Maximilian Eibl, Martin Gaedke. (Hrsg.): INFORMATIK 2017,
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 1357

Entwicklung eines Datenmodells für ein umfassendes Forschungsdatenmanagement zur flexiblen Analyse longitudinaler Daten

Jan-Patrick Weiß[1], Jens Rauch[2], Jens Hüsers[3], Jan-David Liebe[4], Frank Teuteberg[5] und Ursula Hübner[6]

Abstract: Forschungsdatenbanken dienen als gemeinsame Datenbasis für heterogene Datensätze unterschiedlicher Wissenschaftler, um neue Forschungsansätze, Ideen oder Fragestellungen im Forschungsprozess zu identifizieren und zu analysieren. Klassische Datenmodellierungs-Ansätze wie das dimensionale Modell oder das Entity-Attribute-Value (EAV) Modell erweisen sich entweder als unflexibel hinsichtlich neuer Anforderungen und der Erweiterung um neue Datenquellen oder erschweren longitudinale Analysen. In diesem Artikel wird ein grundlegendes Systemdesign für sich häufig ändernder Forschungsdaten vorgestellt und als erster Meilenstein die Implementation des Datenmodells fokussiert. Das EAV-Modell wurde hierzu um das Data-Vault-Modell erweitert. Dieser kombinierte Ansatz ermöglicht die Historisierung beliebiger Merkmalsausprägungen und die Erweiterung um neue Merkmale aus weiteren Datenquellen.

Keywords: Datenmanagement, Datenmodell, longitudinale Daten, Forschungsdatenbank, Data-Vault-Modell

1 Einleitung

Häufig werden in Forschungsprojekten heterogene Daten gesammelt oder erstmalig erzeugt [Po15], [Th14]. Durch eine gemeinsame Datenbasis für unterschiedliche Wissenschaftler können darüber neue Forschungsansätze, Ideen oder Fragestellungen im Rahmen des Forschungsprozesses identifiziert [He09] und analysiert werden. Das Forschungsdatenmanagement sichert somit den langfristigen Austausch von Informationen für zukünftige Forschungsfragen [Me12]. Im Gesundheitswesen hat sich in den letzten Jahren das

[1] Hochschule Osnabrück, Informatik im Gesundheitswesen, Postfach 19 40, 49009 Osnabrück, j.p.weiss@hs-osnabrueck.de
[2] Hochschule Osnabrück, Informatik im Gesundheitswesen, Postfach 19 40, 49009 Osnabrück, j.rauch@hs-osnabrueck.de
[3] Hochschule Osnabrück, Informatik im Gesundheitswesen, Postfach 19 40, 49009 Osnabrück, j.huesers@hs-osnabrueck.de
[4] Hochschule Osnabrück, Informatik im Gesundheitswesen, Postfach 19 40, 49009 Osnabrück, j.liebe@hs-osnabrueck.de
[5] Universität Osnabrück, Unternehmensrechnung und Wirtschaftsinformatik, Katharinenstr. 1, 49069 Osnabrück, frank.teuteberg@uni-osnabrueck.de
[6] Hochschule Osnabrück, Informatik im Gesundheitswesen, Postfach 19 40, 49009 Osnabrück, u.huebner@hs-osnabrueck.de

doi:10.18420/in2017_136

Konzept des „lernenden Gesundheitssystems“ entwickelt, das die Nutzung von Routinedaten für die Forschung vorsieht [OAM07] und diese bei Bedarf mit Forschungsdaten verknüpft. Dabei soll über eine zentrale oder vernetzte Plattform ein Informationsaustausch zwischen Forschung und Praxis gefördert werden, sodass beispielsweise Sekundärdaten weiteren Anwendungsfällen zugeführt werden und das erarbeitete Wissen unter den unterschiedlichen Stakeholdern geteilt wird [FWB10]. Die verteilten, unterschiedlichen, internen und externen Datenquellen führen jedoch zu informationstechnologischen Barrieren [LBK07]. Für einen effektiven und effizienten Informationsaustausch, müssen im Prozess des Forschungsdatenmanagements die Datenbestände integriert sowie konsistent und strukturiert persistiert werden.

Die Forschungsprojekte ROSE und INITIATIVE eHealth an der Hochschule Osnabrück sind Beispiele in denen Daten unterschiedlicher Teilprojekte gesammelt und in einer gemeinsame Forschungsdatenbank zusammengeführt werden sollen, um ein lernendes Gesundheitssystems in der Region Osnabrück zu realisieren [Hü16]. Dazu werden Primärdaten in unterschiedlicher Form erhoben (schriftliche Befragungen, Interviews) und mit Sekundärdaten (z. B. Qualitätsberichte der Krankenhäuser, demographische Daten der statistischen Ämter) zusammengeführt. Die Befragungsdaten werden je nach Teilprojekt einmalig oder in regelmäßigen Intervallen erfasst. Neue Forschungsideen führen dabei oft zu sich ändernden Items und damit zu sich ändernden Schnittstellen. Dennoch sollen die Schnittmengen der in allen Erhebungen gleichen Daten abgebildet werden, d.h. es sollen longitudinale Analysen und Quellenzusammenführung bei gleichem Erhebungszeitpunkt möglich sein. Dies setzt voraus, dass die Daten nicht isoliert abgelegt, sondern in einer zentralen Forschungsdatenbank zusammengeführt werden. Diese soll als eine integrierte Plattform zur Verwaltung, Analyse und Präsentation von Forschungsdaten dienen und unter Nutzung von Open Source Komponenten entwickelt, implementiert und kontinuierlich verbessert werden.

