Angewandte Forschung für die Welt von morgen
Schriftenreihe der Hochschule Hannover (Band 3)

Prof. Dr. Christian Wartena, Prof. Dr. Ralf Bruns (Hrsg.)

Smart Data Analytics

Schriften des Forschungsclusters Smart Data Analytics

2020

HOCHSCHULE HANNOVER
UNIVERSITY OF APPLIED SCIENCES AND ARTS

Prof. Dr. Christian Wartena, Prof. Dr. Ralf Bruns (Hrsg.)
Smart Data Analytics : Schriften des Forschungsclusters Smart Data Analytics
Hannover: Hochschule Hannover, 2020
(Angewandte Forschung für die Welt von morgen – Schriftenreihe der Hochschule Hannover; Band 3)

ISSN 2699-3384
DOI: https://doi.org/10.25968/opus-1789
URN: https://nbn-resolving.org/urn:nbn:de:bsz:960-opus4-17897

Weitere Schriften aus der Reihe Angewandte Forschung für die Welt von morgen – Schriftenreihe der Hochschule Hannover finden Sie unter: **https://serwiss.bib.hs-hannover.de/solrsearch/index/search/searchtype/series/id/12**
Die Schriftenreihe Angewandte Forschung für die Welt von morgen – Schriftenreihe der Hochschule Hannover enthält Schriften zu ausgewählten Fragestellungen angewandter Forschung an der Hochschule Hannover.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über **https://portal.dnb.de** *abrufbar.*

Impressum

Hochschule Hannover
Präsident/in
Expo Plaza 4
30539 Hannover

Dieses Dokument ist lizenziert unter der Lizenz Creative Commons Namensnennung 4.0 (CC BY 4.0):
https://creativecommons.org/licenses/by/4.0/

Inhaltsverzeichnis

Das Forschungscluster Smart Data Analytics an der Hochschule Hannover
Was wissen meine Daten, was ich nicht weiß?
(Ralf Bruns und Christian Wartena) 5

1. Kurzporträts der laufenden Projekte

Skalierbarkeit mobiler mikro-Blockheizkraftwerke 10

NOA: Nachnutzung von Open-Access Abbildungen 13

USEfUL: Untersuchungs-, Simulations- und Evaluations-Tool für Urbane Logistik 16

HiGHmed: Medizininformatik-Konsortium – Bessere Patientenversorgung und Forschung durch innovativen Informations- und Datenaustausch 19

ZURESO: Zufriedenheit von Rehabilitanden in Erfahrungsberichten der sozialen Medien 22

Analyse von Erneuerungsverträgen 24

PasoMe: Patientenzufriedenheit in den sozialen Medien. Handlungsempfehlungen für niedersächsische Kliniken. 25

ViTAWiN: Virtuell-augmentiertes Training für die Aus- und Weiterbildung in der interprofessionellen Notfallversorgung 27

GLACIER: Angriffserkennung durch multidimensionale Analyse sicherheitsrelevanter Datenströme 31

QualiPRO: Qualitätstransparenz in der Hüftendoprothetik durch Patient Reported Outcomes 34

ZLG: Zukunftslabor Gesundheit 36

JuVer: Extraktion von juristischen Verweisstrukturen 38

HsH-PROFIL: Kooperation HsH-Forschungscluster Smart Data Analytics mit Forschungsinstitut CETINIA in Madrid 40

SenAD: Sensorintegration in Asphalt für ein Daten-basiertes Degradationsmonitoring 41

Inhaltsverzeichnis

Kombinom: Datenmodellierung für den Einsatz von autonomen Kleinbussen im ländlichen Raum zum kombinierten Transport von Personen und Gütern 43

LernMINT: Datengestützter Unterricht in den MINT-Fächern 45

Verantwortungsvolle KI in der Digitalen Gesellschaft 47

Orte des Gestapoterrors im heutigen Niedersachsen 48

NFDI4Culture: Nationale Forschungsdateninfrastruktur zu materiellen und immateriellen Kulturgütern 50

