DATA QUALITY KONZEPT

für den Zoo Pirmasens

Müller AG Gruppe 1

Inhaltsverzeichnis

Abkürzungen

Abkürzung	**Bedeutung**
DWH	Datawarehouse
ERM	Entity-Relationship-Model
ETL	Extraction Transformation Loading
ELT	Extraction Loading Transformation
IT	Informationstechnik

Ausgangssituation

Der Zoo Pirmasens befindet sich bereits 36 Jahre im Betrieb. Im Zuge einer Modernisierung wird nun auch die Datenverwaltung digitalisiert. Neben der Erstellung eines operativen Datenbanksystems umfasst dies auch die Implementierung eines Datawarehouses mit IT-Architektur.

Bisher gibt es kein IT-System im Zoo, die bisherige Dateneingabe wurde komplett analog abgewickelt. Das IT-System wird somit neu angeschafft, zur Betreuung des Systems werden 2 IT-Fachkräfte eingestellt, aktuell arbeiten diese noch nicht im Zoo. Die Dateneingabe erfolgt ausschließlich durch nicht IT-Affines Fachpersonal, sprich die Tierpfleger. Das IT-Fachpersonal ist lediglich für die administrative Betreuung der IT-Systems zuständig.

Für die Migration der Altdaten die in analoger Form vorliegen wurde ein externer Anbieter beauftragt. Die Dokumente werden abfotografiert und durch Eingabeprofessionals abgetippt. Die digitalisierten Daten aus der Altdatenmigration fließen ins Datawarehouse und finden keine Anwendung im Operativen System.

In das operative System kommen die aktuellen Daten. Hierzu gehörten Datensätze zu ca. 70 Mitarbeitern, ca. 50 freiberuflich angestellte Ärzte, ca. 120 Lieferanten und ca. 6000 Tiere.

Aktuell betreibt der Zoo noch kein Data Quality System. Alle Maßnahmen starten demnach von Null.

Mittel und Langfristig möchte der Zoo gerne einen Onlineshop betreiben, in dem virtuelle Rundgänge möglich sein sollen und Tierpatenschaften abgeschlossen werden können. Weiterhin ist es geplant Interviews mit den Besuchern durchzuführen, bei denen Sie ihre Erfahrungen sowie Feedback zum Zoo mitteilen können.

Zielsetzung

Die operativen Daten des Zoos müssen den höchsten Aktualisierungsgrad besitzen der möglich ist. Im Zoo wird mit gefährlichen Raubtieren gearbeitet, weshalb die Dokumentation des Aufenthaltsortes jedes einzelnen Tieres von großer Bedeutung ist. Weiterhin werden Daten zur Futterversorgung und Medikamenten im System gespeichert, welche für die Gesundheit der Tiere von außerordentlicher Wichtigkeit sind.

Eine Datenqualität von 100% wäre im Idealfall anzustreben, allerdings ist diese aufgrund von Einschränkungen wie das gleichzeitige entladen mehrerer LWWs auf dem Hof, was zu Verschiebungen bei der Buchung des Bestands führt, nicht zu realisieren. Demnach wird eine Datenqualität mit einem Level von 97% angestrebt.

Die Altdatenmigration

Die Altdaten liegen analog vor. Die Migration wird von einem externen Anbieter vollzogen. Hier werden Dokumente abfotografiert und Daten in das System eingetippt. Hierbei kann die gesamte Bandbreite an Fehlern auftreten, welche die Datenqualität beeinträchtigen.

Durch den gewählten Migrationsprozess an sich können bereits Datenqualitätsprobleme auftreten. Dokumente könnten beschädigt sein oder durch verblassen der Tinte über die Zeit nicht für einen Scanner oder eine Kamera lesbar sein. Weiterhin könnten Seiten fehlen oder beschädigt sein, was zu Unvollständigkeit der Daten führen würde. Gegen derartige Probleme gibt es keine technische Lösung.

Zudem könnten die Dateneingabe Profis Tippfehler begehen. Dies könnten man durch ein Vieraugenprinzip prüfen. Weiterhin wäre der Einsatz von Texterkennungssoftware, als Ersatz für die manuelle Eingabe eine Alternative.

Es besteht bei handschriftlich verfassten Dokumenten die Gefahr, dass die Fachkräfte die Schrift nicht lesen können. Hierfür gibt es ebenfalls keine technische Lösung. Falls der ehemalige Verfasser der Daten noch im Zoo beschäftigt ist, könnte man diesen mit der Eingabe beauftragen.