Mit diesem Artikel wird das grundlegende Systemdesign gezeigt und die Implementation des Datenmodells als erster Meilenstein fokussiert. Zur Erhebung der Anforderungen wurden Interviews mit Anwendern aus den Forschungsgruppen der Teilprojekte durchgeführt. Ein zentrales Anwendungsgebiet ist dabei die seit 2002 bundesweit und in den letzten Jahren auch international durchgeführte Befragung zum Digitalisierungsgrad in Krankenhäusern [Fo02]. Der Digitalisierungsgrad von Krankenhäusern unterliegt einem inkrementellen Wachstum, was zu einem Bedarf an longitudinalen Studien führt [Bu11]. Außerdem führt der kontinuierliche technische Fortschritt zu sich mit der Zeit ändernden und neuen Befragungsitems [Ag10]. Um die Auswirkungen dieses innovativen Prozesses auf Krankenhäuser und deren Umwelt für Praktiker, Politik und Patienten sichtbar zu machen, müssen die Befragungsdaten um Sekundärdaten (z. B. Qualitätsberichte der Krankenhäuser) in den jeweiligen Analyse ergänzt werden [Am09]. Für das Datenmodell ist daher zentral, dass es einerseits robust gegenüber Datenarten aus unterschiedlichen Datenquellen ist und deshalb über ein hohes Maß an Erweiterbarkeit verfügt. Dazu müssen auch semantische Änderungen (z.B. Änderung des Fragetextes bei Befragungsdaten) nicht nur kompatibel mit dem Datenmodell sein, sondern auch nachvollziehbar abgebildet werden können. An-

dererseits soll das Datenmodell verschiedene analytische Sichten auf die integrierten Daten gestatten und daher keine Form der Auswertung, z. B. nach Dimensionen und Fakten, vorschreiben. Außerdem sollen externe Informationen wie beispielsweise Qualitätsdaten oder andere frei verfügbare Daten integriert werden.

Es konnten insgesamt sechs wesentliche Anforderungen konkretisiert werden: (1.) Das Datenmodell soll beliebige Befragungsdatensätze unterschiedlicher Quellformate integrieren, ohne die Datenstrukturen anpassen zu müssen. (2.) Neue Umfrageschnittstellen sollen an das System angeschlossen werden können, ohne die bestehende Struktur des Datenmodells ändern zu müssen. (3.) Das Datenmodell soll sowohl in Bezug auf das Hinzufügen weiterer Merkmale, als auch hinsichtlich vereinzelter Merkmalsausprägungen skalieren. (4.) Zeitlich gesehen sollen Quer- und Längsschnittanalysen möglich sein. Dazu ist es essentiell, identische Teilnehmer und Items aufeinander abzubilden und die Veränderung von Merkmalsausprägungen zu historisieren. (5.) Das Datenmodell soll um zusätzliche Datenquellen erweiterbar sein, die Items oder Teilnehmer beschreiben, ohne die bestehende Struktur des Datenmodells ändern zu müssen. (6.) Das Datenmodell soll sich um die Datenstrukturen zentrieren und auswertungsunabhängig sein. Es soll insbesondere keine Kausalitäten oder gerichtete Zusammenhänge zwischen Merkmalen implizieren, sondern sich auf die reinen Datenrelationen beschränken.

Die Anforderungen implizieren, dass das Datenmodell in generischer Weise Befragungsdaten beschreiben kann, aber auch erweiterbar bezüglich unvorhergesehener Spezifika von Befragungsdaten oder weiteren Datenquellen ist, wie sie etwa durch neue Erhebungsinstrumente oder Sekundärdaten hinzukommen können. Außerdem beinhalten die Anforderungen, dass Items wiederholter Erhebungen, als solche auch verknüpft werden und die Änderungen explizit abgefragt werden können. Umgekehrt sollen ebenfalls befragungsübergreifend sämtliche Ausprägungen von Items für ausgewählte Merkmalsträger zusammengeführt werden. Dieses führt zu den diesem Beitrag zugrundeliegenden Forschungsfragen:

- FF1: Welches Datenmodell eignet sich als konzeptuelle Grundlage für die Umsetzung der genannten Anforderungen?
- FF2: Wie kann am Beispiel mehrjähriger und unterschiedlich strukturierter Umfragedaten ein vereinheitlichtes Datenmanagement implementiert werden?

