Konsolidierte Datenmodellierung von Versorgungsdaten mit dem Entity-Attribute-Value-Modell und Data Vault

Consolidated data modeling of health services research data with the entity-attribute-value model and data vault

Jens Rauch[1]
Jan-Patrick Weiss[1,2]
Frank Teuteberg[2]
Ursula Hübner[1]

1 Hochschule Osnabrück, Fakultät Wirtschafts- und Sozialwissenschaften, Forschungsgruppe Informatik im Gesundheitswesen, Osnabrück, Deutschland

2 Universität Osnabrück, Institut für Informationsmanagement und Unternehmensführung (IMU), Osnabrück, Deutschland

Abstract

Distributed and heterogeneous data must be integrated for health services research in a way, which is open to new requirements and easily expandable for new data sources. For data integration in the health services research domain, mainly data warehouses have been used that model data either as a dimensional or an entity-attribute-value (EAV) model. However, these data models are either not flexible enough or lack data management capabilities, which makes longitudinal data analyses more difficult. We have extended the EAV approach with data vault modelling and hereby modelled the data structures of the hospital quality reports by the Gemeinsamer Bundesauschuss (G-BA) and integrated data from the years 2011 to 2015 accordingly. This makes it possible to historicise metadata of features, in particular those of quality indicators, and establishes a high degree of extensibility towards new heterogeneous data sources. The proposed approach allows a free selection of the abstraction level for the entities to be modelled, so that a completely generic EAV model with historicised metadata can be created.

Keywords: health services research, data warehouse, data collection, common data elements, information storage and retrieval

Zusammenfassung

Für die Versorgungsforschung ist wichtig, dass verteilte und heterogene Daten so integriert werden, dass sie offen für neue Analyse-Anforderungen und leicht um neue Datenquellen erweiterbar sind. Für die Integration von Versorgungsdaten werden bislang hauptsächlich Data-Warehouses eingesetzt, die Daten dimensional oder als Entity-Attribute-Value-Struktur (EAV) modellieren. Diese Datenmodelle sind jedoch entweder unflexibel oder weisen ein zu geringes Maß an Datenorganisation auf, was longitudinale Analysen erschwert. Wir haben den EAV-Ansatz um die Data-Vault-Modellierung ergänzt und damit die Datenstrukturen der Krankenhaus-Qualitätsberichte des Gemeinsamen Bundesausschusses (G-BA) modelliert sowie die Daten der Jahre 2011 bis 2015 integriert. Dies ermöglicht eine Historisierung der Metadaten für Merkmale, insbesondere der Qualitätsindikatoren, sowie ein hohes Maß an Erweiterbarkeit gegenüber neuen heterogenen Datenquellen. Der vorgeschlagene Ansatz erlaubt es, den Abstraktionsgrad für die zu modellierenden Entitäten frei zu wählen, so dass auch ein vollständig generisches EAV-Modell mit historisierten Metadaten erstellt werden kann.

Schlüsselwörter: Versorgungsforschung, Data-Warehouse, Datensammlung, gemeinsame Datenelemente, Informationsspeicherung und -abruf

Einleitung

Die rasant wachsende und verteilte Menge an Datenquellen im Gesundheitswesen birgt großes Potential für die Versorgungsforschung [1]. Die zunehmende Digitalisierung der Gesellschaft macht es aber auch für die Forschung einfacher, Versorgungsdaten im Rahmen von Befragungen und Interviews zu erheben [2], [3], [4]. Diese Vielzahl an internen wie externen, verteilten Datenquellen geht einher mit informationstechnologischen Barrieren und einem hohen Grad semantischer Heterogenität [5]. Daraus resultiert ein hoher Bedarf an integrierten Datenbeständen, konsistenter Datenhaltung und strukturiertem Datenmanagement. Denn von besonderer Bedeutung in der Versorgungsforschung ist die Sichtweise auf Versorgungsdaten im Längsschnitt, um Trends und Prädiktoren aufzuspüren [6], [7], [8].

Für Längsschnittbetrachtungen dieser Art müssen Datensätze, die dieselben Informationsobjekte beschreiben, aufeinander abgebildet und ihre Veränderung erfasst werden. Neue Forschungsideen führen dabei oft zu sich ändernden Erhebungs-Items und damit zu sich ändernden Schnittstellen. Hinzu kommt, dass in größeren Forschungsprojekten eine Vielzahl von Wissenschaftlern verschiedene Fragestellungen untersuchen, auch wenn sie eine gemeinsame Datenbasis verwenden [9]. Dabei entstehen oft erst während des Forschungsprozesses neue Ansätze, Ideen und Fragestellungen, die im Verlauf näher untersucht werden [10].

