MUSTANG: Realisierung eines Analytischen Informationssystems im Kontext der Gesundheitsberichtserstattung

Yvette Teiken, Martin Rohde, Matthias Mertens

OFFIS - Institut für Informatik
Escherweg 2, 26121 Oldenburg, Germany
{teiken|rohde|mertens}@offis.de

Abstract: Die Aufgaben in der Gesundheitsberichterstattung sind vielfältig und komplex, so dass informationstechnische Unterstützung notwendig ist. Zu den Aufgaben gehört die Integration verschiedener Datenquellen und die Berechnung statistischer Kennzahlen auf den integrierten Datenbeständen. Diese Kennzahlen bilden die Grundlage tagesaktueller, wöchentlicher oder jährlicher Gesundheitsberichte. Für diese Aufgaben bietet sich die Verwendung eines Analytischen Informationssystems an, welches multidimensionale Daten mit statistischen Berechnungen und Geo-Informationen verbindet. In diesem Artikel wird mit der MUSTANG eine Plattform für die Entwicklung Analytischer Informationssysteme vorgestellt. MUSTANG stellt die Grundlage für eine Vielzahl von Analyse-Anwendungen für den Gesundheitsmarkt und insbesondere für die automatisierte Gesundheitsberichterstattung im Öffentlichen Gesundheitsdienst dar.

1 Einleitung und Anforderungen an Analytischen Informationssysteme für GBE

Gesundheitsberichterstattung (GBE) bezeichnet die Aufbereitung und Darstellung gesundheitlich relevanter Aspekte mit Bevölkerungsbezug. GBE dient der Information von Akteuren im Gesundheitssystem, von Politikern, Forschern und interessierten Laien. Auf Landesebene bilden die Beratung und Unterstützung der Landesregierungen, Behörden, Einrichtungen oder Kommunen in Fragen der Gesundheit, der Gesundheitspolitik sowie der Sicherheit und des Gesundheitsschutzes in der Arbeitswelt zentrale Ziele der GBE. In Nordrhein-Westfalen wird die GBE vom Landesinstitut für Gesundheit und Arbeit (LIGA. NRW) durchgeführt. Zu den Aufgaben der GBE gehört es, verschiedene Datenquellen zu integrieren, mit Hilfe der Daten den Zustand der Gesundheit und der Versorgung zu beobachten, sowie Analysen und Berichte zur Gesundheitssituation bereitzustellen. Des Weiteren werden Risiken benannt und gegebenenfalls Warnungen ausgesprochen.

In der Fachgruppe „Infektiologie und Hygiene“ werden wöchentliche Berichte mit aufbereiteten Informationen zu meldepflichtigen Infektionen generiert und der Öffentlichkeit auf einem Portal bereitgestellt. In der Fachgruppe „Gesundheitsinformationen“ werden

jährliche Berichte zu den Indikatoren der Ländergesundheitsberichterstattung veröffentlicht. Kommunale Indikatoren werden zusätzlich interaktiv als Gesundheitsatlas und in Form von vergleichenden Kreisprofilen bereitgestellt. Neben der Veröffentlichung der Indikatoren versuchen die Experten in den Fachgruppen, durch die zeitnahe Analyse der Daten - im Kontext der Infektionsepidemiologie werden die Daten tagesaktuell ausgewertet - Gesundheitsrisiken frühzeitig zu erkennen. Fasst man die drei beschrieben Szenarien der Gesundheitsberichtserstattung zusammen, so ergeben sich eine Reihe von Anforderungen, die es in einem ganzheitlichen Analytischen Informationssystem auf Basis von Data Warehouse Technologien zu unterstützen gilt. Diese werden im Folgenden weiter ausgeführt.