In Kapitel 2 wird zunächst ein Überblick über unterschiedliche Datenmodelle gegeben und welche Vor- und Nachteile diese jeweils bieten. Anschließend wird in Kapitel 3 das entwickelte Konzept für die Forschungsdatenbank erläutert. Schließlich werden in Kapitel 4 die konkreten Implementierungen vorgestellt und abschließend in Kapitel 5 zusammenfassend betrachtet.

2 Stand der Forschung

Data Warehouse Systeme sind Informationssysteme, die regelmäßig anfallende Daten aus verschiedenen Quellen integrieren und die Daten organisieren, so dass unterschiedlicher Zustände der Informationsobjekte abgebildet werden und zeitliche Veränderungen abgerufen werden können [In05]. Ein Data Warehouse System umfasst neben der persistenten Datenhaltung sowohl Datenschnittstellen zu Quell- und Zielsystemen als auch Software zur Datenintegration und Monitoring sowie graphische Benutzerschnittstellen zur Datenanalyse [Ba13]. Für den vorliegenden Beitrag wird das zugrundeliegende Datenmodell fokussiert.

Klassische Datenmodelle aus dem Bereich der Wirtschaftsinformatik für Data-Warehouse-Systeme sind normalisierte Relationale Modelle und die Dimensionale Modellierung (Sternschema) [Ba13]. In der Medizinischen und der Gesundheitsinformatik hat sich parallel zu den klassischen Datenmodellen das Entity-Attribute-Value-Datenmodell entwickelt [Lö12]. Ein weiterer Modellierungsansatz, der in den letzten Jahren vermehrt für unterschiedlichste Geschäftsanwendungen angewendet wurde, ist das Data-Vault-Modell (DV) [Ch10]. In normalisierten Modellen werden Informationsobjekte, Attribute und die strukturellen Beziehungen explizit abgebildet. In dimensionalen Modellen werden Daten hingegen als Ereignisse (Fakten) und ereignisbeschreibende Dimensionen abgebildet. Normalisierte Modelle sichern die Konsistenz und referentielle Integrität der Daten. Neue Datenquellen oder die Anpassung bestehender Daten führen jedoch zu aufwendigen strukturellen Änderungen [DN07]. Bei der dimensionalen Modellierung werden die Daten anforderungsgetrieben bereits für die Analyse als abhängige Variablen (Fakten) und unabhängige Variablen (Dimensionen) modelliert [KR13]. Dementsprechend führen neue Analyseanforderungen auch bei dimensionalen Modellen zu aufwendigen strukturellen Änderungen. Daten aus der Medizin und dem Gesundheitswesen stellen diese klassischen Datenmodelle häufig vor neue Herausforderungen, weil dort jedes Informationsobjekt viele Merkmale besitzt und fortlaufend neue hinzukommen können. EAV-Modelle spezifizieren Merkmale daher nicht in einem festgelegten strukturellen Schema, sondern zeilenweise auf Datensatzebene. Dadurch können heterogene Daten mit sich laufend ändernden Datenstrukturen persistiert werden. Dieses führt jedoch zu komplexen Abfragen [DN07]. Bei der DV-Modellierung werden Informationsobjekte, Attribute und Beziehungen voneinander getrennt modelliert. Dadurch können die Datenstrukturen bei neuen Analyseanforderungen erweitert und müssen nicht geändert werden [Bo16], [JSM14]. In der wissenschaftlichen Literatur wird die DV-Modellierung bislang wenig betrachtet.

Es gibt Ansätze die Datenmodelle miteinander zu kombinieren. Dazu wurde beispielsweise ein EAV-Modell in weiteren Architekturschichten eines Data Warehouses in ein (1) Sternschema überführt [Ya12] oder in eine (2) dimensionale Struktur eingebettet [WHM11], [Mu10]. In Ansatz (1) bleibt in der weiteren Architekturschicht das Problem der sich fortlaufend hinzukommender Merkmale und neuer struktureller Anforderungen bestehen. In Ansatz (2) werden EAV-Tabellen als Faktentabellen und Dimensionen als

Attributsdimensionen modelliert. Dadurch wird zwar die Erweiterbarkeit des EAV-Modells bewahrt, jedoch ist die strukturelle Anpassung der Dimensionstabellen weiterhin eingeschränkt. Bisher gibt es keinen Ansatz, EAV-Modelle mit DV-Modellen zu verknüpfen, um für ein besseres Strukturieren der Daten ohne Verlust der Flexibilität des EAV-Ansatzes zu sorgen.

3 Konzept

Klassische Datenmodelle können die oben angeführten Anforderungen (1.), (2.), (3.) und (6.) eines Data Warehouses nicht hinlänglich bedienen. In dieser Arbeit sollen daher alle eingangs aufgestellten Anforderungen durch die Kombination des EAV-Ansatzes mit dem DV-Modell umgesetzt werden (FF1).