TAPIR: Teilautomatisiertes Persistent-Identifier-basiertes Reporting 53

Weitere Veröffentlichungen des Clusters 54

2. Ausgewählte Fachbeiträge

Der rechtliche Schutz von und durch künstliche Intelligenz geschaffener Kunstwerke
(Fabian Schmieder) 57

Standard and Advanced Automated Data Quality Monitoring
(Felix Heine, Carsten Kleiner, Thomas Oelsner) 64

Usability Testing of Visual Policy Evaluation for Network Security Event Detection
(Volker Ahlers, Bastian Hellmann) 94

Screen Reject Data Warehouse – Klinisches Data Warehouse zur Abstoßungsdiagnostik nach NTx: Ein Werkstattbericht
(Matthias Katzensteiner, Nina Schewe, Maximilian Zubke, Kim Nörtemann, Bianca Frenzel und Oliver J. Bott) 102

Unsere Autoren 126

Das Forschungscluster Smart Data Analytics an der Hochschule Hannover

Was wissen meine Daten, was ich nicht weiß?

(Ralf Bruns und Christian Wartena)

Die Hochschule Hannover (HsH) zeichnet sich durch die hohe wissenschaftliche und künstlerische Qualität ihrer Forschungs- und Entwicklungsaktivitäten sowie ihrer Lehre in Kombination mit einer ausgeprägten Praxisrelevanz und Anwendungsnähe aus. Um das Forschungsprofil zu schärfen, die Forschungsaktivitäten hochschulintern zu vernetzen und zu intensivieren, hat die HsH sechs Forschungscluster eingerichtet, in denen Forschende fakultätsübergreifend zentrale Fragen angewandter Forschung bearbeiten. Eines dieser Forschungscluster heißt Smart Data Analytics, das wir Ihnen im Folgenden vorstellen möchten.

Das Thema Smart Data hat sich aufgrund der wirtschaftlichen und gesellschaftlichen Bedeutung von Daten zu einem der wichtigsten Zukunftsthemen weltweit entwickelt. Der Zugang zu Daten und die Fähigkeit, Daten zielbezogen bereitzustellen und auszuwerten, sind entscheidende Wettbewerbsfaktoren der digitalen Wertschöpfung und bilden einen zentralen Baustein der vielfältigen Digitalisierungsinitiativen, die zurzeit laufen.

Das Forschungscluster Smart Data Analytics

Das Forschungscluster Smart Data Analytics bündelt die umfangreichen methodischen, technologischen, rechtlichen und anwendungsbezogenen Kompetenzen der Hochschule Hannover im Bereich Smart Data. Der Fokus liegt hierbei auf der anwendungsorientierten Forschung, also dem Transfer von Forschungsergebnissen in praxistaugliche und sichere Anwendungssysteme. Ein zentrales Anliegen des Forschungsclusters ist es deshalb, Unternehmen und öffentlichen Institutionen in der Region Hannover und darüber hinaus den Zugang zu diesen, für ihre Wettbewerbsfähigkeit entscheidenden Technologien, zu ermöglichen.

Durch Bündelung ihrer Forschungsaktivitäten zu Smart Data Analytics adressiert die HsH ein Zukunftsthema mit weltweit sehr hohem Innovationspotenzial und stellt seine interdisziplinäre Expertise Kooperationspartnern im Rahmen von anwendungsorientierten Forschungsprojekten zur Verfügung. Die thematischen Schwerpunkte des Forschungsclusters liegen in den Kernbereichen Data Integration, Data Management, Data Analytics & Visualization sowie Data Security & Privacy. Zu den erforschten Anwendungsbereichen gehören insbesondere die Medizin und das Gesundheitswesen. Durch dieses Zusammenführen der signifikanten Technologiebereiche deckt das Forschungscluster die zentralen Fragestellungen dieses übergreifenden Themas ganzheitlich ab und entwickelt integrierte Lösungen. Dabei wird das komplette Spektrum adressiert: von konzeptionellen Grundlagen über technologische Lösungen bis hin zu einer Realisierung in konkreten Anwendungen und einer Betrachtung technischer und gesellschaftlicher Auswirkungen.