Grundlage aller Szenarien muss ein qualitätsgesicherter, integrierter Datenbestand sein, der sich aus verschiedenen Datenquellen, insbesondere amtlichen Statistiken, Statistiken der Akteure des Gesundheitssystems (z. B. Krankenkassen u.a.) und Befragungen oder andere Erhebungen zusammensetzt. Darauf basierend können komplexe Kennzahlen definiert werden, die Daten aus verschiedenen Quellen nutzen und somit neue Analysen und Erkenntnisse ermöglichen. Die integrierte Datenhaltung sollte in Form eines multidimensionalen Datenmodells umgesetzt werden, wodurch verschiedene Kennzahlen (z.B. Anzahl meldepflichtiger Infektionen) mit OLAP Operationen in unterschiedlichen Dimensionen (Region, Krankheit, Zeit) und Aggregationsstufen (Monat, Tag) analysiert und bereitgestellt werden können.

Für die Generierung von Berichten ist es erforderlich, dass geeignete Systeme zur Berechnung von Kennzahlen, zur Informationsvisualisierung und zur Veröffentlichung der Informationen in geeigneter Form (PDF, HTML) genutzt werden können. Berichte zu aktuellen Entwicklungen bei meldepflichtigen Infektionskrankheiten sollen in tabellarischer und grafischer Form in kurzen wöchentlichen Zyklen voll automatisiert veröffentlicht werden. Die jährlichen Berichte zu den Indikatoren der Ländergesundheitsberichterstattung enthalten dagegen sehr viele (ca. 400) Kennzahlen, die zum Teil komplexe statistische Verfahren abbilden, weshalb eine Teilautomatisierung der Indikatorenerstellung angestrebt wird.

Um es den Experten der Fachgruppen zu ermöglichen, eigenständige Analysen auf den integrierten Daten durchführen zu können, müssen diese weitestgehend von automatisierbaren Routinetätigkeiten im Kontext der Berichterstellung entlastet werden. Die Datenintegration aus externen Quellen sollte automatisiert und effektiv durchgeführt werden.

Neben Werkzeugen zur Automatisierung der GBE sollten auch Analysewerkzeuge bereitgestellt werden, das es erlaubten, die Daten multidimensional explorativ zu untersuchen und auch räumlich statistische Analyseverfahren beherrscht. Räumliche Clusterverfahren als ein Beispiel von räumlich-statistischen Verfahren sind notwendig, um die Ausbreitung von Epidemien besser beobachten und Gegenmaßnahmen einleiten zu können.

Im LIGA.NRW ist ein analytisches Informationssystem mit MUSTANG als Datenanalyseplattform eingeführt worden, das die oben genannten Szenarien und damit verbundenen Anforderungen erfüllt. Das System wurde am Informatinstinstitut OFFIS entwickelt.

2 Die MUSTANG Plattform

Das Akronym MUSTANG steht für Multidimensional Statistical Data Analysis Engine und beschreibt eine Analyseplattform, die sich durch die folgenden drei Haupteigenschaften auszeichnet.

Multidimensional: Daten, die mittels der MUSTANG Plattform für Analysen verwendet werden, sind multidimensional aufbereitet. Dies ermöglicht die Verwendung des OLAP-Paradigmas und somit die intuitive interaktive Analyse.

Erweiterte Statistik: Für die Analysen stehen vielfältige erweiterte statistische Verfahren zur Verfügung, deren Ursprung in der Epidemiologie liegen. Neben Berechnungen einfacher Kennzahlen für Inzidenzraten sind auch komplexe Verfahren zur Auswertung von zum Beispiel Überlebenszeitwahrscheinlichkeiten oder die Identifizierung von Clustern realisiert.

Geographisch: Auf Daten mit Geografiebezug können räumlich statistische Verfahren angewandt werden. Hierbei unterstützt die Plattform sowohl kleinräumige-, wie auch Flächenanalysen.

Die Ursprünge der MUSTANG Plattform liegen im Projekt CARLOS, welches für das Krebsregister des Landes Niedersachsen zuständig war. In diesem Projekt, welches im Jahr 1993 begann, wurden Komponenten wie OLAP-Server und Geodatenbank als Eigenentwicklung realisiert. Mangels Standardkomponenten und gängigen Austauschformaten konnte auf keine existierende Software zurückgegriffen werden.