Unser Systemdesign umfasst zunächst eine 5-Schichten-Architektur, bestehend aus je einer Schicht für die Datenerfassung, der Datenhaltung, die Datenverarbeitung und die Datenpräsentation. Diese sind über eine fünfte Schicht zur Datenintegration miteinander verbunden (vgl. Abb. 1). Die Datenerfassungsschicht beschreibt alle Instrumente, aus denen die relevanten Daten stammen. Die Datenhaltung wird durch ein Data Warehouse zur Datenkonsolidierung und Speicherung beschrieben und besteht aus drei weiteren Schichten: In der Quellschicht wird zunächst jede der von der Datenerfassungsschicht bereitgestellten Datenquellen als relationale Datenbanktabelle gespiegelt und zur weiteren Verarbeitung gespeichert. Die Primärdaten aus verschiedenen Umfragen und Sekundärdaten (z. B. Qualitätsberichte, Krankenhausverzeichnis, klinische Daten) werden in der Kernschicht in eine konsolidierte Form transformiert. Data Marts bieten in der Analyseschicht optimierte Ansichten mit automatisch aggregierten Daten für vordefinierte Analysen. Um komplexere Berechnungen durchzuführen, werden die Daten in weitere Werkzeuge der Datenverarbeitungsschicht geladen und relevante Ergebnisse in weiteren Data Marts gespeichert. Der Austausch der Daten erfolgt zentral über die Datenintegrationsschicht. In der Präsentationsschicht werden die Daten für standardisierte regelmäßige Berichte oder für weitere Untersuchungen angezeigt.

In diesem Artikel soll auf die Datenhaltung der Kernschicht fokussiert werden. Zur Repräsentation der Umfragedaten wird die Datenstruktur in die Informationsobjekte Merkmalsträger, Merkmal und Merkmalsausprägung unterteilt. Nach den eingangs spezifizierten Anforderungen sollen Merkmalsträger, neben den Umfragedaten, mit Daten aus weiteren Datenquellen (z.B. Krankenhausverzeichnis, Qualitätsberichte) verknüpft werden. Außerdem sollen Merkmale und Merkmalsausprägungen durch Metadaten näher beschrieben werden. Daher werden in diesem Konzept, im Gegensatz zum EAV-Schema, die Informationsobjekte nicht als Entität, Attribut und Wert, sondern jeweils als eigenständige Entitäten modelliert. Die Forschungsdaten werden somit in Entitäten für Merkmalsträger (z. B. Befragungsitems, Klinikstandorte), Gruppen von Merkmalen (z. B. thematische Itemblöcke, Diagnosesysteme) und Ausprägungen (z. B. Antwortoptionen, Fallzahlen) organisiert.

