

Integrationsnachgelagertes Datenmanagement am Beispiel des Gesundheitswesens

Ralph Stuber¹, Hans-Jürgen Appelrath²

¹ OFFIS Institut für Informatik - Bereich Gesundheit
stuber@offis.de

² Carl von Ossietzky Universität Oldenburg - Abteilung Informationssysteme
appelrath@informatik.uni-oldenburg.de

Abstract: Um Synergieeffekte aus der Verschneidung verschiedener Datenbestände nutzen und auf diese Weise neue Informationen gewinnen zu können, werden heutzutage zunehmend integrierte Datenbestände beispielsweise in Form von Data Warehouses erzeugt. Häufig erfolgt die Erzeugung integrierter Datenbestände jedoch nicht durch die Urheber der verschiedenen Datenquellen selbst, sondern durch andere Personen, Institutionen oder Dienstleister. In solchen Fällen ergibt sich für die jeweiligen Quelldatenurheber oftmals erst nach Veröffentlichung des integrierten Datenbestandes eine Möglichkeit zur Einsichtnahme und Validierung der von Ihnen verantworteten Daten, so dass ihnen das klassische Vorgehen zur Durchführung von Datenqualitätsmanagement (DQM) im Rahmen der ETL-Prozesse [Inm92, BG01] nicht ermöglicht wird. Vielmehr kann ein Bedarf an integrationsnachgelagertem Datenmanagement (INDM) identifiziert werden.

Zur Deckung dieses Bedarfs wird das Vorgehensmodell VD2M (Vorgehensmodell zur *Delegation von Datenmanagement*) als neuartiges Konzept für den Umgang mit qualitativen Unzulänglichkeiten der Daten in integrierten Datenbeständen und Data Warehouses entwickelt. Es gibt den Urhebern der Quelldaten die Möglichkeit, integrationsnachgelagert Datenmanipulationen zu initiieren und somit die Datenqualität integrierter Datenbestände und Data Warehouses auch nach Abschluss der ETL-Phase zu erhöhen. Am Beispiel der Weissen Liste, eines Internet-Portals zur Darstellung der Qualität der Versorgung durch Krankenhäuser in Deutschland, wird aufgezeigt, wie eine werkzeuggestützte Umsetzung von VD2M zur Lösung der dargestellten Problemstellung beitragen kann.

1 Motivation von integrationsnachgelagertem Datenmanagement

Um Potentiale der in verschiedenen Datenbeständen enthaltenen Informationen nutzen zu können, indem beispielsweise zusätzliche Informationen durch Synergieeffekte gewonnen werden, werden diese oftmals aus verschiedenen Quellen unter Einsatz von Data Warehouse (*DWH*)-Technologien [Inm92] integriert und so in einem gemeinsamen Datenbestand zusammengeführt. Die auf diese Weise zusammengetragenen Kennzahlen und Informationen ermöglichen es den Institutionen und Unternehmen unterschiedlicher Branchen, Analysen zum Zwecke der Unterstützung in Management-Entscheidungen durchzuführen, oder die Informationen z.B. zur Präsentation von Produkten,

Dienstleistungen und Unternehmensstatistiken in aufbereiteter Form beispielsweise über das World Wide Web zu nutzen.

Sollten Datenqualitätsmängel in den Datenquellen vorliegen, so werden diese in der Regel in einer Phase des Datenqualitätsmanagements (*DQM*) behoben. Der klassische Ansatz des *DQM* ist im DWH-Prozess in der sog. *ETL*-Phase [BG01] verortet. Diese Phase umfasst sowohl die Erfassung der Daten aus den Quellen (*E*) als auch deren Transformation zur Vereinheitlichung und Qualitätssicherung (*T*) sowie das Laden der vereinheitlichten, qualitätsgesicherten Daten in das DWH (*L*). Das *DQM* erfolgt somit im Vorfeld der Erzeugung eines integrierten Datenbestandes, bereits bevor die Daten im DWH zusammengeführt werden. Schreibende Zugriffe auf integrierte DWH-Datenbestände sind nach der in der Praxis etablierten Definition des DWH-Prozesses nach Inmon [Inm92] explizit nicht vorgesehen.