Mit der Neuentwicklung von MUSTANG als Plattform wurde im Jahr 2007 begonnen. Da sich der Markt in der Zwischenzeit verändert hat, konnte bei der Neurealisierung auf Standardkomponenten zurückgegriffen werden. Zu diesen gehören eine Geographie-Datenbank für die Speicherung geografischer Daten, ein OLAP-Server für die Speicherung und Auswertung multidimensionaler Daten und eine Statistik-Komponente für die Realisierung komplexer statistischer Verfahren.

2.1 Beschreibung der MUSTANG Plattform

Bei der MUSTANG Plattform handelt es sich um eine rekonfigurierbare serviceorientierte Architektur [KMR03]. Das zentrale Element der Plattform stellt der sog. MUSTANG Servicelayer dar, der die drei Anwendungsblöcke *Geo Services*, *Multidimensional Data Services* und *Statistical Services* miteinander verknüpft. Jeder der drei Anwendungsblöcke ist eine Komponente mit abgeschlossenem Funktionsumfang.

Die Services innerhalb des Anwendungsblocks *Multidimensional Data Services* kapseln den Zugriff auf den OLAP-Server. Der Dimension Service ist für die Abfragen von Dimensionen und deren Elementen zuständig, der Cube Service für das Abfragen von Cubes innerhalb des OLAP-Servers. Der eigentliche Zugriff auf den OLAP-Server erfolgt mittels der Abfragesprache XMLA. Dies ermöglicht es auf einfache Weise andere OLAP-Server

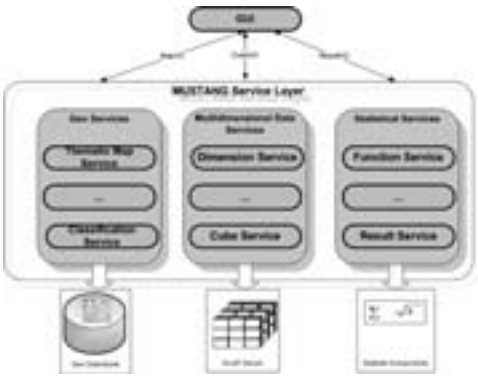

Abbildung 1: Architektur der MUSTANG Services

an die Plattform anzubinden. Für manche Aspekte des Zugriffs müssen jedoch über XMLA hinaus Anpassungen vorgenommen werden. Diese werden ebenfalls innerhalb dieses Anwendungsblocks realisiert. Unterstützte OLAP-Server der OLAP-Plattform sind zur Zeit Microsoft Analysis Services, Palo und Mondrian.

Innerhalb des Anwendungsblocks *Statistical Services* werden die von MUSTANG unterstützten statistischen Verfahren realisiert. Diese Verfahren basieren auf Berechnungen mit OLAP Cubes. Bei komplexen Kennzahlen bzw. Verfahren können verschiedene Cubes miteinander verrechnet werden. Diese Verrechnung findet nicht nur auf Zell-Ebene statt, sondern kann auch Teile von Cubes beinhalten. Zur Beschreibung der Eigenschaften werden in diesem Anwendungsblock die grundsätzlichen strukturellen Abbildungen definiert. Die eigentlichen Berechnungen werden nicht innerhalb der Plattform durchgeführt, sondern mit der Statistik-Komponente R. Deswegen enthält dieser Anwendungsblock auch Services zur Umwandlung von Cubes in R-Strukturen und umgekehrt. Die Verwendung von R hat den Vorteil, dass R eine große Anzahl von relevanten statistischen Funktionen bereits enthält, und dass Verfahren von Statistikern direkt in R realisiert werden können.

Im Anwendungsblock *Geo Services* werden Funktionalitäten zum Umgang mit geografischen Daten umgesetzt. Hierzu gehört neben der Anfrage von Geoobjekten aus einer Geodatenbank auch Services zum Erzeugen einer thematischen Karte, bei der Kennzahlen mit Geoobjekten verknüpft werden. Zurzeit wird PostGIS als Geodatenbank verwendet.