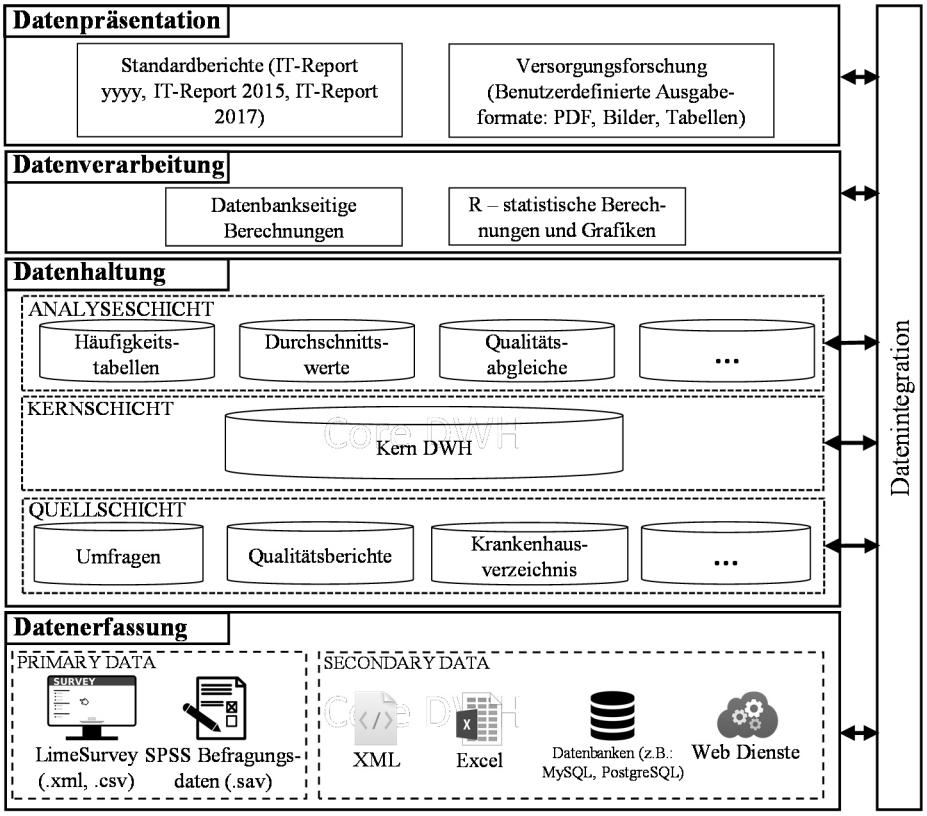

Abb. 1: 5-Schichten-Architektur - Systemdesign

Für jede Entität wird nach der DV-Modellierung eine *Hub*-Tabelle erstellt. Jeder *Hub* beinhaltet sowohl den natürlichen Schlüssel der Entität als auch einen generierten technischen Schlüssel [LO16]. Die technischen Schlüssel werden in *Link*-Tabellen als Referenz genutzt, um die *Hubs* bzw. Entitäten miteinander zu verknüpfen. Es ergibt sich also an Stelle der Entität-Attribut-Wert-Tabelle ein ternärer Link "Merkmalsträger-Merkmal-Ausprägung", der auf die entsprechenden Entitätsschlüssel verweist. Die unmittelbare Abhängigkeit zwischen Ausprägungen und Merkmalen, die sich beispielsweise aus der Beziehung zwischen Fragebogenitems und Antwortvorgaben ergibt, wird ebenfalls als *Link* "Merkmal-Ausprägung" modelliert. Das Datenschema bildet ab, welche Entitäten miteinander in Beziehung stehen. *Hubs* und *Links* sind nach der Data Vault-Spezifikation [LO16] zeitinvariant und beschreiben daher lediglich, welche Entitätsinstanzen in der Historie gemeinsam aufgetreten sind. In der DV-Modellierung werden alle zeitvarianten Daten (z.B. Entitätsattribute) innerhalb sogenannter *Satellites* modelliert. Die Historisierung der Datensätze erfolgt über Zeitstempel, wobei Datensätze weder überschrieben noch ge-

löscht werden. Die Teilnahme an Forschungsumfragen ist anonymisiert bzw. pseudonymisiert, sodass der natürliche Schlüssel eines Merkmalsträgers abhängig von der jeweiligen Befragungsrunde ist. Der Merkmalsträger ist daher als "Teilnahme" und nicht als "Teilnehmer" zu verstehen. Die longitudinale Verknüpfung von Teilnahmen zu Merkmalsträger, deren zeitlich variierende Ausprägungen analysiert werden sollen, erfolgt durch einen weiteren *Link* "Teilnahme-Merkmalsträger". Dadurch werden Teilnahmen mit den dazugehörigen Merkmalsträgern identifiziert. Auf den *Hub* "Merkmalsträger" kann nun mit beliebigen zusätzlichen Entitäten über *Links* und die *Links* beschreibenden *Satellites* referenziert werden. *Hubs* sind somit passive Kataloge, sodass das Schema für die Befragungsdaten nicht modifiziert werden muss.

4 Implementierung

Das Systemdesign wurde vollständig mit Open Source Komponenten umgesetzt. Das System selbst wurde auf einem Ubuntu Server 16.04 mit PostgreSQL 9.6 implementiert. Die Datenintegration zwischen den einzelnen Schichten wurde mit Pentaho Data Integration 7 implementiert. Die Statistiksoftware R 3.3.2 wurde für erste longitudinale Berechnungen und Visualisierungen mit ggplot2 2.2.1 eingesetzt [WC16]. Primärdaten wurden aus LimeSurvey 2.54.3 und aus archivierten Umfragedatensätzen der Statistiksoftware SPSS von IBM aus vorherigen Jahren extrahiert. Sekundärdaten wurden aus den Krankenhaus-Qualitätsberichten des Gemeinsamen Bundesausschuss (G-BA) und dem deutschen Krankenhausverzeichnis entnommen.

SPSS bietet drei Komponenten, mit denen sowohl die Fragebogenstruktur, als auch die Umfrageergebnisse abgebildet werden können: (1) Das Datenblatt, das die Umfrageergebnisse enthält, besteht aus dem Teilnehmer (Zeile), dem Item bzw. der Variable (Spalte) und der gewählten Antwortoptionen (Zelle). (2) Die Spezifikation der Items liefert die Item-Label, Skalenniveau und Formatierungsattribute. (3) Die Spezifikation der Itemausprägungen bezeichnet Label und Kodierungen. Diese drei Komponenten können als separate CSV-Dateien exportiert werden. Das Datenblatt (1) wird durch die Metadaten aus (2) und (3) näher beschrieben.

LimeSurvey bietet einen Export der Fragebogenstruktur und der Umfrageergebnisse. Das Datenblatt für die Umfrageergebnisse ist strukturell identisch mit dem SPSS-Datenblatt und wird als CSV-Datei exportiert. Die Fragebogenstruktur wird als XML-Dokument exportiert. In diesem sind Fragetexte in Kombination mit Antworttexten zu Items instanziiert, sodass keine 1:1 Beziehung zwischen Fragetext und Item besteht. Eine Multiple-Choice-Frage mit vier Antwortoptionen korrespondiert beispielsweise mit fünf Item-Spalten im LimeSurvey-Datenblatt. Daraus folgt, dass die exportierte Fragenbogenstruktur zunächst in die semantische Struktur von Merkmalsträger, Merkmal und Ausprägung transformiert werden musste, damit Metadaten für das LimeSurvey-Datenblatt extrahiert werden können.