Die Migration der Altdaten wird durch eine externe Firma durchgeführt. Die dort tätigen Mitarbeiter sind nicht mit den Fachbegriffen aus dem Zoo vertraut. Deshalb besteht die Gefahr, dass implausible Daten nicht als solche erkannt werden, da das Fachwissen für eine derartige Überprüfung fehlt. Eine Lösung für dieses Problem wäre die zusätzliche Überprüfung der eingegebenen Daten durch eine Fachkraft des Zoos.

Auch bei reibungslosem Ablauf des Prozesses der Altdatenmigration, besteht die Gefahr das die Inhalte der digitalisierten Dokumente zu Qualitätsproblemen bei den Daten führen.

Der Eingabe der Altdaten lag kein definierter Geschäftsprozess zugrunde, weshalb Daten in uneinheitlicher Form zu erwarten sind. Uneinheitlichkeit in Format, Schreibweise, Messgrößen und Granularität können nachträglich im ETL-Prozess vor dem Laden ins Datawarehouse angepasst werden. Eine Vereinheitlichung direkt bei der Eingabe wäre allerdings effektiver, in dem Wertebereiche und Feldformate eingesetzt werden.

Es wäre möglich das Daten mehrfach im Archiv als Kopien vorliegen, was bei der Eingabe der Daten zu Duplikaten führen würde. Hier wäre es möglich die Daten im ETL-Prozess vor dem Laden ins Datawarehouse zu bereinigen. Alternativ können Dokumente bereits vor der Digitalisierung auf Duplikate geprüft werden.

Das Operativsystem

Die Daten im neuen operativen System sollen durch die Zoo Mitarbeiter (Tierpfleger) eingegeben werden. Dies birgt viele Gefahren für die Datenqualität.

Da es bis Dato keine digitalen Systeme im Zoo gibt, ist es möglich die IT-Systeme von Anfang an, im Sinne einer bestmöglichen Datenqualität zu entwickeln und dementsprechend auszurichten.

Durch vielfältige Ursachen kann Datenverlust entstehen. Beispielsweise können Nutzer die Daten versehentlich löschen oder es kann ein Stromausfall entstehen. Das regelmäßige erstellen von Backups in der operativen Datenbank bietet die Möglichkeit derartige Schäden zu minimeren. Die Backups sind von den neu einzustellenden IT-Fachkräften zu verwalten. Die Daten im Zoo sollen möglichst in Echtzeit vorhanden sein, daher ist es sinnvoll einen kurzen Zyklus für die Backups einzustellen. So soll z.B. alle 20 Minuten ein Backup erstellt werden. Die Folge wäre, dass maximal die Arbeit von 20 Minuten verloren gehen kann. Dies wäre den Mitarbeitern auch an Aufwand im Fall eines Datenverlustes zumutbar.

Die Mitarbeiter des Zoos sind für die Datenpflege zuständig, was ein hohes Fehlerpotential durch menschliches Fehlverhalten bietet. Zudem sind die Mitarbeiter keine IT-Fachkräfte, was die Fehleranfälligkeit nochmals steigert. Zur Begrenzung der möglichen Fehleingaben ist eine Limitation der technisch möglichen Eingaben sinnvoll. Dies könnte durch den Einsatz von Referenztabellen erreicht werden. Durch die Referenztabelle sind nur bereits vorgegebene Werte anwählbar, was die Fehleranfälligkeit reduziert. Die Eingabe abweichender Werte ist durch das System nicht zulässig. Eine Referenztabelle kann auf zwei Weisen implementiert werden. Die erste Möglichkeit ist die Implementierung der Referenztabelle im Entity-Relationship-Modell (ERM). Die Tabelle wird in diesem Fall mit der Tabelle für die sie Werte bereitstellt mittels Fremdschlüssel verbunden. Dies birgt den Nachteil, dass das ERM erheblich vergrößert wird. Weiterhin ist die Referenztabelle auch im Datawarehouse-Model zu berücksichtigen. Bei der zweiten Möglichkeit kann die Referenztabelle ohne Beziehungen im ERM inkludiert werden und als Look-Up-Tabelle für Drop-Down-Felder in der Eingabemaske dienen. Diese zweite Option führt zum gleichen Ergebnis bzgl. Datenqualität aber vereinfacht das DWH-Modell.

