Dissertation Dipl.-Inform. Ralph Stuber

Titel: Integrationsnachgelagertes Datenmanagement in Data Warehouses unter Berücksichtigung verteilter Verantwortlichkeit

Tag der Disputation: 12.10.2011

Erstgutachter: Prof. Dr. Dr. h.c. H.-Jürgen Appelrath

Zweitgutachter: Prof. Dr. Reinhold Haux

Um die aus der gezielten Integration verschiedener Datenbestände entstehenden Synergieeffekte zu nutzen und so zusätzliche Informationen zu gewinnen, werden zunehmend integrierte Datenbestände beispielsweise in Form von DataWarehouses erzeugt, die sich aus verschiedenen Datenquellen speisen. Die so gewonnenen Informationen dienen verschiedenen Anwendungsfeldern, so z.B. zur Veröffentlichung von Informationen im Intranet oder imWorldWideWeb, oder zur Durchführung von Analysen, deren Ergebnisse zur Entscheidungsunterstützung verwendet werden können.
Häufig erfolgt die Erzeugung integrierter Datenbestände jedoch nicht durch die Urheber der verschiedenen Datenquellen selbst, sondern durch andere Personen, Institutionen oder Dienstleister, so dass Szenarien verteilter Verantwortlichkeit entstehen. Dadurch bedingt ergibt sich für die jeweiligen Quelldatenurheber oftmals erst nach Veröffentlichung der Daten des integrierten Datenbestands eine Möglichkeit zur Einsicht in solche Datenbestände. Haben Quelldatenurheber nun ein Interesse einer „korrekten“ Darstellung der von ihnen verantworteten Daten in fremdverwalteten integrierten Datenbeständen, so entfällt für sie das klassische Vorgehen zur Durchführung von Datenqualitätsmanagement im Data Warehousing im Rahmen üblicher ETL-Prozesse (Extraktion, Transformation, Laden), da sie durch die fehlende Einbeziehung in den Aufbau der integrierten Datenbestände keine Möglichkeit der Einbringung von Änderungs- oder Korrekturwünschen im vorgelagerten Integrationsprozess haben. Auch andere Umstände motivieren Datenqualitätsmanagement auf dem bereits integrierten Datenbestand, so z.B. eine potentielle Aufwandseinsparung gegenüber einer erneuten vollständigen Quelldatenintegration, fehlende Kenntnis über den Ursprung der Daten (Data Provenance) oder eine nachträglich nicht sichergestellte Datenverfügbarkeit in den Quellen. Insgesamt kann daher ein Bedarf an integrationsnachgelagertem Datenmanagement identifiziert werden.
Zur Deckung dieses Bedarfs wird im Rahmen der vorliegenden Arbeit das Vorgehensmodell VD2M (Vorgehensmodell zur Delegation von Datenmanagement) entwickelt. Es gibt den Urhebern der Quelldaten die Möglichkeit, integrationsnachgelagert Datenmanipulationen direkt am integrierten Datenbestand nachvollziehbar und reversibel zu initiieren und somit die Datenqualität integrierter Datenbestände und Data Warehouses auch nach Abschluss der ETL-Phase zu erhöhen und die Korrektheit der eigenverantworteten Daten in fremdverantworteten integrierten Datenbeständen sicherzustellen.
Am Beispiel eines Internet-Portals aus der Domäne des Gesundheitswesens wird aufgezeigt, wie eine Werkzeug gestützte Umsetzung WD2M (Werkzeuge zur Delegation von Datenmanagement) für das Vorgehensmodell VD2M zur Lösung der dargestellten Problemstellung beitragen kann, indem sie die Delegation integrationsnachgelagerter Datenmanagement-Aktivitäten auf fremdverantworteten Datenbeständen ohne Kenntnis der diesen zu Grunde liegenden informationstechnologischen Strukturen und unter Nutzung natürlichsprachlicher Beschreibungen der Anforderungen ermöglicht.