Entgegen dem oben skizzierten klassischen Vorgehen existieren jedoch auch verschiedene Umstände, die eine Durchführung des *DQM* direkt auf dem integrierten Datenbestand, also dem DWH sinnvoll erscheinen bzw. erforderlich werden lassen [Zeh03]. Die Praxis zeigt z.B., dass die funktionellen Rollen Datenurheber einer Datenquelle und Betreiber eines DWH vielfach nicht durch ein und dieselbe Person oder Institution repräsentiert werden; es lassen sich vielmehr Szenarien verteilter Verantwortlichkeiten im Hinblick auf Urheberschaft der Datenquelle und Betrieb des DWH beobachten. In solchen Situationen werden die Datenurheber der Quelldaten selten in den Prozess der Erstellung oben beschriebener Datenbestände durch Dritte eingebunden, so dass sie erst nach Veröffentlichung des DWH die Möglichkeit zur Einsicht erhalten. Sollten sie nun Bedarf an *DQM* identifizieren, so haben sie keine Möglichkeit mehr, Änderungs- und Korrekturwünsche an den Quelldaten in den vorgelagerten und bereits abgeschlossenen Integrationsprozess einzubringen. Entgegen dem klassischen *DQM*-Vorgehen entsteht so der Bedarf an integrationsnachgelagertem Datenmanagement (*INDM*).

Darüber hinaus existieren noch weitere Gründe, die *INDM* auch aus Sicht des DWH-Betreibers motivieren. So ist Kenntnis über die Herkunft bzw. den aktuellen Speicherort der im integrierten Datenbestand enthaltenen Informationen erforderlich, um Korrekturen auf den Datenquellen durchführen zu können [BKT00]. Sind Daten eines integrierten DWH nicht mit Metainformationen über die Herkunft der Datenquellen annotiert, so ist eine Korrektur in den Quellen nicht möglich; Korrekturen oder Manipulationen können nur integrationsnachgelagert am DWH vorgenommen werden. Weiter gestaltet sich die Durchführung der Integrationsprozesse (bedingt durch die integrationsinhärenten Schritte der Beschaffung, Bereinigung, Aufbereitung und Vereinheitlichung der Daten [BG01]) zudem ggf. so aufwändig, dass die Durchführung einzelner Datenkorrekturen in den Quellen im Hinblick auf die nach den Korrekturen erneut zu investierenden Integrationsaufwände nicht verhältnismäßig erscheint. Schließlich existieren integrierte Datenbestände, die zeitbehaftete Datenabzüge der Datenquellen beinhalten. Fallen beispielsweise in den Datenquellen kontinuierlich große Datenmengen an, so werden die Daten ggf. nach Integration ins DWH in den Quellen gelöscht [Zeh03], um unbegrenztem Wachstum der Datenquellen entgegenzuwirken. Eine Korrektur im DWH ist in solchen Fällen somit alternativlos.

2 Bedarf nach integrationsnachgelagertem Datenmanagement am Beispiel der Weissen Liste

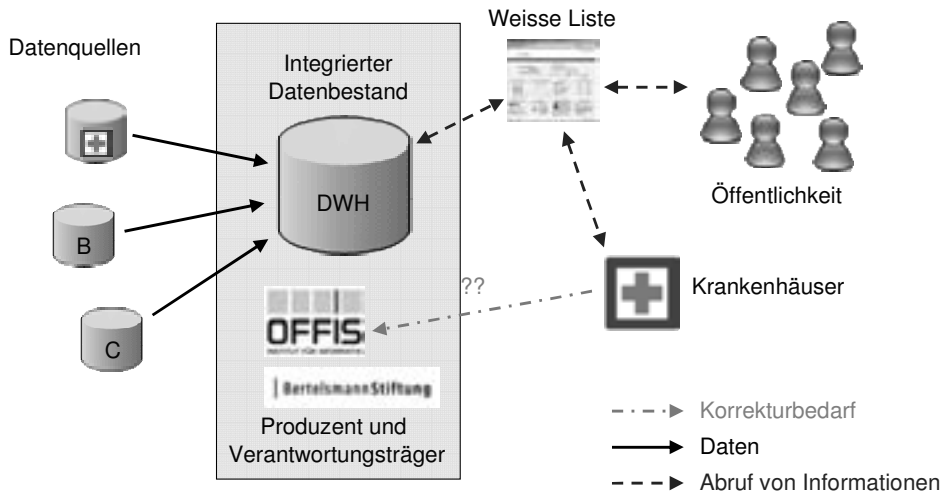


Abbildung 1: Bedarf an integrationsnachgelagertem Datenmanagement unter verteilten Verantwortlichkeiten am Beispiel der Weissen Liste