Gemäß der Data-Vault-Spezifikation wurden *Hubs*, *Links* und *Satellites* eins-zu-eins als

relationale Tabellen erstellt (vgl. Abb. 2). Primär- und Fremdschlüssel-Beziehungen wurden als Constraints in PostgreSQL angelegt. Der MD5-Hashing-Algorithmus wurde eingesetzt, um auf Basis des natürlichen Schlüssels (z.B. Fragencode, Antwortoption, Qualitätskennziffer) den technischen Schlüssel der Entitäten zu erzeugen. Die Datensätze werden um die Attribute *load_dts* und *rcd_src* ergänzt, welche das Erhebungsdatum der Befragung zur Historisierung der unterschiedlichen Umfragen und den vollständigen Dateipfad der jeweiligen Quelldatei enthalten. Alle Fragen werden, unabhängig vom Quellsystem, im Hub *h_item* konsolidiert. Dieser Hub hat insgesamt drei *Satellites*. Der generische *Satellite s_item* enthält die Attribute, die für Merkmale universell sind. Zusätzlich gibt es für die beiden Quellsysteme SPSS und LimeSurvey je einen *Satellite*, der systemspezifische Merkmalsattribute erfasst. Als Beispiel für die Einbeziehung von Sekundärdaten, können die frei zugänglichen Qualitätsberichte (XML-Dateien) deutscher Krankenhäuser angeführt werden. Diese können mit demographischen Daten aus dem Krankenhausverzeichnis verknüpft werden. Wie bereits bei den Befragungsdaten wurden auch hier gemäß der Data Vault-Spezifikation *Hubs* für die Entitäten, *Links* für die Beziehungen und *Satellites* für zeitabhängige und deskriptive Attribute erstellt. Die zentrale Beziehung ergibt sich hier zwischen den Entitäten Klinikstandort, Auswertungseinheit, Leistungsbereich und Qualitätsindikator.

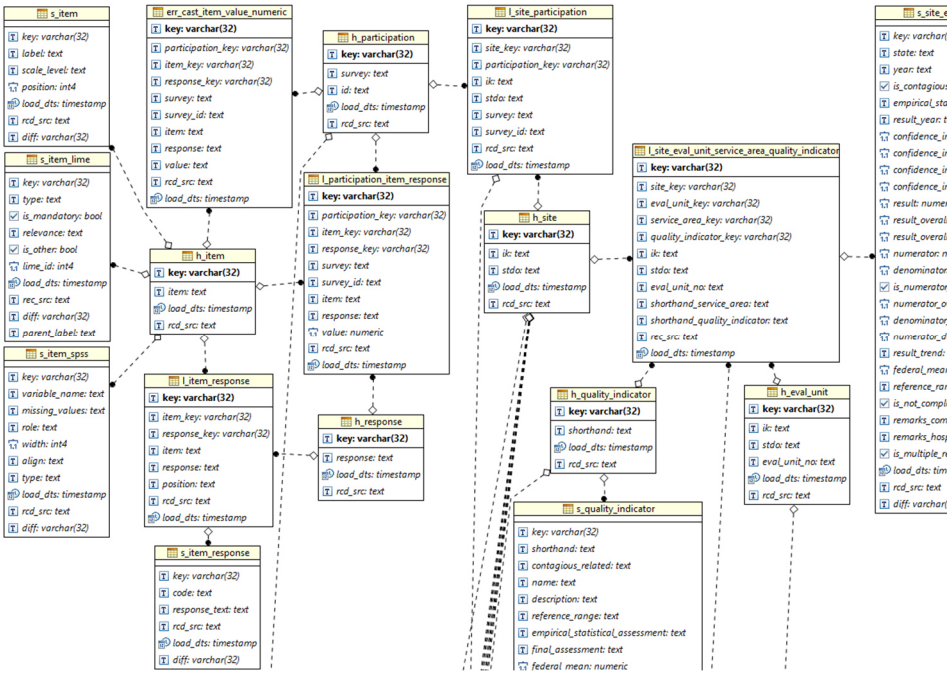

Abb. 2: Ausschnitt des Datenmodells zur Abbildung der Befragungsdaten und Qualitätsindikatoren

Das diesem Datenmodell zugrundeliegende Kern DWH (vgl. Abb. 1) enthält Daten von Befragungen aus den Jahren 2011, 2013, 2015 und 2017 (2011: 339 Teilnahmen mit 203

Items, 2013: 259 Teilnahmen mit 521 Items, 2017: 194 Teilnahmen mit 226 Items). Für die Qualitätsberichte der Jahre 2011 bis 2014 liegen insgesamt 733.000 Datenwerte für 381 Qualitätsindikatoren vor, verteilt auf insgesamt knapp 20.000 Auswertungseinheiten und 39 Leistungsbereiche.

