Zypern ; Details siehe Datei Wertebereiche.txt
- MwSt. Satz in Prozent:19,7,,,
- PLZ: alle in Deutschland gültigen, ; Details siehe Datei Wertebereiche.txt
- Titel (für Mitarbeiter, Ärzte, Lieferanten):Dr.,Dr. med.,Prof.,Prof. Dr.,
- Vertretungsgründe (für Mitarbeiter):Krankheit,Urlaub,Schwangerschaft,Fortbildung,Sonstige
- Währungen (mit der das eingekaufte Futter abgerechnet wird):Euro,US-Dollar,Australischer Dollar,Pfund Sterling,

Es ist zu empfehlen für Parameter, die eine Meldepflicht nach sich ziehen (zB. Meldepflichtige Krankheiten), diese in die Datenbank aufzunehmen um eine Eingabevalidierung zu erzielen. Es ist ebenfalls für eine gute Datenqualität vorteilhaft, möglichst viele Eingabevalidierungen einzurichten, insbesondere für Tierkrankheiten, Tiermedikamente und alle Stammdaten (Tierärzte, Mitarbeiter, Gebäude, Gehege, Rundweg, Tiere, Tierarten). Für den Lieferanten ist zu überlegen ob eine Eingabevalidierung sinnvoll ist; das könnte der Fall sein wenn wenig neue Lieferanten auftreten können. Sollten neue Werte hier notwendig sein, ist zu klären welche und wieviele Personen hier Stammdatenerweiterungen vornehmen dürfen.

Die Datenquellen (Excellisten, Worddokumente, Informationen aus Emails, etc.) sollen in einem zentralen Dokument aufgelistet werden, sodass die Vollständigkeit und die notwendige Dokumentation sichergestellt werden können.

Die Reihenfolge und das Vorgehen für die Art und Weise des Datenimports muss festgelegt werden.

Folgende Maßnahmen sind wichtig zur Verbesserung der manuellen Dateneingabe durch die Datentypisten in Zusammenarbeit mit dem Data Owner:

Nach dem Datenimport soll eine Qualitätskontrolle und ein Vergleich der Daten aus den Datenquellen und dem Datenbestand im neuen IT-System vorgenommen werden.

Sicherstellung der Datenqualität im laufenden Betrieb

Warum?

Für die Verwaltung und den Betrieb des Zoos sind korrekte und vollständige Daten unabdingbar, sie bilden das informationelle Fundament, ermöglichen den Geschäftsbetrieb und sind teils rechtlich notwendig. Sie sind die Basis für geschäftliche Entscheidungen und für betriebliche und gesellschaftliche Informationsbedarfe.

Bestimmte Parameter wie Krankheiten und Medikamente erfordern eine vollständige Korrektheit, u.a. aufgrund gesetzlicher Regelungen (z.B. das TkrMeldpflV).

Im täglichen Betrieb entstehen täglich neue relevante Daten, die gemäß der definierten Struktur vollständig und korrekt integriert werden müssen.

Wer ?

Alle Personen, die Daten erfassen oder verwenden

Die Datenqualitätskriterien sollen allen bekannt sein, jeder soll Zugriff auf die inhaltliche Erläuterung des Data Dictionary haben.

Festlegung von Data Owner mit der Gesamtverantwortung

Hierbei ist eine Entscheidung über die Installation eines oder mehrerer Data Owner zu treffen. Es ist festzulegen, wieviele Personen mit Data Owner Rollen vergeben werden und ob es praktikabel ist diese Rolle gemäß der Zuständigkeiten zu verteilen. Die Zuständigkeiten könnten folgendermaßen aufgeteilt werden: Tierverwaltung (Zu- und Abgänge, Unverträglichkeiten), Fütterung Tier, Krankheitsbehandlung Tier, Personenverwaltung, Buchhalterische und Rechtliche Themen, Gebäude/Gehegeverwaltung, Rundwegeverwaltung,

Was sind generell die Aufgaben von Data Owner?

- **Definition der Datenanforderungen:** Der Data Owner definiert die Anforderungen an die Daten, die für den Geschäftsbetrieb benötigt werden.
- **Auswahl von Datenquellen:** Der Data Owner wählt die Datenquellen aus, aus denen die benötigten Daten abgerufen werden.