Abbildung 1 zeigt die zuvor beschriebene Problemstellung am Beispiel der Weissen Liste (WL) [WL10]. Das Projekt WL wurde von der Bertelsmann Stiftung gemeinsam mit den größten Patienten- und Verbraucherorganisationen initiiert und stellt Informationen zur Versorgungsqualität durch Krankenhäuser allgemeinverständlich und öffentlich in einem Web-Portal zur Verfügung. Die Umsetzung der WL erfolgte in Zusammenarbeit mit verschiedenen Dienstleistern, darunter das OFFIS Institut für Informatik als Erzeuger und Betreiber des zu Grunde liegenden DWH. Die WL veröffentlicht u.a. Informationen aus den sog. „Strukturierten Qualitätsberichten“, die nach §301 SGB V¹ von allen in Deutschland zugelassenen Krankenhäusern gesetzlich verpflichtend in einem XML-Format veröffentlicht und an zentrale Annahmestelle außerhalb der WL geliefert werden. Stichtagbezogen erfolgt die Integration der Strukturierten Qualitätsberichte alle 2 Jahre.

OFFIS integriert im Auftrag der Bertelsmann Stiftung die Inhalte verschiedener Datenquellen, und stellt die Ergebnisse im Portal der Öffentlichkeit zur Verfügung. Sollten Krankenhäuser nun einen Datenmanipulationswunsch identifizieren, so kann dieser erst integrationsnachgelagert an OFFIS als Betreiber des integrierten Datenbestandes übermittelt werden, da die Krankenhäuser nicht in den Prozess der Datenintegration involviert sind.

¹ Fünftes Buch Sozialgesetzbuch - Gesetzliche Krankenversicherung - (Artikel 1 des Gesetzes vom 20. Dezember 1988, BGBl. I S. 2477), das zuletzt durch Artikel 3 des Gesetzes vom 17. März 2009 (BGBl. I S. 534) geändert worden ist, URL: http://www.gesetze-im-internet.de/sbg_5/index.html, letzter Besuch: 22.04.2010.

3 Umsetzung von integrationsnachgelagertem Datenmanagement in der Weissen Liste: VD2M und WD2M

Im Rahmen der ETL-Prozesse des Data Warehousing in der WL erfolgt eine Reihe an DQM-Maßnahmen gemäß klassischen Vorgehens bereits während der Integration der Daten aus den Quellen. Automatisiert korrigierbare Datenqualitätsmängel werden so im Vorfeld der Veröffentlichung im Portal behoben. Dazu zählen beispielsweise der Abgleich der Daten gegen die per inhaltlichen Restriktionen auferlegten Anforderungen, oder Datenbereinigungsmaßnahmen wie Typ- und Formatprüfungen, Erkennung von Duplikaten sowie die Behandlung von fehlenden Werten oder Verletzungen referenzieller Integritäten [Stu03]. Neben diesen integrationsvorgelagerten DQM-Maßnahmen existieren jedoch auch Anforderungen hinsichtlich weiterer Maßnahmen zur Qualitätssicherung, die integrationsnachgelagert erfolgen müssen. Um diese auf dem DWH der WL umsetzen zu können, wird das Vorgehensmodell VD2M (Vorgehensmodell zur *Delegation* von *Datenmanagement*) als neuartiges Konzept für den Umgang mit qualitativen Unzulänglichkeiten der Daten entwickelt, das die Durchführung integrationsnachgelagerter Datenmanipulationen ermöglicht.

Vorgehensmodell VD2M

Dem externen Melder eines integrationsnachgelagerten Datenmanagement-Bedarfs - am Beispiel der WL ist dies z.B. ein Krankenhaus - ist der Ort der Speicherung eines realweltlich beschriebenen, zu manipulierenden Datums im Modell des integrierten Datenbestandes unbekannt. Als Lösungsansatz wird in VD2M ein Suchindex (SI) eingeführt, der eine semantische Annotation der Relationen und Attribute mit Texten zur realweltlichen Beschreibung ermöglicht. Das Ziel besteht in der Ermöglichung einer Stichwortsuche, um modellierte Strukturen auf Basis der realweltlichen Beschreibung auffinden zu können. Um einen solchen Suchindex aufbauen zu können, bedarf es der Erfassung der Struktur und ggf. potentiell im Datenschema enthaltener Semantik wie z.B. Kommentartexten oder Relationen- bzw. Attributbezeichnungen. Nach der grundlegenden Generierung des Suchindexes ist anschließend eine zusätzliche manuelle semantische Annotation der erzeugten Einträge durch einen Domänenexperten des Datenschemas erforderlich, um zusätzliche Semantik in den Suchindex einzupflegen.