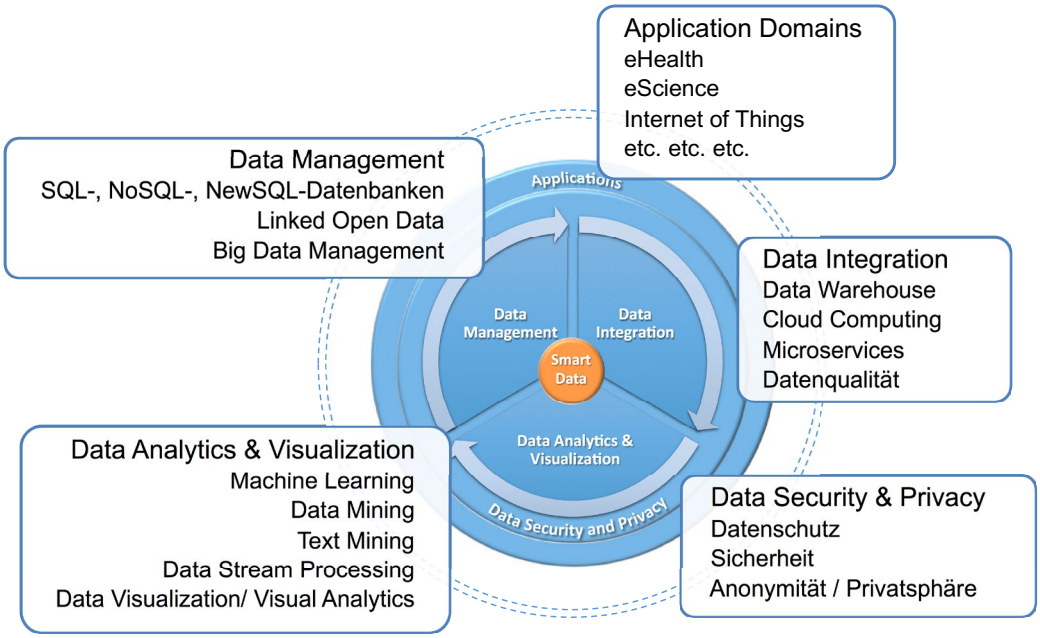


Abb. 1. Forschungsthemen im Cluster Smart Data Analytics

Schwerpunkte und ausgewählte Technologien/Themen, die im Cluster bearbeitet werden, sind:

- **Data Integration: Data Warehouse, Cloud Computing, Microservices,Datenqualität**
- **Data Management: SQL- und NoSQL-Datenbanken, Linked Open Data, Big Data Management**
- **Data Analytics & Visualization: Machine Learning, Data Mining, Text Mining, Data Stream Processing, Data Visualization, Virtual Reality**
- **Data Security & Privacy: Datenschutz, Sicherheit, Anonymität, Privatsphäre**
- **Anwendungsbereiche: Medizin, Gesundheitswesen, Logistik, Internet of Things (IoT), eScience, IT-Sicherheit, Elektromobilität etc.**

Das Team des Clusters besteht aus 13 Professorinnen und Professoren der Fakultät III – Medien, Information und Design, Abteilung Information und Kommunikation (F3) sowie der Fakultät IV – Wirtschaft und Informatik, Abteilungen Informatik und Wirtschaftsinformatik (F4), die ihre jeweilige Expertise in das Forschungscluster einbringen.

