Richtlinie zum „Software-Sprint“

Klinik-DB:
Die freie Krankenhaus Datenbank

Schlussbericht

Zuwendungsempfänger: **Köther-Schubert-Rentzsch-Molitor-GbR**

Das diesem Bericht zugrundeliegende Vorhaben wurde mit Mitteln des Bundesministeriums für Bildung und Forschung unter dem Förderkennzeichen **01IS21S08** gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autoren (Ulf Köther, Malte Schubert, Steven Rentzsch und Philipp Molitor).

Kurze Darstellung der Aufgabenstellung und Motivation

Im Jahresverlauf 2020 wollten wir uns aus Eigeninteresse einen umfassenden Einblick in die deutsche Krankenhauslandschaft der letzten Jahre verschaffen, insbesondere durch die Nutzung der sogenannten Qualitätsberichte der Krankenhäuser (QBK). In kurzen Worten zusammengefasst beinhalten diese Berichte detaillierte Angaben jeder deutschen Klinik, die in die Planung zur Versorgung gesetzlich versicherter Patienten einbezogen ist. Eine jede solche Klinik muss diese Berichte standardisiert und jährlich an den Gemeinsamen Bundesausschuss (GBA) übermitteln. Von diesem kann auf schriftlichen Antrag hin eine Kopie dieser Berichte bezogen werden. In der Gesundheitswirtschaft wurde zwar schon lange kolportiert, dass einige technische Hürden auf dem Weg zur sinnvollen Nutzung dieser Daten existieren, aber bis zu unserem eigenen Versuch in dieser Angelegenheit war uns nicht bewusst, wie aufwändig sich dieser Prozess eigentlich darstellt. Da die Berichte nur im XML-Format auf DVD an die Beantragenden geliefert werden, liegen sie nicht in einem für eine berichtsübergreifende Recherche oder weitere Analyse geeigneten Format vor. Andere Quellen für QBK-Daten existieren zwar, sind aber entweder nur in aggregierter Form (z.B. Destatis), oder nicht Klinikvergleichend bzw. nicht maschinenlesbar abrufbar (z.B. „Weisse Liste“). Daher haben wir uns mit dem Projekt Klinik-DB folgende Ziele gesetzt: Wir wollten möglichst viele Berichtsjahre vom GBA anfordern (die ursprüngliche Planung sah 10 Jahre vor), um sie automatisiert und validiert in eine SQL-Datenbankstruktur zu überführen, und diese SQL-Datenbank dann Dritten zur Verfügung stellen, einmal per Datenbank-Dump, also zur Reproduktion der gesamten Daten auf einem eigenen Datenbankserver, und einmal per REST-API, also als strukturierte Programmierschnittstelle, damit z.B. Softwareanbieter oder andere interessierte Dritte Daten aus den QBK für eigene Anwendungen nutzen können. Die Software für diesen Prozess der Datenextraktion aus dem originalen XML-Dateien wollten wir veröffentlichen, so dass in Zukunft auch Andere daran partizipieren bzw. davon profitieren können. Zusätzlich wollten wir über vorgefertigte Analyseumgebungen interessierten Nutzern Beispiele an die Hand geben, was dann mit den Rohdaten an Analysen möglich ist. Die ursprünliche Planung sah vor, hierfür containerisierte Beispiele mit R (Shiny) und in Python (Bokeh oder Dash) zur Verfügung zu stellen.

Geplante Meilensteine im Förderzeitraum:

- Anfordern von 10 Jahren an Qualitätsberichten der Krankenhäuser vom GBA
- Erstellung eines Software-Parsers, der das automatisierte Einlesen und Überführen der Daten in eine SQL-Datenbank übernimmt
- Damit einhergehend Erarbeitung eines entsprechenden Datenbankmodells
- Identifikation und Bereitstellung von zusätzlichen Datensätzen, die im Zusammenspiel mit den QBK inhaltlich relevante Analysen ermöglichenn (z.B. Bevölkerungszahlen)
- Bereitstellen der gesamten Daten als Download und über eine REST-API
- Bereitstellung von Beispielanalysen in R und Python als Docker-Container, so dass interessierte Nutzer die Analysen innerhalb einer lokalen Analyseumgebung nutzen können