- **Implementierung von Datenprozessen:** Der Data Owner implementiert Prozesse für die Erfassung, Speicherung, Bereinigung und Transformation der Daten.
- **Überwachung der Datenqualität:** Der Data Owner überwacht die Qualität der Daten und stellt sicher, dass sie für den vorgesehenen Zweck geeignet sind.
- **Bereitstellung von Datenzugriff:** Der Data Owner stellt den berechtigten Nutzern den Zugriff auf die Daten zur Verfügung.
- **Einhaltung von Richtlinien:** Der Data Owner stellt sicher, dass die Daten gemäß den Richtlinien des Unternehmens verwendet werden.

Quelle: https://www.datainstitute.io/glossar/data-owner

Wie?

Es sollen Vorgehensweisen und gemeinsame Ressourcen geschaffen werden, um ein gemeinsames Verständnis von Geschäftsbegriffen im Unternehmen sicherzustellen. Ein kontinuierlicher Prozess des Datenqualitätsmanagements ist entscheidend, um langfristig hohe Datenqualität zu gewährleisten.

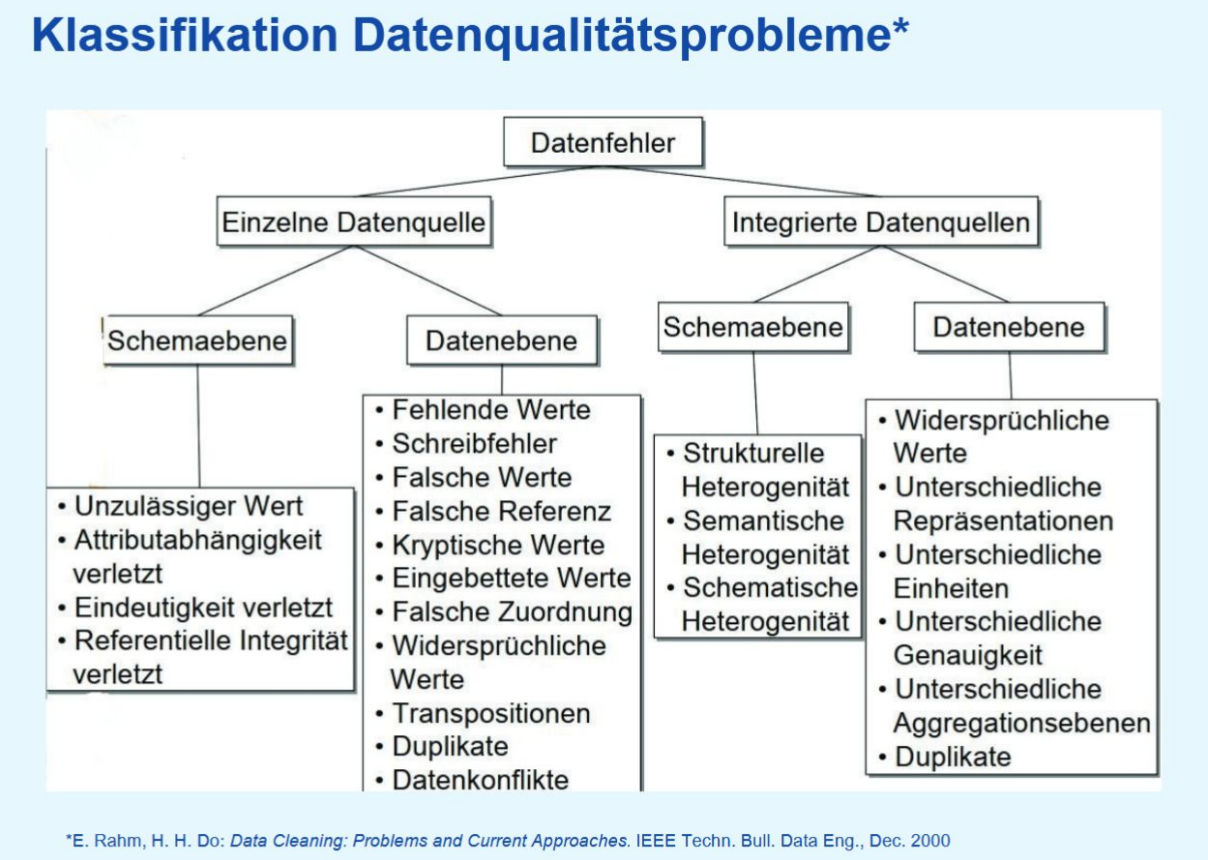
Regelmäßige Überprüfungen der Daten durch den zuständigen Data Owner stellen eine gleichbleibende Datenqualität sicher.