Daraus leiten sich zwei Anforderungen für das zentrale Datenmodell eines Forschungs-Data-Warehouse ab: Für die laufend neu zu erschließenden Datenquellen ist ein Datenmodell zu entwickeln, das in hohem Maß *erweiterbar* ist und die Integration von Daten aus neuen Datenstrukturen möglichst einfach macht. Andererseits sollte diese Integration in zweifacher Hinsicht *offen* sein. Wenn zu erwarten ist, dass Daten ihre Semantik ändern, weil beispielsweise die Items eines Fragebogens umformuliert werden, ist es wichtig, dass das Datenmodell dies adäquat abbilden kann, ohne dafür gesondert angepasst zu werden (Offenheit gegenüber geänderter Semantik der Datenquellen). Darüber hinaus soll das Datenmodell keine analytischen Entscheidungen vorwegnehmen, etwa weil seine Struktur bestimmte Analysedimensionen vorsieht, die bestimmen, wonach Daten aggregiert werden können. Die Datenbasis sollte daher gegenüber anderen „Sichten“ auf die Daten, neuen Fragestellungen, aber auch ungeplanten Analysen offen sein (Offenheit gegenüber geänderten Analyseanforderungen) [11].

Diese Anforderungen sind typisch für komplexe langfristige Forschungsvorhaben wie das Projekt ROSE [12] in dem neben Versorgungsdaten aus Interviews und Befragungen ein breites Spektrum an externen Daten anfallen. Im Projekt ROSE an der Hochschule Osnabrück fallen neben Versorgungsdaten aus Interviews und Befragungen ein breites Spektrum an externen Daten an. Dazu gehören die Krankenhaus-Qualitätsberichte des Gemeinsamen Bundesausschusses (G-BA) [13], das Krankenhausverzeichnis [14] sowie soziodemographische Daten der statistischen Landesämter. Für die Integration dieser Versorgungsdaten lassen sich die Anforderungen Erweiterbarkeit und Offenheit konkretisieren:

1. Das Datenmodell soll sowohl einer hohen und wachsenden Anzahl von Merkmalen standhalten, als auch die Vereinzelung von Merkmalsausprägungen handhaben können (vgl. [15]).
2. Das Datenmodell soll Analysen im Quer- und Längsschnitt ermöglichen. Dies erfordert insbesondere, dass identische Teilnehmer und übereinstimmende Merkmale aufeinander abgebildet werden und dass erfasst wird, wie sich Items über die Zeit verändert haben.
3. Das Datenmodell soll um Kontextdaten erweiterbar sein, die Merkmale und Teilnehmer beschreiben, wiederum ohne die bestehende Struktur des Datenmodells ändern zu müssen.
4. Das Datenmodell soll datengetrieben und nicht auswertungsgetrieben sein. Es soll insbesondere nicht zwischen abhängigen und unabhängigen Variablen unterscheiden oder Analysen auf eine Auswahl von Informationsobjekten (z.B. Qualitätsindikatoren vs. Diagnosen) beschränken.

Wir stellen im Folgenden ein Konzept für ein Datenmodell vor, das diese Anforderungen umsetzt, und dessen Implementierung anhand der Krankenhaus-Qualitätsberichte des G-BA.

Stand der Forschung

Informationssysteme, die regelmäßig anfallende Daten aus verschiedenen Quellen integrieren und so organisieren, dass jeder beliebige Zustand von Informationsobjekten in der Vergangenheit, aber auch ihre Veränderung über die Zeit abrufbar sind, werden als Data-Warehouses bezeichnet [16]. Sie haben als Systeme zur Datenintegration in die Versorgungsforschung bereits Einzug gehalten [17], [18].

Das Datenmodell eines Data-Warehouse gibt vor, wie die Daten so zu organisieren sind, dass die genannten Anforderungen bestmöglich erfüllt werden. Die am weitesten verbreiteten Datenmodelle für Data-Warehouse-Systeme sind normalisierte Relationale Modelle und die Dimensionale Modellierung (Sternschema) [19], [20]. Ein normalisiertes Modell sichert zwar Konsistenz und referentielle Integrität der Daten, erzwingt aber in aller Regel tiefgreifende strukturelle Anpassungen, wenn neue Datenquellen integriert werden müssen oder Datenspezifikationen sich ändern [21]. Dimensionale Modelle unterteilen Daten in Ereignisse („Fakten“) und ereignisbeschreibende Dimensionen. Sie setzen mithin voraus, bestimmte Daten als zentrale Ereignisse zu identifizieren und vorab festzulegen, anhand welcher Dimensionen diese ausgewertet werden sollen. Diese Festlegungen sind stark anforderungsgetrieben und erfordern es, festzulegen, welche Daten für die Analyse als abhängige Variablen (Fakten) gelten und welche unabhängig (Dimensionen) sind [22].