Ein weiteres Problem besteht in dem Umstand, dass dem externen Melder eines integrationsnachgelagerten Datenmanagement-Bedarfs auch die Kombination der Primärschlüsselbelegungen der zu manipulierenden Tupelinstanz unbekannt ist, die zudem vom Melder nur durch eine sprachliche Beschreibung der Realwelt-Eigenschaften spezifiziert werden kann. Als Lösungsansatz sieht VD2M die Einführung einer Suchfunktionalität vor, die sämtliche Attributausprägungen der zuvor bestimmten Relation sowie deren Elternrelationen auf Vorkommen eines anzugebenden Stichwortes untersucht. Das Ziel besteht in der Ermöglichung einer Stichwortsuche anhand realweltlicher Beschreibung zur Spezifikation der Primärschlüssel-Belegungen. Zur Umsetzung bedarf es der Erfassung der Referenzierungshierarchien, damit ausgehend von einer gefundenen Startrelation alle Elternrelationen in den Suchprozess zur Bestimmung von Primärschlüsselbelegungen sukzessive einbezogen werden können.

Sollen nun Datenmanipulationen durchgeführt werden, so müssen alle Änderungen am Datenbestand nachvollziehbar sein, um Reversibilität zu ermöglichen. Als Lösung sieht VD2M die Einführung einer Historisierungsinfrastruktur (HI) in Form eines Abbildes des Datenmodells des integrierten Datenbestandes vor, in welche die enthaltenen und zu manipulierenden Originalbelegungen - erweitert um Zeitstempelung und Dokumentationsverweise - geordnet und nachvollziehbar abgelegt werden können. Das Ziel besteht in der zeitannotierten Speicherung von ursprünglichen Belegungen beliebiger Relationen und Attribute. Zur Umsetzung bedarf es der Erfassung der Struktur in Form von Relationen- und Attributbezeichnern, Wertebereichen und Vernetzungshierarchien zum initialen Aufbau der Historisierungs-Infrastruktur.

Damit das Vorgehensmodell eine weitgehend generische Anwendbarkeit auf beliebige relationale Datenbestände aufweist, erfolgt zudem die Einführung einer Phase der initialen Konfiguration zur Anpassung aller schemaspezifischen Informationen, die zum Aufbau von Suchindex und Historisierungsinfrastruktur benötigt werden. Das Ziel dieser Phase besteht darin, eine strukturierte Anpassung der zum Vorgehen erforderlichen Informationen unter möglichst hohem Automatisierungsgrad erreichen zu können. Schließlich muss eine Berücksichtigung struktureller Veränderungen der zugrunde liegenden Datenbankschemata erfolgen, da Datenquellen sich im Laufe der Zeit strukturell verändern können. Als Lösungsansatz für dieses Problem erfolgt eine teilanaloge Nutzung der Funktionalitäten aus der Konfigurationsphase zur Restrukturierung und Anpassung an strukturelle Änderungen mit dem Ziel, eine strukturierte Anpassbarkeit der zum Vorgehen erforderlichen Informationen unter möglichst hohem Automatisierungsgrad und insbesondere unter Erhalt vorhandener Informationen wie z.B. Suchindex-Annotationen und bereits historisierter Datentupel zu erreichen.

Zusammenfassend können in VD2M sechs Phasen identifiziert werden. Abbildung stellt das Vorgehensmodell VD2M dar.

1. **Quellenanalyse:** Sammeln struktureller Informationen über den Datenbestand.
2. **Infrastruktur-Generierung:** Erzeugung einer Infrastruktur zur Historisierung der Instanzen.
3. **Suchindex-Generierung:** Semantische Indexierung des Datenbestandes (automatisierte Erzeugung, nachgelagert manuelle Anreicherung).
4. **Assistiertes Datenmanagement:** Ermöglichung der Durchführung des Datenmanagements mit Historisierung durch Manipulationswerkzeug.
5. **Dokumentation:** Dokumentation der durchgeführten Datenmanagement-Schritte.
6. **Reorganisation:** Erweiterung/Anpassung des Suchindexes an Struktur-Änderungen.