Beitrag des Projektes zu den Zielen der Förderinitative „Software-Sprint“

Spätestens durch die Corona-Krise wurde nochmal der unmittelbare Zusammenhang zwischen politischen Entscheidungen über die deutsche Kliniklandschaft und den daraus resultierenden gesellschaftlichen Auswirkungen klar ersichtlich. Es wurde aber auch deutlich, dass es nur auf Basis von Daten und Fakten möglich ist, solche politische Entscheidungen fundiert von außen zu beurteilen, und die dafür notwendige Datengrundlage muss für die Allgemeinheit transparent und einfach verfügbar sein.

Beim Förderprogramm „Software Sprint“ liegt der Fokus[1] auf „kreativen Ideen mit Bezug zu gesellschaftlich relevanten Lösungen in einer datengetriebenen Welt, […] bei der innovative Modelle und Werkzeuge zur Erzeugung, zum Teilen und zur Nutzung von Daten gefragt sind. Die Lösungen sollen es ermöglichen, dass sich die Öffentlichkeit stärker an gesellschaftlichen Entwicklungen beteiligen kann und beim souveränen Umgang mit Daten und Informationen unterstützt wird.“

Vor dem Hintergrund dieser Definition glauben wir klar sagen zu können, mit dem Projekt Klinik-DB den Anspruch an durch den „Software Sprint“ zu fördernde Projekte zu erfüllen: Wir haben gesellschaftlich relevante Daten einer breiteren Öffentlichkeit zugänglich gemacht, haben die Allgemeinheit in die Lage versetzt, mit diesen Daten souverän umzugehen, und haben Werkzeuge veröffentlicht, die diesen Prozess in Bezug auf die QBK-Daten auch in Zukunft vorantreiben können.

Ausführliche Darstellung der Ergebnisse

Der erste Schritt war das Schreiben eines R-Pakets (https://klinik-db.gitlab.io/qbgbaReader), womit die XML-Berichtsdateien automatisiert eingelesen und geparsed werden können. Parallel haben wir ein einheitliches SQL-Datenbankschema erstellt (ebenfalls im Sourcecode von qbgbaReader verfügbar) und die Daten in einer entsprechenden Datenbank hinterlegt. Dieser Vorgang wurde für die letzten fünf verfügbaren Berichtsjahre (2015 bis 2019) vollzogen. Viel Arbeit erforderte hierbei das Matchen eines Krankenhauses über die verschiedenen Jahre hinweg, da sich herausstellte, dass dies nicht automatisiert möglich ist, da die einzelnen Berichte keine reliable übergreifende Identifikationsmöglichkeit bieten. Hier war leider „Handarbeit“ notwendig, um die Informationen so miteinander zu verknüpfen, dass die Kliniken über die Jahre verfolg- und vergleichbar sind. Einen

[1] Adaptiert von https://www.softwaresysteme.pt-dlr.de/de/software-sprint.php, abgerufen am 13.09.2021.

Anfangspunkt lieferten uns zwar die sogenannten „Positivlisten“ des GBA, die zwar einen übergreifenden ID-Code liefern, diesen aber leider wiederum nicht auf die Daten in den eigentlichen Berichtsdateien matchen, so dass dies per Hand erfolgen musste.

Da uns zur Zeit der Förderbeantragung nur ein Berichtsjahr zur Verfügung stand, haben wir leider im Vorfeld den Umfang dieser Arbeit eklatant unterschätzt, vor allem auch deshalb, weil ältere Berichtsjahre sich in ihrer Datenqualität umso schlechter darstellen desto älter sie sind. Als Nebenbemerkung sei angeführt, dass wir im Nachhinein auch erfahren haben, dass der GBA seit dem Berichtsjahr 2015 mit einem neuen IT-Dienstleister zusammenarbeit, was ab diesem Berichtsjahr auch die allgemeine Datenqualität spürbar verbessert hat. Jedoch mussten wir aufgrund dieses Umstands unsere ursprüngliche Zeitplanung hinsichtlich des leistbaren Umfangs der Projektarbeit anpassen, weswegen wir in Absprache mit unserer Projektbegleitung und dem Projektträger die einzulesenden Jahre von zehn auf fünf (2015 bis 2019) reduziert haben. Vor dem gleichen Hintergrund haben wir die ursprüngliche Planung, die geplanten Beispielanalysen in R und in Python zu erstellen, auf eine Programmiersprache reduziert, d.h. auf die Nutzung von R und Shiny. So konnten wir zwar nicht alle Meilensteine in ihrem vollem Umfang wie geplant erfüllen, haben aber doch jede der avisierten Grundfunktionalitäten bereitstellen können.