Ändern sich die Anforderungen müssen Dimensionale Modelle erheblich angepasst werden.
In jüngster Zeit konnten sich für Geschäftsanwendungen aber auch einige neuere Modellierungstechniken wie Data Vault (DV) etablieren [23], [24]. Sie trennen in strikter Weise Informationsobjekte, deren Attribute und Beziehungen voneinander, so dass schematische Änderungen immer nur eine Erweiterung und niemals eine Anpassung bestehender Datenbankstrukturen zur Folge haben [20], [25]. Diese jüngeren Ansätze fanden bislang in der wissenschaftlichen Literatur nur zögerlich Beachtung.
In der Versorgungsforschung hat parallel zu den genannten Data-Warehouse-Datenmodellen, das Entity-Attribute-Value-Datenmodell (EAV) Verbreitung gefunden [21], [26], [27]. Diese Art der Modellierung liefert ein einfaches und flexibles Datenmodell, das sich besonders robust bei stark *vereinzeltem* und *veränderlichem* Auftreten von Merkmalen zeigt [15]. Die Informationsobjekte von Versorgungsdaten verfügen über eine hohe Anzahl an potentiellen Merkmalen, aber oft ist nur eine kleine Teilmenge ihrer Ausprägungen erfasst. Es kommen außerdem laufend neue Merkmale hinzu. Diese Eigenschaften vertragen sich nur schwer mit den klassischen Data-Warehouse-Modellen, die voraussetzen, dass Merkmale erschöpfend und abschließend vorab festgelegt werden. EAV begegnet dem Problem, indem die erfassten Merkmale nicht strukturell im relationalen Schema festgeschrieben werden, sondern auf Datensatzebene (zeilenweise) repräsentiert werden [21].
Reine EAV-Datenmodelle bieten also einen robusten Ansatz bei hoher Datenheterogenität und rascher Evolution von Datenstrukturen. Sie leisten dies aber zu Lasten der Datenorganisation, was sich vor allem durch komplexe Abfragen bemerkbar macht [21]. Bisher wurde deshalb versucht, die EAV-Daten entweder in späteren Architekturschichten des Data-Warehouse (DW) anforderungsgetrieben doch wieder in ein Sternschema-Modell zu überführen [28] oder sie in eine dimensionale Struktur einzubetten [29], [30]. Da der erste Ansatz die ursprünglichen Probleme der sich wandelnden Merkmale und unterspezifizierter Anforderungen in spätere Schichten verlagert, ist hier nur der zweite interessant. Es werden dazu EAV-Tabellen wie Faktentabellen im Sternschema behandelt und durch Dimensionstabellen beschrieben. Die Dimensionen beschreiben in dem Fall keine Auswertungsdimensionen, sondern Attributdimensionen, je nachdem, welche Attributtypen als Spalten der Faktentabelle modelliert sind [30]. Dieses Vorgehen bewahrt im Wesentlichen die Flexibilität des EAV-Modells für die eigentlichen Messdaten, bringt aber die bekannten Einschränkungen des Sternschemas für die Dimensionstabellen mit sich. Bisher gibt es keinen Ansatz, der EAV-Modelle mit neueren DW-Datenmodellen wie DV verknüpft, um auf diese Weise Daten besser zu strukturieren, ohne die Flexibilität von EAV aufzugeben.

Konzept

Die Anforderungen 2. und 3. an die Verarbeitung von Versorgungsdaten im ROSE-Projekt sprechen für einen Data-Warehouse-Ansatz. Allerdings wird an den Anforderungen 1. und 4. deutlich, dass die gängigen Datenmodelle für Data-Warehouses nicht im gewünschten Maße offen und erweiterbar gegenüber neuen oder sich ändernden Quellsystem sind, ohne umfangreiche und aufwändige Modifikationen des Datenschemas nach sich zu ziehen. Deshalb wurde für die vorliegende Arbeit ein EAV-Ansatz um ein DV-Modell erweitert.
Zentral für die Modellierung von Versorgungsdaten ist die Beziehung zwischen den Informationsobjekten Merkmalsträger, Merkmal und Merkmalsausprägung. In einem einfachen EAV-Schema würden diese Informationsobjekte entsprechend als Entität, Attribut bzw. Wert modelliert. Gemäß Anforderungen sind zu Merkmalen und ihren Ausprägungen jedoch Metadaten zu erfassen und Merkmalsträger mit Daten aus weiteren Datenquellen zu verknüpfen. Aus diesem Grund abstrahieren wir von Attributen und Werten, indem diese ebenfalls als eigenständige Entitäten modelliert werden. Die Versorgungsdaten organisieren sich somit in Entitäten für Merkmalsträger (z.B. Klinikstandorte, Personen), Klassen von Merkmalen (z.B. Qualitätsindikatoren, Personalausstattung) und Ausprägungen (z.B. ICD-Codes, Antwortmöglichkeiten für Befragungsitems). Die DV-Modellierung sieht für jede Entität im logischen Datenbankmodell eine Hub-Tabelle vor, die einmalig das Vorkommen einer konkreten Entitätsinstanz anhand ihres natürlichen Schlüssels erfasst und ihr durch einen Hashwert einen technischen Schlüssel zuweist [31]. Die Verknüpfung von Entitäten erfolgt über Link-Tabellen, die lediglich Referenzen auf die jeweiligen technischen Schlüssel in den Hubs enthalten. Es ergibt sich also an Stelle der Entität-Attribut-Wert-Tabelle ein n-facher Link, der bestimmte Merkmalsträger mit mehreren Merkmalsklassen und Ausprägungen verknüpft und auf die entsprechenden Entitätsschlüssel verweist. Nach der DV-Spezifikation [31] geben Hubs und Links Aufschluss darüber, welche Entitätsinstanzen und -beziehungen jemals aufgetreten sind. Entitätsattribute, und damit auch alle zeitabhängigen Daten, werden im DV-Modell in Satellites modelliert. Datensätze werden hierbei niemals verändert oder gelöscht. Die Historisierung erfolgt über Zeitstempel.