In der Abbildung 1 ist die MUSTANG-Architektur abgebildet. Die Services sind zustandslos. Die Daten werden innerhalb der Plattform über so genannte Datentransferobjekte nach dem DTO-Pattern ausgetauscht. Diese Objekte beschreiben den Zustand des Systems.

2.2 MUSTANG als Basis von Analyseanwendungen

Die Informationslogistik für die GBE im LIGA.NRW basiert auf einer Hub-and-Spoke-Architektur mit einem Data Warehouse (DWH) als zentrale, integrierte, bereinigte, qua-

litätsgesicherte Datenbank. Dieses DWH beinhaltet alle notwendigen Daten für die Indikatoren der Ländergesundheitsberichterstattung und die Infektionsberichte, und bildet die Grundlage der MUSTANG-Plattform, sowie der auf Basis der MUSTANG-Plattform erstellten Anwendungen (vgl. Abb. 2). Technologisch ist das DWH im Rahmen einer SQL Server 2005-Infrastruktur im LIGA.NRW umgesetzt worden. MUSTANG bildet die Plattform zur Konfiguration sogenannter Berichtsmappen, in denen verschiedene Analysen zusammengefasst werden. Analysen in MUSTANG beinhalten Kennzahlen wie standardisierte Inzidenzraten, die Dimensionalität wie z.B. die Einschränkung auf die Krankheit Masern. Weitere Beispiele für die Dimensionalität einer Analyse bilden die Landkreise in NRW, das Diagnosejahr 2007 und die Einschränkung auf die Altersgruppe der 8-10-jährigen Kinder. Zu einer Analyse gehört auch die Art der Visualisierung wie Diagramme, Tabellen und thematische Karten. Kommunale Indikatoren, also Indikatoren auf Landkreisebene, werden in der GBE häufig über thematische Karten dargestellt, die Indikatoren der Ländergesundheitsberichterstattung stellen dagegen meist mehrere Kennzahlen in Form von Tabellen mit einem länderübergreifend vorgegebenen Tabellenlayout nebeneinander dar. Auch die Layoutinformationen werden in die Analysen hineinkodiert.

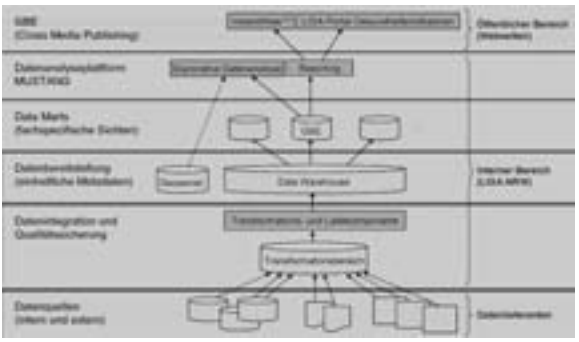

Abbildung 2: Architektur der Informationslogistik für die GBE im LIGA.NRW

Für das LIGA.NRW sind vier Anwendungen auf Basis der MUSTANG-Plattform realisiert worden: Mit AIM+ (Automatisiertes Infektionskrankheiten Meldesystem) werden wöchentliche Infektionsberichte erstellt. Mit der Anwendung „Gesundheit NRW“ wird ein Großteil der Indikatoren der Ländergesundheitsberichterstattung erstellt. Daneben sind mit EARL (Early Warning System) ein Expertenwerkzeug zur Überwachung der Infektionskrankheiten und mit dem INITIAL-System ein Analysewerkzeug zur Beantwortung von Ad-Hoc-Anfragen und für explorative Datenanalysen im LIGA.NRW umgesetzt worden.