Weiterhin können die Datenbanktabellen auf Wertebereiche begrenzt werden. Hierbei wird verhindert, dass implausible Daten eingetragen werden. So könnte man beispielsweise ein Geburtsdatumsfeld anlegen. Hierbei ist es im Jahr 2023 nicht möglich, dass ein Mensch vor 1900 geboren wurde. So könnte man das Feld auf >1900 begrenzen. Das gleiche Prinzip wäre bei weiteren Attributen wie Gewicht, Menge, Größe möglich.

Zudem führt die Normalisierung zu einer höheren Konsistenz und damit zu einer besseren Datenqualität. So gibt es beispielsweise bei einem Adressfeld die Möglichkeit von Inkonsistenzen, wenn Hausnummer und Straße im gleichen Feld gespeichert werden. z.B. Musterstraße 83 oder 83 Musterstraße. Derartige Inkonsistenzen könnten Duplikate begünstigen, da nicht erkannt werden konnte, dass sich die beiden Eintragungen auf die gleiche Information beziehen.

Neben dem Aufbau der operativen Datenbank bietet auch die Auswahl der Hardware Vorteile für eine kontinuierlich hohe Datenqualität.

Die Datenpflege soll parallel zu den Tätigkeiten der Mitarbeiter durchgeführt werden können. Mobile Devices wie Barcode Scanners oder Handy-Apps bieten sich hierfür an. Ein mögliches Beispiel wäre die Dokumentation einer Arztbehandlung. In der Datenbank ist in diesem Fall die Tier ID und die Arzt ID zu dokumentieren. Jedes Tier (z.B. mittels Halsband) und jeder Arzt (z.B. mittels Zoo-Ausweis) erhält einen Barcode. Der Zuständige Mitarbeiter würde die Barcodes bei der Durchführung der Behandlung scannen und die passenden ID-Nummern würden dem Datensatz hinzugefügt werden. Dies würde zum einen eine Datenpflege in Echtzeit ermöglichen, zum anderen bietet der Einsatz von mobilen Geräten auch die Minimierung von Eingabefehlern. Wichtig ist, dass die Barcodes und die Scanner im Vorfeld richtig eingestellt werden, damit es keine Möglichkeit gibt, damit der Mitarbeiter versehentlich die falsche ID ins System einträgt.

Auch bei Einsatz aller möglichen Hardware- und Software-Lösungen zur Verbesserung der Datenqualität, ist das System noch anfällig für Datenqualitätsverluste durch die Mitarbeiter die im Einsatz sind.

Derzeit sind keine Geschäftsprozesse und somit auch keine Data Quality Maßnahmen definiert. Bei der Neuanschaffung des IT-Systems könnten somit auch neue Geschäftsprozesse geschaffen werden, die im Sinne der Datenqualität hochwertig sind.

In einem ersten Schritt sollten die Geschäftsprozesse einmal festgelegt sowie schriftlich fixiert werden. Allein hierdurch würde sich schon die Datenqualität verbessern, da die Prozesse nun einheitlich sind, was die Wahrscheinlichkeit erhöht, dass der einzelne Prozess jedes Mal gleich abläuft. Hiermit werden unterschiede in der Einheitlichkeit der Daten vermieden.

Somit sollten überall, wo keine Geschäftsprozesse vorhanden sind, welche geschaffen werden. Hierfür benötigen die Mitarbeiter eine ausreichende Schulung, damit die Einhaltung der Prozesse sichergestellt wird.

Ändert sich ein Prozess nur geringfügig durch die neuen IT-Systeme, so kann er entsprechend angepasst werden, damit datenqualitätsrelevante Themen berücksichtigt werden. So würde sich der Prozess der Tierarztbehandlung (z.B. Termin mit Tierarzt vereinbaren -> Tier in ein leeres Gehege bringen -> Tierarzt zum Gehege bringen -> Behandlung Dokumentieren -> usw.) nur geringfügig durch den Einsatz der neuen IT ändern. Der Schritt „Behandlung dokumentieren“ käme hier noch hinzu. Dieser Subprozess müsste demnach neu definiert werden. Bei der Definition des Subprozess könnten bereits Schritte zur Datenqualitätsverbesserung mit einfließen. Beispielsweise könnte geprüft werden, ob der Datensatz bereits in der Datenbank existiert, um so Duplikate zu vermeiden. Diese Prüfung kann manuell oder auch automatisch durchgeführt werden. Ein manueller Prüfvorgang wäre beispielsweise, dass der Mitarbeiter zuerst nach einer Kombination aus Tier ID und dem heutigen Datum sucht, bevor er anfängt einen neuen Datensatz anzulegen. Ein automatischer Prüfvorgang wäre, wenn das System eine derartige Prüfung selbst durchführt nachdem der Mitarbeiter die Daten eingegeben hat. Dies würde das Anlegen von duplizierten Daten verhindern.