Implementierung

Das Datenmodell wurde ausgehend von den Qualitätsberichten des G-BA für die Jahre 2011 bis 2014 in PostgreSQL 9.6 implementiert, die im Laufe der Entwicklung mit Befragungsdaten (z.B. [3]) und anderen externen Daten verknüpft wurden. Ein Ausschnitt des resultierenden Datenbankschemas ist in Abbildung 1 dargestellt. Gezeigt sind Tabellen und Fremdschlüsselbeziehungen, die die Ergebniskennzahlen des Teil C der Qualitätsberichte [13] enthalten: Die zentrale Beziehung besteht hier

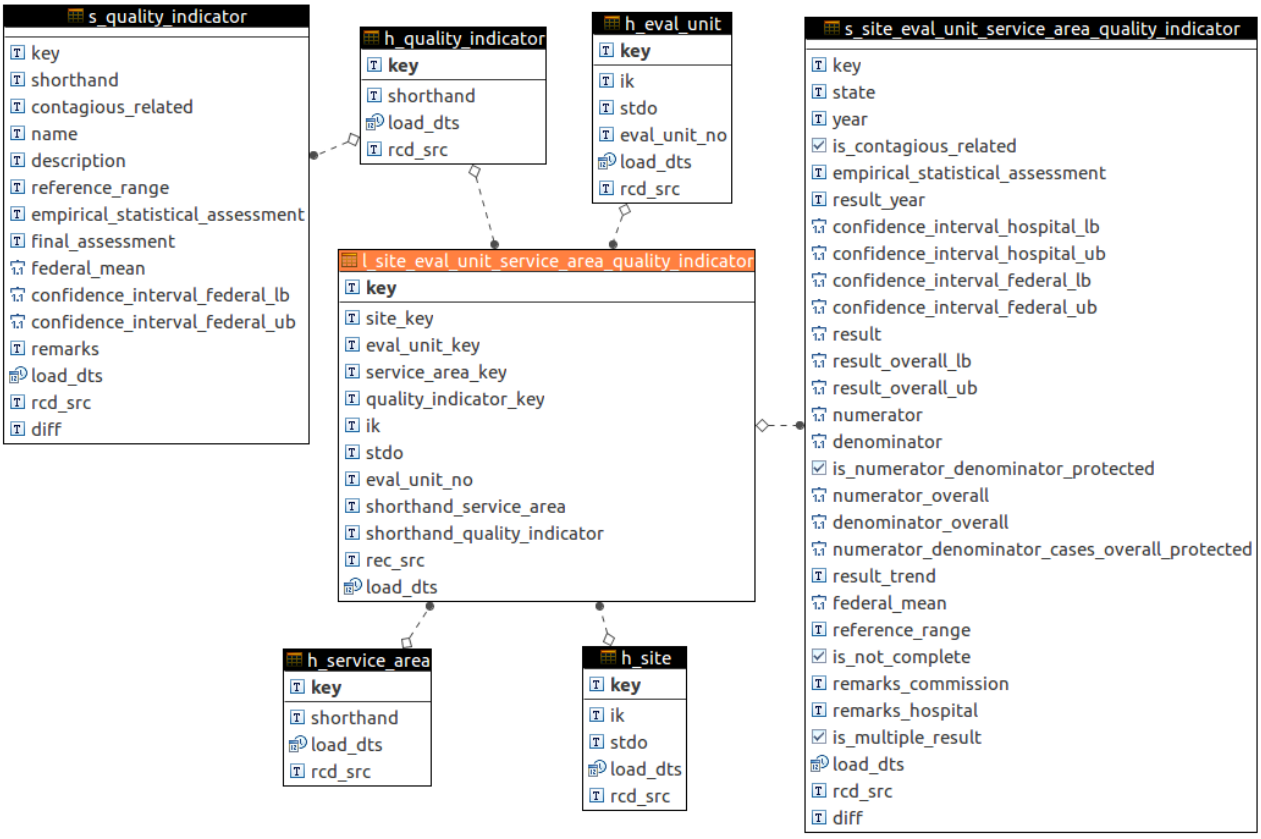

Abbildung 1: Der Ausschnitt des Gesamt-Datenmodells, der die Daten für Qualitätsindikatoren umfasst. Die gezeigten Hubs und Links sind mit weiteren nicht abgebildeten Hubs und Links verknüpft, die weitere Daten der Qualitätsberichte und aus anderen Quellen (z.B. Befragungsdaten) enthalten.