Die Erstellung der wöchentlichen Infektionsberichte erfolgt mit AIM+ vollautomatisch. Alle Prozessschritte werden von einem Prozess-Scheduler verwaltet und automatisch angestoßen: die Extraktion aus der SurvNet-Anwendung, einer vom Robert-Koch-Institut (RKI) zur Bearbeitung der Daten nach dem Infektionsschutzgesetz entwickelten Programm; das Laden der Daten in das DWH; die Kennzahlenberechnung; der Ergebnisexport; die Erstellung von HTML-Seiten. Die Erstellung der zu veröffentlichenden HTML-Seiten erfolgt über einen XSLT-Prozessor, welcher die von MUSTANG im XML-Format exportie-

ren Analyseergebnisse mit Hilfe eines XSLT-Skripts, in HTML rendert.[1]

Für die Fachgruppe „Gesundheitsinformationen“ ist ein Datenmanagementwerkzeug „Gesundheit NRW“ entwickelt worden, über das die Prozessschritte zur Erstellung der Indikatoren der Ländergesundheitsberichterstattung gesteuert werden können. Der erste Prozessschritt ist die Extraktion, Transformation und das Laden der Daten in das DWH, also die Implementierung sogenannter ETL-Prozesse für die verschiedenen Rohdaten, die zur Berechnung benötigt werden - Daten wie Todesursachen-, Diagnose-, Pflegestatistiken, sowie über die Arbeitsunfähigkeit von Arbeitnehmern. Datenquellen für diese Rohdaten sind der Landesbetrieb Information und Technik (IT.NRW), die Deutschen Rentenversicherungsträger, die Betriebskrankenkassen und andere Einrichtungen. Die ETL-Prozesse sind mit SQL Server-Technologien als parametrisierte SSIS-Packages (SQL Server Integrations Service) realisiert worden. Die Veröffentlichung der Indikatoren erfolgt ähnlich wie bei AIM+ auf fest definierten Analysen, die auf dem integrierten Datenbestand durchzuführen sind und somit vorkonfiguriert werden. Weitere Prozesschritte im Rahmen der Indikatorerstellung sind die Anpassung von Analyseparametern wie dem Berichtsjahr und - wie bei AIM+ - die Berechnung, der Export und die HTML-Ausgabe der Ergebnisse.

Während die beiden beschriebenen Anwendungen nur wenig Interaktion zulassen und zur Automatisierung der Berichterstellung verwendet werden, stellen EARL und das INITIAL-System genau diese Interaktionsfunktionalität zur Verfügung. Sie ermöglichen das Monitoring und die Exploration des integrierten Datenbestands innerhalb des DWH.

3 Zusammenfassung und Ausblick

In diesem Artikel wurde gezeigt, wie die Anforderungen an die GBE in Form eines Analytischen Informationssystem nach [CG06] umgesetzt werden kann. Hierbei wurde gezeigt, wie auf Basis der MUSTANG Plattform verschiedene Anwendungen zur Verfügung gestellt worden sind, die alle auf demselben integrierten Datenbestand basieren. Die genannten Anwendungen werden im LIGA verwendet. In Zukunft sollen weitere Daten in das System integriert werden. Hierzu zählen Daten zur Arbeitswelt in NRW bzw. Daten nach dem Psychisch-kranken-Gesetz und Betreuungsrecht. Weiterhin soll der jetzt schon hohe Grad der Automatisierung von Aufgaben noch weiter erhöht werden.

Literatur

[CG06] Peter Chamoni und Peter Gluchowski. *Analytische Informationssysteme. Business Intelligence-Technologien und -Anwendungen: Business Intelligence-Technologien Und -Anwendungen*. Springer, Berlin, 2006.

[KMR03] S. Koch, J. Meister und M. Rohde. MUSTANG – A framework for Statistical Analyses of Multidimensional Data in Public Health. In A. Gnauck und R. Heinrich, Hrsg., *17th International Conference Informatics for Environment Protection*, Seiten 635–642, 2003.

[1] siehe http://www.liga.nrw.de/themen/gesundheit_berichte_daten/gesundheitsindikatoren/