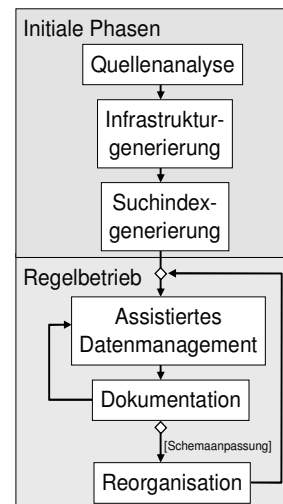


Abbildung 2: VD2M

Werkzeuge WD2M

Zur Anwendung des vorgestellten Vorgehensmodells VD2M im Kontext des Projekts Weisse Liste wird eine Softwarewerkzeug-Sammlung WD2M (Werkzeuge zur Delegation von Datenmanagement) implementiert. Die Implementierung gliedert die beiden Bereiche zur Konfiguration des Datenbestandes sowie zum Regelbetrieb in zwei getrennte Werkzeuge. Das Konfigurationswerkzeug setzt dabei die Aufgaben um, die in den Phasen der Quellenanalyse, der Suchindex- und Historisierungsinfrastruktur-Erzeugung sowie der Reorganisation anfallen. Das Werkzeug zur Umsetzung der Phasen des Regelbetriebes ermöglicht die Durchführung und Dokumentation von assistiertem Datenmanagement. Durch den Einsatz der Werkzeuge ist es den Krankenhäusern im Kontext der Weissen Liste nun möglich, die Durchführung integrationsnachgelagerter Datenmanagement-Anforderungen den Restriktionen des G-BA folgend nachvollziehbar und reversibel ohne Kenntnisse der Struktur des DWH der WL zu delegieren.

Zusammenfassung und Ausblick

Dieser Beitrag motiviert zunächst den Bedarf an integrationsnachgelagertem Datenmanagement. Am Beispiel des Projektes Weisse Liste, in dem zudem verteilte Verantwortlichkeiten hinsichtlich der Rollen Datenurheber und DWH-Betreiber vorliegen, wird zur Lösung das Vorgehensmodell VD2M vorgestellt, welches es externen Quelldatenverantwortlichen erlaubt, Datenmanagement-Bedarfe an den Betreiber des DWH zu delegieren und diese dokumentiert und reversibel durchzuführen.

Ausblickend ist zu untersuchen, wie die einzelnen Phasen des Vorgehensmodells optimiert werden können, um eine möglichst hohe Abdeckung erfolgreich durchgeführter Datenmanagement-Anfragen erreichen zu können. Die Teilprobleme der Identifikation des Speicherungsortes sowie die Bestimmung einer Tupelinstanz anhand einer natürlichsprachlichen Beschreibung der Realwelt sind hier vertieft zu betrachten.

Literaturverzeichnis

- [BG01] Bauer, A. and Günzel, H. 2001. „Data Warehouse Systeme. Architektur, Entwicklung, Anwendung“. Dpunkt Verlag. ISBN 3932588762.
- [BKT00] Buneman, Peter and Khanna, Sanjeev and Tan, Wang Chiew 2000. “Data Provenance: Some Basic Issues”. In: FSTTCS, Seiten 87-93.
- [Inm92] Inmon, William H. 1992. “Building the Data Warehouse”. John Wiley & Sons, Inc., New York, NY, USA. ISBN 0471569607.
- [Stu03] Stuber, Ralph 2003. “Data Preprocessing - Datenvorverarbeitungsschritte des Prozessmodelles“. Technischer Bericht. URL: <http://www-is.informatik.uni-oldenburg.de/publications/2954.pdf>, letzter Besuch: 22.04.2010.
- [WL10] „Weisse Liste - Patienten unabhängig informieren“. URL: <http://www.weisse-liste.de/>, letzter Besuch: 22.04.2010.
- [Zeh03] Zeh, T. 2003. „Data Warehousing als Organisationskonzept des Datenmanagements: Eine kritische Betrachtung der Data-Warehouse-Definition von Inmon“, Inform., Forsch. Entwickl., Volume 18 , Nummer 1, 32-38.