zwischen dem 4-Tupel der Entitäten Klinikstandorte, Auswertungseinheiten, Leistungsbereiche und Qualitätsindikatoren. Gemäß der Data-Vault-Spezifikation wurden für Entitäten Hubs und für deren Beziehungen Links als relationale Tabellen erstellt. Ihre zeitabhängigen und deskriptiven Attribute werden in Satellites ausgelagert. Da die Ausprägungen der Qualitätsindikatoren numerisch sind, gibt es für diese Klasse von Merkmalen keine eigene Ausprägungsentität. Die zeitabhängige Ausprägung für den 4-Tupel Standort-Auswertungseinheit-Leistungs-Bereich-Qualitätsbericht ist in den Satellite des Links dieser Beziehung ausgelagert. Es wurden alle Informationsobjekte als Hubs modelliert, deren Beschreibung durch Metadaten vorliegt bzw. fachlich zu erwarten ist (z.B. Adressen) oder die mit mehr als einem Informationsobjekt in Beziehung stehen.

Die technischen Schlüssel der Entitäten wurden über Anwendung des MD5-Hashing-Algorithmus auf die natürlichen Schlüsselattribute gebildet. Da alle Datenquellen als Flatfiles vorliegen, wird das Feld *rcd_src*, das die Datenquelle eines Datensatzes enthält, mit dem vollständigen Dateipfad der jeweiligen Quelldatei befüllt. Das Feld *load_dts* enthält – entgegen der DV-Spezifikation – das Erhebungsdatum der Befragung und nicht den Zeitpunkt, zu dem der Datensatz in die Datenbank geschrieben wurde. Grund ist, dass die Historisierung von technisch bedingten Datenänderungen im vorliegenden Fall nicht von Interesse ist, so dass die dafür erforderliche zusätzliche Komplexität (zweite Datumsspalte oder Indextabelle) zu rechtfertigen wäre. Entsprechend der Zielsetzung übergreifend historisierter Merkmale werden die Qualitätsindikatoren in dem Hub *h_quality_indicator* abgelegt, der somit offen gegenüber neuen Qualitätsindikatoren ist. Dieser Hub hat den Satellite s_ *quality_indicator*, der der Indikatorspezifikation des G-BA [13] genügt (vgl. Abbildung 1).

Dieses Datenmodell enthält Daten zu 2.883 Klinik-Standorten. Für diese liegen hier für den Ergebnisteil der Qualitätsberichte insgesamt 733.000 Datenwerte für 381 Qualitätsindikatoren vor, verteilt auf insgesamt knapp 20.000 Auswertungseinheiten, 39 Leistungsbereiche und die Jahre 2011 bis 2014. Das gesamte Datenmodell findet sich in Anhang 1.

Lessons learned

Das DV-Modell erwies sich im Verlaufe des Projekts als natürliche Erweiterung des EAV-Ansatzes [21]. Die DV-Spezifikation [31] gibt einfache konzeptionelle Bausteine vor, mit denen von Entitäten, Attributen und Werten abstrahiert werden kann, so dass diese als eigenständige Informationsobjekte mit lose gekoppelten Beziehungen untereinander rückverfolgbar integriert und historisiert werden. Die wichtigsten Designentscheidungen bestehen darin, festzulegen, welche Daten als Entitäten (Hubs) und