Um die Nutzbarkeit und den Informationsgehalt von Klinik-DB weiter zu steigern, wurden die GBA-Daten wie geplant mit weiteren Drittdaten „angereichert”, die sich für weitere Analysen einfach als weiteres R-Paket (https://klinik-db.gitlab.io/qbgbaExtraData) installieren lassen. Damit sich alle User einen Eindruck von den Möglichkeiten der Klinik-DB machen können, haben wir bereits verschiedene Beispielanalysen erstellt, die sich mit den Daten realisieren lassen. Diese haben wir über einen Docker-Container bereitgestellt, in dem eine R-Analyse-Umgebung und eine Shiny-basierte interaktive Weboberfläche enthalten ist (https://gitlab.com/klinik-db/qbgbaAnalysis). Damit besteht z. B. die einfache Möglichkeit, sich die Kliniken auf einer interaktiven Karte anzeigen und sich zu jeder Klinik zusätzliche Informationen über die verfügbaren Jahre ausgeben zu lassen. Neben diesem nutzer*innenfreundlichen Analysetool steht ein kompletter Daten-Download für die Jahre 2015 bis 2019 im SQL-Format zur Verfügung: Sämtliche eingelesenen Daten können als vollständiger SQL-Dump (https://gitlab.com/klinik-db/qbgbaSQLData) heruntergeladen werden, so dass sich die SQL-Datenbank mit allen Informationen problemlos reproduzieren lässt. Da die Gesamtdaten mit mehr als 2 GB jedoch zu groß für einen Docker-Container sind, haben wir einen kleineren Auszug aus den Daten für die oben genannten Analysen zusammengestellt und bieten diesen auch in unserem R-Datenpaket (https://klinik-db.gitlab.io/qbgbaExtraData) an. Darüber hinaus haben wir für die gesamte Datenbank eine maschinenlesbare REST-API erstellt, über die (aus Gründen der Serverkapazität mit gewissen Limits pro Request) jede Tabelle der Datenbank abgefragt werden kann. Für den API-Zugang genügt eine einfache Anmeldung (https://klinik-db.de/site/klinikdb/api). Schlussendlich stellen wir eine Websuche zur Verfügung, über die einzelne Krankenhäuser gesucht und verschiedene Datenpunkte aus den Berichten direkt eingesehen werden können (https://klinik-db.de). Die Dokumentation für alle diese Softwarebestandteile ist auf den jeweiligen Webseiten verfügbar.

Wir haben in der Förderzeit sehr viel Detailwissen über die Struktur (und die Sinnhaftigkeit?) der QBK gewinnen können, und auch unsere Fähigkeiten im Umgang mit XML-Strukturen wurde zwangsweise und über alle Maßen trainiert. Insgesamt müssen wir klar sagen, dass wir nicht nur über die Förderung durch das BMBF, sondern auch über die Unterstützung durch die Open Knowledge Foundation und die Begleitung durch den Projektträger DLR sehr dankbar waren und von der strukturierten Anleitung profitiert haben.

Zielgruppe, Nutzen und mögliche Weiterentwicklungen

Welcher Nutzen ergibt sich für die Zielgruppe aus den Ergebnissen Deines Projekts? Welche weitergehenden Effekte ergeben sich aus der Open-Source-Stellung der Ergebnisse? Gibt es Ideen für die Weiterentwicklung Deiner Lösung und Pläne zu deren Umsetzung?