Datenqualitätsschulungen soll die Mitarbeiter für die Bedeutung der Datenqualität sensibilisieren und sie befähigen, aktiv zur Qualitätssicherung beizutragen.

Technologische Lösungen wie Datenbereinigungstools und Monitoring-Tools können helfen, die Datenqualität aufrechtzuerhalten.

Anhang

Qualitätskriterien



- Konsistenz: Widerspruchsfreiheit
- Korrektheit: Übereinstimmung mit der Realität
- Vollständigkeit:
  - Abwesenheit von fehlenden Werten
  - Ein Datensatz muss alle notwendigen Attribute enthalten.
  - Attribute müssen alle notwendigen Daten enthalten
  - Beispiel: Vollständige Adresse inklusive Hausnummer
- Genauigkeit und Granularität:

    - Daten müssen in der geforderten Exaktheit vorliegen.
    - Beispiel: Anzahl der Nachkommastellen, tagesgenaue Daten
- Einheitlichkeit
- Zuverlässigkeit und Glaubwürdigkeit: Nachvollziehbarkeit der Entstehung, Vertrauenswürdigkeit des Lieferanten
- Verständlichkeit: inhaltlich und technisch, strukturell für die jeweilige Zielgruppe
- Verwendbarkeit und Relevanz: geeignetes Format, Zweckdienlichkeit
- Redundanzfreiheit
    - innerhalb der Datensätze dürfen keine Dubletten vorkommen
- Schlüsseleindeutigkeit: Eindeutigkeit von Primärschlüsseln
- Referentielle Integrität: zu jedem Fremdschlüssel existiert ein Primärschlüssel in der referenzierten Relation

Fehlerarten und Vorgehen zur Fehlerbereinigung

Um die genannten Fehlerarten in Datensätzen zu beheben, empfiehlt es sich die Datenaufbereitung systematisch in mehreren Schritten abzuarbeiten.

NULL Wert / Missing Values (kein Eintrag)

Fehlende Werte in einem Datensatz, die entweder als NULL oder als leere Felder angezeigt werden.

Fehlerbehebung:

- **Einsetzen/Ersetzen**: Fehlende Werte mit statistischen Methoden wie dem Mittelwert, Median oder Modus ersetzen.
- **Vorheriges Entfernen**: Zeilen mit zu vielen fehlenden Werten oder irrelevante Spalten ganz entfernen.
- **Default-Werte/ Standardwerte setzen (zb. 0 für Nicht zutreffend)**
- **Statistische Vorhersage**

Unvollständige Felder (Eintrag)

Inkomplette oder teilweise ausgefüllte Felder (z. B. Name und Adresse, aber die Stadt fehlt).

Fehlerbehebung:

- **Datenbereinigung**: Manuelle Prüfung und Ergänzung fehlender Felder, wenn möglich.

- **Standardisierung**: Definition von Standardwerten für obligatorische Felder, um sicherzustellen, dass sie korrekt ausgefüllt werden.

Fehlerhafte Eingaben

Fehlerhafte Eingaben können nur mit Fachwissen oder Historienwissen erkannt werden

- **Eingabevalidierung:** mit Referenzwerten bei der Dateneingabe
- **Inhaltliche Validierung und Prüfung auf Plausibilität:** mit Fachpersonal oder externen Datenquellen

Spalte nicht in 1. Normalform (mehrfache Belegung)

Daten in einer Spalte sind nicht atomar, d.h., mehrere Werte sind in einer einzelnen Zelle gespeichert (z. B. mehrere Telefonnummern oder Adressen in einer einzigen Zelle).

Fehlerbehebung:

- **Normalisierung**: Die Daten so umstrukturieren, dass jede Spalte nur einen Wert enthält (d.h., Werte in mehrere Spalten oder Tabellen aufteilen).
- **Verwendung von Relationen**: Erstellen einer separaten Tabelle in der DB für die wiederholten Daten und Referenzierung dieser durch Fremdschlüssel.