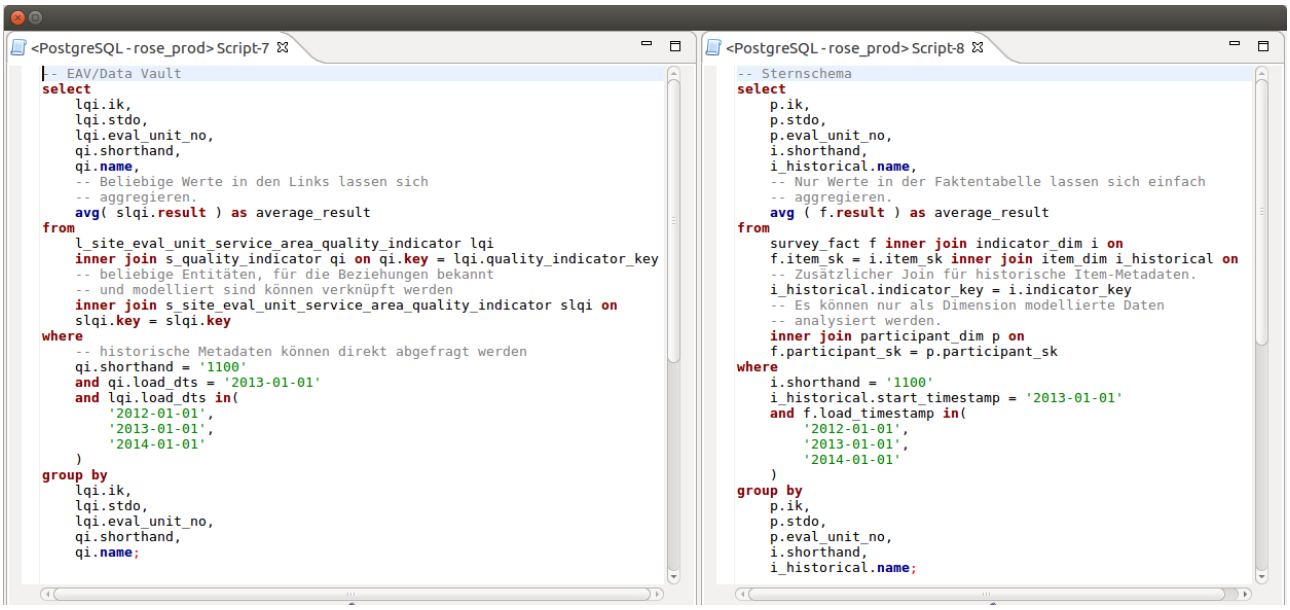

Abbildung 2: Vergleich von SQL-Abfragen gegen das implementierte DV-Modell und gegen ein vergleichbares Sternschema

welche als elementare Attribute (Spalten der Satellites) zu modellieren waren. Diese Entscheidung beeinflusst, wie flexibel das Datenmodell für jene Daten ist. Es empfiehlt sich daher, Informationsobjekte, für die heterogene oder inkonsistente Daten erwartet werden, von vornherein als vollwertige Entitäten zu modellieren. Dies zieht jedoch für jede solche Entität die Definition eines Hubs und ggf. mehrerer Links und Satellites nach sich. Die nachträgliche Aufwertung von Attributspalten im Satellite zu eigenen Hubs ist in jedem Fall mit erheblichem Aufwand verbunden.

Bei der Modellierung von Beziehungen war besonders zu beachten, welche Datenzusammenhänge zeitvariant und welche zeitinvariant sind. Links für Teilnahmen an Befragungen sind beispielsweise zeitinvariant, da eine Teilnahme niemals rückgängig gemacht werden kann. Der Link für Qualitätsindikatoren dagegen ist zeitvariant, weil ein Krankenhaus nicht zwangsläufig immer alle erforderlichen Indikatoren an den G-BA meldet. Die Information über die Existenz einer Beziehung zu gegebenem Zeitpunkt muss also im Satellite stehen. Da das Konzept des Links zeitinvariant ist, ist es erforderlich, dass es für zeitabhängige Zusammenhänge einen eigenen Satellite gibt, auch wenn der Link über keine Attribute verfügt.

Die ausgeprägte Flexibilität des DV-Modells erzwingt, dass klar abzugrenzen ist, welche Datenänderungen historisiert werden, da sonst die Gefahr von „Overengineering“ besteht. Im Hinblick auf den Grad der Datenorganisation stellt sich die DV-Modellierung als solider Kompromiss zwischen EAV und Sternschema heraus. EAV ist ein vollends generisches Datenmodell mit nur rudimentärer Datenorganisation, da nur zwischen Entitäten, Attributen und Werten unterschieden wird [15], [21]. Entsprechend hoch ist die Komplexität der Abfragen. Das Sternschema auf der anderen Seite definiert inhaltlich festgelegte deskriptive Dimensionen und zu analysierende Faktenwerte, was die Daten fachlich und leicht zugänglich strukturiert und Abfragen für die vorgegebenen Anwendungsfälle minimiert. Das DV-Modell bewegt sich zwischen diesen Polen und kann nach Bedarf generischer oder konkreter definiert werden [23]. Dies illustriert unsere Wahl der Qualitätsindikatoren als Hubs. Denkbar wäre stattdessen auch ein gänzlich generischer Hub „Krankenhausmerkmal“ oder aber noch konkretere Hubs für bestimmte Indikatoren gewesen.

Abbildung 2 illustriert, wie sich das implementierte Schema gegenüber einem vergleichbaren Sternschema verhält. Das EAV-/DV-Modell ermöglicht insbesondere die Aggregation beliebiger Kennzahlen und erlaubt eine flexiblere Zusammenführung historischer Versionen von Datensätzen. Mit wachsender Datenmenge und steigenden Anfragen wird man zukünftig Aussagen über die Performanz im Echtbetrieb tätigen können.

Fazit

Im vorliegenden Beitrag wurde ein kombiniertes EAV-/DV-Modell für die historisierte und flexible Integration von semantisch heterogenen Befragungsdaten entwickelt. Es liefert eine integrierte Datensicht auf Versorgungsdaten und historisiert vollständig Metadaten zu Merkmalen und ggf. Ausprägungen, wodurch Analysen auf frei zusammenstellbaren Teilmengen der Daten bestehen können. Darüber hinaus ist das Datenmodell beliebig um neue Entitäten und Beziehungen erweiterbar.