Mitarbeiter die das System nutzen sind vorab umfangreich zu schulen. Zum einen sollten Sie sich mit der Technik vertraut machen, zum anderen werden Sie auf die Wichtigkeit der Datenqualität hingewiesen. Datenfehler, die durch eine falsche Nutzung des Systems zurückzuführen sind, können so vermieden werden. Weiterhin wird eine höhere Akzeptanz für datenqualitätsbezogene Schritte gewonnen.

Durch den Einsatz von mobilen Endgeräten wird das Potential von Datenqualitätsverlust durch menschliches Fehlverhalten reduziert. Jedoch kann nicht alles per Barcodescan abgewickelt werden. Einige Daten sind noch händisch einzutragen. Ein großes Problem ist hierbei, dass Datensätze doppelt in die Datenbank eingetragen werden. Neben der Definition der Geschäftsprozesse ist es auch notwendig klare Verantwortungsbereiche zu definieren, um damit möglichst viele vorhersehbare Datenqualitätsthemen abzufangen. So wäre beispielsweise die Ausarbeitung eines Dienstplans für die Datenpflege wichtig oder die Regel, dass der Mitarbeiter der gerade ein TO DO ausführt auch gleich die dazugehörige Datenpflege erledigt, sinnvoll.

Ein weiteres wichtiges Werkzeug für die Datenqualität ist das Vier-Augen-Prinzip. Auch wenn klare Verantwortlichkeiten zu einer besseren Datenqualität führen, ist der Faktor Mensch nach wie vor eine Gefahr für die Datenqualität. Deswegen ist das Vier-Augen-Prinzip als zusätzliches Werkzeug bei den Geschäftsprozessen zu inkludieren, mit der Zielsetzung das eben der zweiten Person der Fehler auffällt, welcher der ersten entgangen ist.

Werden Daten nicht schriftlich aufgenommen, so ist es möglich das diese von der zu bearbeiteten Person falsch verstanden oder fehlinterpretiert werden. Das Risiko steigt noch, wenn mit fremdsprachlichen und fachspezifischen Daten gearbeitet wird. Dies betrifft im Zoo die Lieferanten und die Tierarztdaten. Damit mögliche Fehler bei der Dateneingabe vermieden werden, empfiehlt es sich immer die mündlich aufgenommenen Daten noch schriftlich per E-Mail bestätigen zu lassen, bevor diese in die Datenbank aufgenommen werden.

Die Prüfung auf Plausibilität ist ein weiterer wichtiger Punkt für die Datenqualität. Um zu prüfen, ob ein Datensatz mit einem gewissen Kontext plausibel ist, werden Kenntnisse benötigt, die das Datenbanksystem nicht liefern kann. Deshalb muss die Prüfung auf Plausibilität als manueller Schritt dazu genommen werden. Diesen Schritt haben die Mitarbeiter bei jedem Dateneingabeprozess zu beachten.

Trotz der oben genannten Maßnahmen ist es möglich, dass ein Datensatz bezüglich Formats, Wertebereichs und der Plausibilität korrekt scheint, aber dennoch nicht korrekt ist. Hier besteht in manchen Fällen die Möglichkeit den Datensatz mit externen Quellen abzugleichen. So könnte eine Lieferadresse mit der Schufa Auskunft verglichen oder die Mitarbeiteradresse mit dem Personalausweis nachgeprüft werden.

Der ELT-Prozess (Datawarehouse)

Ziel ist es künftig die Altdaten und die Daten aus dem operativen System in ein Datawarehouse zu übertragen, um so Analysen über Tiere, Lieferanten usw. durchführen zu können. Hierfür werden die Daten aus den relevanten Quellen extrahiert und ins DWH geladen. Hierbei findet eine Transformation der Daten statt, damit die Daten in der gewünschten Form vorliegen, damit die Analysen durchgeführt werden können.

Um verschieden Werte im DWH miteinander vergleichen zu können oder Werte aggregieren zu können, werden gleiche Einheiten benötigt. Im operativen System des Zoos liegen die Werte für Futter allerdings in verschiedenen Einheiten vor. So sind für Lieferant A die Werte in Kisten im System hinterlegt, während für Lieferant B die Werte in Säcken eingepflegt sind. Da die Werte so nicht vergleichbar sind, müssten diese in der gleichen Einheit vorliegen, also sprich in Kilogramm umgewandelt werden. Weiterhin sollten die Einheiten für Größe und Gewicht der Tiere sowie die Währung der Bestellungen vereinheitlicht werden.