5 Zusammenfassung, Fazit und Ausblick

Die Herausforderungen an unterschiedliche Datenmodelle für ein umfassendes Datenmanagement wurden in diesem Beitrag kritisch betrachtet. Die DV-Modellierung ist zwischen dem EAV-Ansatz und dem Sternschema einzuordnen. Dem EAV-Ansatz liegt ein generisches Datenmodell zugrunde, was jedoch zu komplexen Abfragen führt. Ferner ist das EAV-Modell sehr flexibel, kann jedoch keine semantischen Änderungen von Daten verfolgen. Sobald Attribute von Entitäten geändert werden müssen oder die Semantik von Attributen geändert wird, muss eine Erweiterung des EAV-Modells stattfinden. Das Sternschema hat hingegen festgelegte Dimensionen und Faktenwerte, wodurch Abfragen einfacher sind, jedoch die Analyseoptionen bereits festgelegt sind. Mit der DV-Modellierung steht es frei, beispielsweise die Qualitätsindikatoren als einen *Hub* zu modellieren oder unterschiedliche Qualitätsindikatoren über mehrerer *Hubs* abzubilden. Die DV-Modellierung ermöglicht umfassende Historisierungen der Datenänderungen, jedoch steigt mit der Flexibilität auch die Komplexität des Systems, sodass diese Designentscheidung konkret spezifiziert werden muss, um einem zu hohen Entwicklungsaufwand ohne zugrundeliegender Anforderungen entgegenzuwirken.

In diesem Beitrag wurde daher das EAV mit dem DV-Modell zum Persistieren semantisch heterogener Befragungsdaten (FF2) kombiniert. Das vorgestellte Modell liefert somit eine integrierte Sicht auf Datensätze unterschiedlicher Quellen, wie am Beispiel von SPSS- und LimeSurvey-Datensätzen gezeigt wurde und historisiert vollständig Befragungs- und Metadaten. Dadurch kann beispielsweise eine veränderte Semantik durch abweichende Wortlaute eines Items nachvollzogen werden und die Item-Gruppen, die sich in natürlicher Weise aus der Fragebogenstruktur aus LimeSurvey ergeben, für die zugeordneten SPSS-Items übernommen werden. Das Datenmodell kann durch neue Entitäten und Beziehungen erweitert werden. Neue Attribute oder Umfragesysteme werden durch zusätzliche *Satellites*, neue Teilnehmertypen durch neue *Links* zur zentralen Relation "Teilnahme-Merkmal-Ausprägung" implementiert.

Mit dem DV-Modell konnte der EAV-Ansatz, bestehend aus Entitäten, Attributen und Werten, generalisierbar erweitert werden, sodass alle Informationsobjekte als eigenständige Entitäten modelliert wurden. Durch diese, im Vergleich zum EAV-Ansatz, lose Kopplung der Beziehungen können unterschiedliche Informationsobjekte zusammen integriert und historisiert werden. Jede Entität wird als *Hub* mit mehreren *Links* und *Satellites* implementiert. Limitationen ergeben sich durch Änderungen von Attributspalten eines Satelliten zu eigenständigen *Hubs*, weil diese Transformationen sehr aufwändig sind. Da-

ten, welche als *Hubs* implementiert werden sollen, müssen daher im Designprozess frühzeitig definiert werden. Entitäten mit heterogenen und inkonsistenten Daten sollten daher grundsätzlich als *Hubs* modelliert werden. Das resultierende Modell ist im Ergebnis soweit generalisierbar, dass es sich auf Messergebnisse aller Art allgemein übertragen ließe, da die Relation Merkmalsträger-Merkmal-Ausprägung auf fast alle statistisch auszuwertenden Daten anwendbar ist.

Die Datenhaltung ist der erste Meilenstein in Richtung einer Gesamtarchitektur für ein umfassendes Forschungsdatenmanagement. Zukünftige Arbeiten liegen in der Vereinfachung von standardisierten, regelmäßigen Abfragen durch vordefinierte Prozesse. Dadurch soll mittelfristig der Zeitraum zwischen Erhebung der Daten und Bereitstellung der Informationen für wissenschaftliche Analysen und die Präsentation der Ergebnisse für Praxis und Politik minimiert werden. Weitere zukünftige Schritte sind die Entwicklung von Web-Front-Ends für die unterschiedlichen Teilprojekte (klinische Entscheidungsunterstützung [Hü16], Best Practices im Rahmen eines Benchmarkings zur Digitalisierung deutscher Krankenhäuser [Th14]), mit denen wissenschaftliche Ergebnisse in einer nutzerorientierten Sichtweise in die Praxis überführt werden können.