Anmerkungen

Interessenkonflikte

Die Autoren erklären, dass sie keine Interessenkonflikte in Zusammenhang mit diesem Artikel haben.

Danksagung

Die Arbeiten werden in dem durch das Land Niedersachsen geförderten Projekt „Das Lernende Gesundheitssystem in der Region Osnabrück Emsland (ROSE)“ (Förderkennzeichen ZN 3103) durchgeführt.

Anhänge

Verfügbar unter
http://www.egms.de/en/journals/mibe/2017-13/mibe000170.shtml

1. supplement-complete-data-model.png (298 KB)
   Gesamtes Datenmodell

Literatur

1. Kuo MH, Sahama T, Kushniruk AW, Borycki EM, Grunwell D. Health big data analytics: current perspectives, challenges and potential solutions. Int J Big Data Intelligence. 2014;1(1/2):114-26. DOI: 10.1504/IJBDI.2014.063835
2. Thye J, Straede MC, Liebe JD, Hübner U. IT-benchmarking of clinical workflows: concept, implementation, and evaluation. Stud Health Technol Inform. 2014;198:116-24. DOI: 10.3233/978-1-61499-397-1-116
3. Hübner U, Liebe JD, Straede MC, Thye J. IT-Report Gesundheitswesen: Schwerpunkt IT-Unterstützung klinischer Prozesse. Hannover: Niedersächsisches Ministerium für Wirtschaft, Arbeit und Verkehr; 2014. (Schriftenreihe des Niedersächsischen Ministeriums für Wirtschaft, Arbeit und Verkehr).
4. Pommerening K, Deserno TM, Ingenerf J, Lenz R, Schmücker P. Der Impact der Medizinischen Informatik. Informatik Spektrum. 2015;38(5):347-69. DOI: 10.1007/s00287-014-0767-7
5. Lenz R, Beyer M, Kuhn KA. Semantic integration in healthcare networks. Int J Med Inform. 2007 Feb-Mar;76(2-3):201-7. DOI: 10.1016/j.ijmedinf.2006.05.008
6. Länsisalmi H, Kivimäki M, Aalto P, Ruoranen R. Innovation in healthcare: a systematic review of recent research. Nurs Sci Q. 2006 Jan;19(1):66-72; discussion 65. DOI: 10.1177/0894318405284129
7. Farré N, Vela E, Clèries M, Bustins M, Cainzos-Achirica M, Enjuanes C, Moliner P, Ruiz S, Verdú-Rotellar JM, Comín-Colet J. Real world heart failure epidemiology and outcome: A population-based analysis of 88,195 patients. PLoS One. 2017 Feb 24;12(2):e0172745. DOI: 10.1371/journal.pone.0172745
8. Sacker A, Ross A, MacLeod CA, Netuveli G, Windle G. Health and social exclusion in older age: evidence from understanding society, the UK household longitudinal study. J Epidemiol Community Health. 2017 Feb;71:681-90. DOI: 10.1136/jech-2016-208037
9. Lenz R, Beyer M, Meiler C, Jablonski S, Kuhn KA. Informationsintegration in Gesundheitsversorgungsnetzen – Herausforderungen an die Informatik. Informatik Spektrum. 2005;28(2):105-19. DOI: 10.1007/s00287-005-0467-4
10. Abbott MR. A new path for science? In: Hey T, Tansley S, Tolle K, editors. The fourth paradigm: data-intensive scientific discovery. Redmond, WA: Microsoft Research; 2009. p. 111-6.
11. Patel NV, editor. Adaptive evolutionary information systems. Hershey, PA: Idea Group (IGI); 2003. DOI: 10.4018/978-1-59140-034-9
12. Hübner U, Babitsch B, Kortekamp S, Egbert N, Braun von Reinersdorff A. ROSE – das lernende Gesundheitssystem in der Region Osnabrück-Emsland [ROSE – the learning health care system in the Osnabrück-Emsland]. International Journal of Health Professions. 2016;3(1):14-20. DOI: 10.1515/ijhp-2016-0006
13. Regelungen des Gemeinsamen Bundesausschusses gemäß § 137 Abs. 3 Satz 1 Nr. 4 SGB V über Inhalt, Umfang und Datenformat eines strukturierten Qualitätsberichts für nach § 108 SGB V zugelassene Krankenhäuser (Regelungen zum Qualitätsbericht der Krankenhäuser, Qb-R). Stand: 19. Juni 2014. Available from: http://www.g-ba.de/downloads/62-492-906/Qb-R_2014-06-19.pdf
14. Deutsche Krankenhausgesellschaft. Deutsches Krankenhausverzeichnis. 2010. Available from: http://www.deutsches-krankenhaus-verzeichnis.de
15. Löper D, Klettke M, Bruder I, Heuer A. Integrating healthcare-related information using the entity-attribute-value storage model. In: He J, Liu X, Krupinski EA, Xu G, editors. Health Information Science. First International Conference, HIS 2012, Beijing, China, April 8-10, 2012. Proceedings. Berlin: Springer; 2012. (Lecture Notes in Computer Science; 7231). p. 13-24. DOI: 10.1007/978-3-642-29361-0_4
16. Inmon WH. Building the data warehouse. 4th ed. Indianapolis, IN: Wiley; 2005.
17. Khan SI, Hoque ASML. Towards development of health data warehouse: Bangladesh perspective. In: 2015 International Conference on Electrical Engineering and Information Communication Technology (ICEEICT); 21-23 May 2015; Dhaka, Bangladesh. IEEE; 2015. p. 1-6. DOI: 10.1109/iceeict.2015.7307514
18. Turley CB, Obeid J, Larsen R, Fryar KM, Lenert L, Bjorn A, Lyons G, Moskowitz J, Sanderson I. Leveraging a Statewide Clinical Data Warehouse to Expand Boundaries of the Learning Health System. EGEMS (Wash DC). 2016 Dec 6;4(1):1245. DOI: 10.13063/2327-9214.1245
19. Bauer A, Günzel H, editors. Data-Warehouse-Systeme: Architektur, Entwicklung, Anwendung. 4th ed. Heidelberg: dpunkt-Verlag; 2013.
20. Jovanovic V, Subotic D, Mrdalj S. Data modeling styles in data warehousing. In: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO); 26-30 May 2014; Opatija, Croatia. IEEE; 2014. p. 1458-63. DOI: 10.1109/mipro.2014.6859796
21. Dinu V, Nadkarni P. Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. Int J Med Inform. 2007 Nov-Dec;76(11-12):769-79. DOI: 10.1016/j.ijmedinf.2006.09.023
22. Kimball R, Ross M. The data warehouse toolkit: the definitive guide to dimensional modeling. 3rd ed. Indianapolis, IN: Wiley; 2013.
23. Gluchowski P, Chamoni P, editors. Analytische Informationssysteme. Business Intelligence-Technologien und -Anwendungen. 5th ed. Berlin, Heidelberg: Springer; 2016. DOI: 10.1007/978-3-662-47763-2
24. Linstedt D, Graziano K, Hultgren H. The business of data vault modeling. 2nd ed. Lulu.com; 2009.
25. Bojicic I, Marjanovic Z, Turajlic N, Petrovic M, Vuckovic M, Jovanovic V. A comparative analysis of data warehouse data models. In: 2016 6th International Conference on Computers Communications and Control (ICCCC); 10-14 May 2016; Oradea, Romania. IEEE; 2016. p. 151-9. DOI: 10.1109/icccc.2016.7496754