Dubletten

Duplikate von Datensätzen, die mehrmals vorkommen.

Fehlerbehebung:

- **Manueller Abgleich**
- **Deduplizierung**: Durchführen von Duplikaterkennungsalgorithmen (z. B. durch Abgleich von Schlüsselfeldern oder durch Vergleich von gesamten Datensätzen).
- **Gruppierung**: Daten nach eindeutigen Identifikatoren gruppieren und nur einen Datensatz pro Gruppe behalten; eignet sich jedoch nicht z.B. für Kreditoren-/Debitorenstammdaten

Transposition (Daten vertauscht)

Daten wurden vertauscht, d.h., Zeilen und Spalten sind vertauscht worden oder die Daten wurden in falschen Feldern gespeichert.

Fehlerbehebung:

- **Datenprüfung und Korrektur**: Manuell oder automatisch die Zeilen und Spalten in die richtige Reihenfolge bringen.
- **Skripte zur Transformation**: Wenn die Transposition systematisch ist, kann ein Skript zur Bereinigung der Daten verwendet werden.

Codepage-Problematik / Messy Value

Fehlerhafte Zeichen aufgrund unterschiedlicher Zeichencodierungen (z. B. Umlaute oder Sonderzeichen werden falsch angezeigt).

Fehlerbehebung:

- **Encoding-Konvertierung**: Die Daten in eine konsistente Zeichencodierung (z. B. UTF-8) umwandeln.
- **Fehlerhafte Zeichen reparieren**: Suchen und Ersetzen von fehlerhaften Zeichen durch korrekte Zeichen. Beispielsweise bietet Excel die Suchen und Ersetzen Funktion.

Wertebereich unklar

Unklare oder nicht spezifizierte Wertebereiche, was zu Inkonsistenzen führt (z. B. ein Alter von 200 oder -5 Jahren).

Fehlerbehebung:

- **Validierungsregeln definieren**: Regeln zur Eingabe von Werten festlegen (z. B. Alter zwischen 0 und 120).
- **Datenbereinigung**: Anomalien manuell oder automatisch erkennen und korrigieren.

Falsches Datenformat

Daten sind im falschen Format (z. B. Datum im Textformat anstelle des Datumsformats).

Fehlerbehebung:

- **Formatkonvertierung**: Die Daten ins richtige Format bringen (z. B. "DD/MM/YYYY" in "YYYY-MM-DD").
- **Automatisierung**: Skripte oder Datenbankfunktionen nutzen, um Formate zu standardisieren.

Schlechte Erweiterbarkeit

Die Datenstruktur ist nicht flexibel genug, um zukünftige Änderungen oder Erweiterungen aufzunehmen.

Fehlerbehebung:

- **Datenmodellierung verbessern**: Datenbankdesign nach Best Practices und Normen (z. B. Normalformen) anpassen.
- **Flexibilität erhöhen**: Nutzung von flexiblen Datentypen oder Hierarchien (z. B. JSON oder XML) in der Datenbank, wenn notwendig.

Kein Surrogatschlüssel

Fehlender eindeutiger Identifikator (z. B. eine ID) für Entitäten, was die Referenzierung und Beziehung zwischen Tabellen erschwert.

Fehlerbehebung:

- **Surrogatschlüssel hinzufügen**: Einen eindeutigen, nicht biologischen Schlüssel (z. B. eine Auto-Inkrement-ID) in die Datenbank einfügen.
- **Primärschlüssel nutzen**: Sicherstellen, dass jede Entität einen klar definierten, eindeutigen Identifikator hat.

Unerkannte und/oder ungenutzte Datensilos

Doppelte Datenhaltung kann zu widersprüchlichen Informationen führen

Fehlerbehebung:

- **Datensilos abschaffen**
- **Sinnvolles Berichtswesen einrichten**: über die zentrale Datenbasis erhält jeder zeitgleich die gleiche Information

Tools, Programme

- Microsoft SQL Server Integration Services (SSIS)
- Oracle Warehouse Bilder
- SAS Data Quality Server
- WinPure List Cleaner Pro