DataPrepHealth

Creating a process for data preparation for modeling projects based on examples from the health-care sector

Programm / Ausschreibung	FORPA, Forschungspartnerschaften NATS/Ö-Fonds, FORPA NFTE2018	**Status**	abgeschlossen
Projektstart	01.10.2018	**Projektende**	30.09.2021
Zeitraum	2018 - 2021	**Projektlaufzeit**	36 Monate
Keywords	Data Preparation, Health Care, Modelling, Simulation, Processes		

Projektbeschreibung

Modellbildung und Simulation wird heutzutage in sehr vielen verschiedenen Anwendungsgebieten, wie Ingenieurwesen, Sozialwissenschaften, Medizin, Logistik und vielen anderen,verwendet, wobei in jedem dieser Gebiete unterschiedliche Modellbildungsansätze verwendet werden, die die jeweiligen Anforderungen erfüllen.
Zu Beginn jedes Modellierungsprojektes stehen Daten. Diese kommen oft aus unterschiedlichen Quellen und sind sehr heterogen. Um eine für einen Modellierung verwendbare Datenbasis zu schaffen, müssen die Daten zuerst entsprechend aufbereitet werden. Dafür sind mehrere Schritte notwendig:
• Es müssen die Inhalte der Daten geklärt werden und die benötigten Daten genau definiert werden.
• Mögliche Datenfehler (Tippfehler, Inkonsistenzen) müssen erkannt und bereinigt werden.
• Kommen die Daten aus mehreren verschiedenen Datenquellen, müssen mehrfach vorkommende Daten eliminiert werden. Die restlichen Daten müssen so zusammengefügt werden, dass ein einheitliches Datenmodell entsteht.
• Je nach gewünschter Modellierungsart müssen die Daten unterschiedlich aufbereitet werden. Es muss ein Datensatz so erstellt werden, dass dieser gut für das Modell genutzt werden kann.

Für viele dieser Schritte gibt es bereits Methoden. Das Ziel dieser Arbeit ist es einen ganzheitlichen Prozess zu entwickeln, der alle diese Schritte umfasst und auch Schnittstellen dazwischen definiert. Dazu soll einerseits auf vorhandene Methoden zurückgegriffen werden, andererseits sollen auch neue Methoden entwickelt werden. Der letzte Schritt, die Datenaufbereitung, soll im Hinblick auf komplexe Simulationsmodelle entwickelt werden.
Die Ansätze, die es bisher in diese Richtung gibt, zielen vor allem auf eine Dokumentation des Prozesse ab (zum Beispiel in Form von DMPs). Einen ganzheitlichen, methodischen Ansatz gibt es noch nicht.
Ein besonderes Augenmerk soll dabei auf Dokumentierbarkeit und Reproduzierbarkeit gelegt werden. Vor allem die Reproduzierbarkeit wurden in der Vergangenheit immer wieder vernachlässigt.
Die Funktionsweise des Prozesses soll anhand einiger Beispiele aus dem Gesundheitsbereich getestet werden. Gerade im Gesundheitsbereich müssen Daten aus verschiedensten Quellen verknüpft werden, da hier, zum Beispiel aus Datenschutzgründen, von jeder Quelle nur sehr spezifische Daten gesammelt werden dürfen.
Das Ergebnis dieser Arbeit soll ein ausgefeiltes Toolkit sein, das den ganzen Prozess der Datenverarbeitung und -

aufbereitung inklusive aller Schnittstellen darstellt. Auch eine ausführliche Dokumentation des Prozesses inklusive einer Definition der möglicherweise notwendigen Einschränkungen soll in Zuge der Arbeit entstehen.

Projektpartner

- Verein DEXHELPP zur Forschungsförderung im Gesundheitssystem