26. Nadkarni PM, Marenco L, Chen R, Skoufos E, Shepherd G, Miller P. Organization of heterogeneous scientific data using the EAV/CR representation. J Am Med Inform Assoc. 1999 Nov-Dec;6(6):478-93. DOI: /10.1136/jamia.1999.0060478

27. Nadkarni PM, Brandt C. Data extraction and ad hoc query of an entity-attribute-value database. J Am Med Inform Assoc. 1998 Nov-Dec;5(6):511-27. DOI: 10.1136/jamia.1998.0050511

28. Yamamoto K, Sumi E, Yamazaki T, Asai K, Yamori M, Teramukai S, Bessho K, Yokode M, Fukushima M. A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. BMJ Open. 2012 Oct 31;2(6). pii: e001622. DOI: 10.1136/bmjopen-2012-001622

29. Wade TD, Hum RC, Murphy JR. A Dimensional Bus model for integrating clinical and research data. J Am Med Inform Assoc. 2011 Dec;18 Suppl 1:i96-102. DOI: 10.1136/amiajnl-2011-000339

30. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar-Apr;17(2):124-30. DOI: 10.1136/jamia.2009.000893

31. Linstedt D, Olschimke M. Building a scalable data warehouse with Data Vault 2.0. 1st ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2015.

Korrespondenzadresse:
Jens Rauch
Hochschule Osnabrück, Fakultät Wirtschafts- und Sozialwissenschaften, Forschungsgruppe Informatik im Gesundheitswesen, Postfach 1940, 49009 Osnabrück, Deutschland
j.rauch@hs-osnabrueck.de

Bitte zitieren als
Rauch J, Weiss JP, Teuteberg F, Hübner U. Konsolidierte Datenmodellierung von Versorgungsdaten mit dem Entity-Attribute-Value-Modell und Data Vault. GMS Med Inform Biom Epidemiol. 2017;13(1):Doc03.
DOI: 10.3205/mibe000170, URN: urn:nbn:de:0183-mibe0001706

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/mibe/2017-13/mibe000170.shtml

Veröffentlicht: *29.08.2017*