Da wir den Programmcode für unseren XML-Parser unter einer Open-Source-Lizenz verfügbar machen (R-Package „qbgbaReader“, https://klinik-db.gitlab.io/qbgbaReader/), kann das Einlesen der Daten aus den XML-Dateien in eine Datenbank von allen, die den Code nutzen, reproduziert werden, solange sie die GBA-Rohdaten zur Verfügung haben. Außerdem kann der Parser an zukünftige Änderungen an der Berichtsstruktur einfach angepasst und entsprechend auch in der Zukunft verwendet werden. Damit ist der zukünftige Arbeitsaufwand für uns und auch andere interessierte Nutzer, die Daten ab dem Berichtsjahr 2020 einlesen wollen, stark reduziert. Wir gedenken, diese Arbeit auch für zukünftige Berichtsjahre zu leisten (das Berichtsjahr 2020 wird in den kommenden Monaten vom GBA veröffentlicht), und auch die aktualisierte Fassung des Parsers frei verfügbar zu machen, so dass eine Weiterentwicklung des Projekts gegeben ist.

Die Arbeit im Projekt hat uns auch in fachlicher Hinsicht weitergebracht, denn wir haben aufgrund des Zeitdrucks doch relativ viele Bausteine in kurzer Zeit und enger Zusammenarbeit erstellen müssen. Das konsequente Nutzen der entsprechenden Tools und ein engmaschiges Projektmanagement waren dabei von großem Vorteil, und da wir dazu konsequent gezwungen wurden, haben wir uns auch hier inhaltlich und fachlich weiterentwickeln können.

Kurze Darstellung der Arbeiten, die zu keiner Lösung geführt haben

In dieser Hinsicht gibt es bei unserem Projekt nichts zu berichten, da die technischen Probleme alle gelöst werden konnten, und wir keine Lösungsansätze aufgrund mangelnder Eignung verwerfen mussten.

Kurze Angabe von Präsentationsmöglichkeiten für mögliche Nutzer

Wir haben unsere Ergebnisse in mehrfacher Form veröffentlicht: Als erstes ist die Website https://klinik-db.de zu nennen, wo Nutzer anhand einer Web-GUI nach Krankenhäusern suchen und sich aktuelle Daten aus den QBK anzeigen lassen können. Als zweites stehen unter https://gitlab.com/klinik-db alle erstellten Softwarepakete inkl. Dokumentation zur Verfügung.[2]

Kurze Erläuterung zur Einhaltung der Arbeits- und Kostenplanung

Wie schon oben beschrieben, haben wir aufgrund der Tatsache, dass wir erst nach Projektbeginn alle Rohdaten einsehen konnten, und somit weder von der fehlenden Identifizierungsmöglichkeit der Kliniken über die Berichtsjahre hinweg und von der schlechten Datenqualität älterer Berichtsjahre keine Kenntnis hatten, den eigentlichen Projktaufwand vorab deutlich unterschätzt. Daher war im Rahmen der Förderung eine vollständige Abbildung auch von Berichtsjahren vor 2015 und auch die Bereitstellung der Analysen in zwei verschiedenen Programmiersprachen nicht realistisch. Durch die dann erfolgte Reduizierung der zu verarbeitenden Berichtsjahre und der Fokusierung auf eine Programmiersprache für die Beispielanalysen konnten wir die Kostenplanung aber letztlich einhalten.

[2] Für weitere Details der einzelnen Pakete siehe oben.

Kurze Darstellung von etwaigen Ergebnissen bei anderen Stellen

Die einzige Veränderung in der Verfügbarkeit der QBK-Daten an anderer Stelle waren die vom GBA selbst bereitgestellten „Referenzberichte“, die erstmals mit den 2019er Daten im Förderzeitraum verfügbar gemacht wurden (https://g-ba-qualitaetsberichte.de/#/search). Da diese jedoch nur einzeln pro Klinik als PDF-Version abgerufen werden können und teilweise 800-1.000 Seiten pro Klinik umfassen, sind auch sie für übergreifende Analysen vollständig ungeeignet. Wir haben die Referenzberichte jedoch (wie auch vom GBA gedacht) als Hilfestellung für die Überprüfung unserer Parsing-Software nutzen können, um zu sehen, dass wir alle relevanten Rohdaten im Ergebnis richtig extrahieren konnten.