Der Transformationsprozess könnte auch eine Quelle für Datenqualitätsprobleme werden. Beispielsweise könnte bei der Transformation von Preisen der falsche Wechselkurs benutzt werden, was zu falschen Daten im DWH und damit auch zu falschen Analysen führen würde. Damit dieses Risiko reduziert wird, sollten auch im Transformationsprozess Referenztabellen benutzt werden.

Die neuen IT-Systeme beinhalten die Möglichkeit für Datenanalysen, weiterhin werden auch neue Geschäftsprozesse erstellt. Die Daten in den alten Dokumenten gehen möglicherweise mit alten Geschäftsprozessen einher, die anderen Daten als die heutigen Prozesse benötigen. Zudem liegen die Altdaten auch in einem anderen Format vor. Für das operative System ist dies nicht relevant, aber im Data Warehouse könnte die Zusammenführung von alten und neuen Daten zu Inkonsistenzen führen. Beispielsweise wurden früher für Mitarbeiter keine E-Mailadressen oder Handynummern gespeichert, was zu leeren Feldern im DWH führt, wenn diese Daten zusammengelegt werden. Es sollte eine Überprüfung von Leerfeldern ins DWH eingeführt werden, da diese Felder in diesem Fall nicht fehlerhaft sein müssen.

Auch das Entstehen von Duplikaten ist hier möglich, z.B. bei Tieren und Mitarbeitern, bei denen sowohl Altdaten als auch neue Daten im operativen System existieren. Hier wäre es Sinnvoll eine Duplikats-prüfung in die Geschäftsprozesse einzubauen, damit diese erkannt und entfernt werden.

Trotz aller Bemühungen eine hohe Datenqualität bei der Eintragung ins operative System zu erzielen, sind Fehler bei der Dateneingabe nicht komplett vermeidbar. Für das Aufdecken von Formatverstößen, können Regular Expressions eingesetzt werden. Hiermit können Inkonsistenzen, Fehler und Informationsdefizite in den Daten gefunden und behoben werden.

Falls die verschiedenen Quellen unterschiedliche Encodings haben, so sollte sichergestellt werden, dass das gleiche Encoding für die Daten benutzt wird. Dies ist wichtig, damit alle Sonderzeichen erkannt werden können.

Das DWH speichert die Daten historisiert. Hierfür wird ein Datum für den Anfang und für das Ende der Gültigkeit eines Datensatzes mitgespeichert. Das von-Datum sollte immer das Datum des Loading ins DWH geladen werden und als bis-Datum immer ein Dummy Wert, der weit in

der Zukunft liegt. Erst wenn die neue Version eines Datensatzes geladen wird, was den alten ungültig macht, soll das bis-Datum auf das Datum dieses neuen Loadings geändert werden. Hierdurch wird eine lückenlose Historie der Daten ermöglicht.

Fazit

Der Zoo Pirmasens wird nun umfangreich modernisiert. Bestandteil dieser Modernisierung ist auch das Digitalisierungsprojekt, bei welchem ein IT-System mit operativem Datenbanksystem und Datawarehouse implementiert wird.

Bis Dato arbeitete der Zoo weder IT gestützt, noch gab es Maßnahmen zu Data Quality. Demnach bedeutet dies einen großen Schritt für den Zoo, welcher jedoch notwendig ist, um im aktuellen Geschäftsalltag mithalten zu können und wettbewerbsfähig zu bleiben. Für den Zoo ist insbesondere eine hohe Aktualität der Daten und ein mit 97% sehr hohes Qualitätslevel vonnöten, um den künftigen Geschäftsalltag ordentlich und zeitgemäß bestreiten zu können.

Diese Lektüre soll als Guideline dienen, um Data Quality fachgerecht als Konzept im Zoo Primasens implementieren zu können. Mit den beschriebenen Empfehlungen zu den Bereichen Altdatenmigration, Operativsystem und ELT-Prozess (Datawarehouse) kann eine IT-Infrastruktur implementiert werden, die ein zum Zoo Pirmasens passendes Qualitätslevel abbildet.

Ist das IT-System erstmal implementiert, können darauf aufbauend weitere Schritte, wie der geplante Onlineshop mit virtuellen Rundgängen sowie der Abschluss von Tierpatenschaften, erfolgen. Mit der dargestellten Lektüre hat der Zoo das Potential sich IT-gestützt langfristig und erfolgreich weiterzuentwickeln.