Literaturverzeichnis

[Ag10] Agarwal, R. et al.: Research Commentary —The Digital Transformation of Healthcare. Current Status and the Road Ahead. In Information Systems Research, 2010, 21; S. 796–809.

[Am09] Amarasingham, R. et al.: Clinical information technologies and inpatient outcomes: a multiple hospital study. In Archives of internal medicine, 2009, 169; S. 108–114.

[Ba13] Bauer, A. Hrsg.: Data-Warehouse-Systeme. Architektur, Entwicklung, Anwendung. dpunkt-Verl., Heidelberg, 2013.

[Bo16] Bojicic, I. et al.: A comparative analysis of data warehouse data models. In (Dzitac, I.; Filip, F. G.; Manolescu, M.-J. Hrsg.): 2016 6th International Conference on Computers Communications & Control (ICCCC). Hotel President, Băile Felix, Oradea, Romania, May 10-14, 2016. IEEE, Piscataway, NJ, 2016; S. 151–159.

[Bu11] Buntin, M. B. et al.: The benefits of health information technology: a review of the recent literature shows predominantly positive results. In Health affairs (Project Hope), 2011, 30; S. 464–471.

[Ch10] Chamoni, P. Hrsg.: Analytische Informationssysteme. Business Intelligence-Technologien und -Anwendungen. Springer, Berlin u.a., 2010.

[DN07] Dinu, V.; Nadkarni, P.: Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. In International journal of medical informatics, 2007, 76; S. 769–779.

[Fo02] Forschungsgruppe Informatik im Gesundheitswesen der Hochschule Osnabrück: IT-Report Gesundheitswesen. http://www.it-report-gesundheitswesen.de/, 03.05.2017.

[FWB10] Friedman, C. P.; Wong, A. K.; Blumenthal, D.: Achieving a nationwide learning health system. In Science translational medicine, 2010, 2; 57cm29.

[He09] Hey, T. Hrsg.: The fourth paradigm. Data-intensive scientific discovery. Microsoft Research, Redmond Washington, 2009.

[Hü16] Hübner, U. et al.: ROSE – the learning health care system in the Osnabrück-Emsland / ROSE – das lernende Gesundheitssystem in der Region Osnabrück-Emsland. In International Journal of Health Professions, 2016, 3.

[In05] Inmon, W. H.: Building the data warehouse. Wiley, Indianapolis Ind., 2005.

[JSM14] Jovanovic, V.; Subotic, D.; Mrdalj, S.: Data modeling styles in data warehousing: 2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). IEEE, 2014; S. 1458–1463.

[KR13] Kimball, R.; Ross, M.: The data warehouse toolkit. The definitive guide to dimensional modeling. Wiley, Indianapolis Ind., 2013.

[LBK07] Lenz, R.; Beyer, M.; Kuhn, K. A.: Semantic integration in healthcare networks. In International journal of medical informatics, 2007, 76; S. 201–207.

[Lö12] Löper, D. et al.: Integrating Healthcare-Related Information Using the Entity-Attribute-Value Storage Model. In (He, J. et al. Hrsg.): Health information science. First international conference, HIS 2012, Beijing, China, April 8 - 10, 2012 ; proceedings. Springer, Berlin, 2012; S. 13–24.

[LO16] Linstedt, D.; Olschimke, M.: Building a scalable data warehouse with Data Vault 2.0, 2016.

[Me12] Meineke, F. et al. Hrsg.: Medizinische Forschungsdatenbanken als Baustein des Forschungsdatenmanagements an der Universität Leipzig. German Medical Science GMS Publishing House, 2012.

[Mu10] Murphy, S. N. et al.: Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). In Journal of the American Medical Informatics Association JAMIA, 2010, 17; S. 124–130.

[OAM07] Olsen, L.; Aisner, D.; McGinnis, J. M. Olsen, L.; Aisner, D.; McGinnis, J. M.: Institute of Medicine (US). Roundtable on Evidence-Based Medicine. The learning healthcare system. Workshop summary. Washington, DC: National Academies Press, 2007.

[Po15] Pommerening, K. et al.: Der Impact der Medizinischen Informatik. In Informatik-Spektrum, 2015, 38; S. 347–369.

[Th14] Thye, J. et al.: IT-benchmarking of clinical workflows: concept, implementation, and evaluation. In Studies in health technology and informatics, 2014, 198; S. 116–124.

[WC16] Wickham, H.; Chang, W.: ggplot2: Create Elegant Data Visualisation Using the Grammar of Graphics. Springer, 2016

[WHM11] Wade, T. D.; Hum, R. C.; Murphy, J. R.: A Dimensional Bus model for integrating clinical and research data. In Journal of the American Medical Informatics Association JAMIA, 2011, 18 Suppl 1; S. 96-102.

[Ya12] Yamamoto, K. et al.: A pragmatic method for electronic medical record-based observational studies: developing an electronic medical records retrieval system for clinical research. In BMJ open, 2012, 2.