- **Prof. Dr. Volker Ahlers (F4): Visualisierung & Simulation**
- **Prof. Dr. Ina Blümel (F3): Informationsmanagement & e-Science**
- **Prof. Dr. Oliver J. Bott (F3): Medizinische Informatik**
- **Prof. Dr. Ralf Bruns (F4 – stellvertretender Sprecher): Softwareentwicklung & Künstliche Intelligenz**
- **Prof. Dr. Jürgen Dunkel (F4): Softwarearchitektur & Intelligente Systeme**
- **Prof. Dr. Felix Heine (F4): Datenbanken & Informationssysteme**
- **Prof. Dr. Carsten Kleiner (F4): Sichere Informationssysteme**
- **Prof. Dr. Arne Koschel (F4): Verteilte Informationssysteme**
- **Prof. Dr. Uwe Sander (F3): Medizinisches Informationsmanagement**
- **Prof. Dr. Jonas Schild (F4): Virtual Reality & Serious Games**
- **Prof. Dr. Fabian Schmieder (F3): IT- & Medienrecht**
- **Prof. Dr. Christoph von Viebahn (F4): Wirtschaftsinformatik**
- **Prof. Dr. Christian Wartena (F3 – Sprecher): Sprach- & Wissensverarbeitung**

Abb. 2. Mitglieder des Forschungsclusters Smart Data Analytics beim ersten Doktorandenkolloquium am 13.6.2017 in Hannover-Linden

An den Projekten des Clusters forschen zahlreiche, junge wissenschaftliche Mitarbeiterinnen und Mitarbeiter, von denen viele auch die Möglichkeit einer kooperativen Promotion wahrnehmen. Weitere Informationen zum Forschungscluster finden Sie unter: **https://forschungscluster.hs-hannover.de/smart-data-analytics/**

Über dieses Buch

Dieses Buch bietet einen Überblick über ausgewählte Cluster-Arbeiten aus den Jahren 2019 und 2020.

In der ersten Hälfte des Buches gibt eine Reihe von Kurzporträts von laufenden oder kürzlich abgeschlossenen Projekten einen Überblick über die Forschung im Cluster und dokumentiert die wissenschaftliche Leistung des Clusters. Enthalten in den Kurzporträts ist eine vollständige kommentiere Liste aller wissenschaftlichen Veröffentlichungen der Cluster-Mitglieder in den letzten beiden Jahren.

In der zweiten Hälfte dieses Bandes finden Sie ausgewählte wissenschaftliche Beiträge über aktuell im Cluster untersuchte Fragestellungen. Diese Beiträge zeigen exemplarisch die thematische Breite des Clusters.

Ein Originalartikel von Fabian Schmieder eröffnet den zweiten Teil dieses Bandes. In diesem Beitrag mit dem Titel **Der rechtliche Schutz von und durch künstliche Intelligenz geschaffener Kunstwerke** findet die Frage Erörterung, welchen rechtlichen Schutz Kunstwerke (und damit auch andere grundsätzlich schutzwürdige Werke) beanspruchen können und wem diese Rechte gebühren.

Drei ausführliche Übersichtsartikel führen uns weiter durch den Kreis der Arbeitsgebiete (siehe Abbildung 1) und geben tiefere Einblicke in Projekte zu **Data Management, Data Analytics & Visualization** und zu einer Anwendung von **Datenintegration in der Medizin.**

Der Beitrag **Standard and Advanced Automated Data Quality Monitoring** stellt ein System zur Überwachung der Datenqualität in Datenbanken vor. Das System basiert auf einer im Projekt iqm4hd entwickelten Sprache, radar, um komplexe Regeln über erwartete Werte in einer Datenbank zu formulieren. Diese Sprache ist unabhängig von der spezifischen Datenbank und kann sowohl für relationale als für verschiedenen no-sql-Datenbanken verwendet werden. Die Regeln werden entweder von einer eigenen rule engine ausgeführt oder in native Datenbankanfragen übersetzt und führen zu Warnhinweisen, wenn Werte oder statistische Verteilungen von Werten von den Erwartungen abweichen. Eine Benutzeroberfläche und automatisch generierte Vorschläge für Qualitätsregeln runden das System ab.

In den Projekten **VisITMeta** und **SIMU** werden ebenfalls Regeln genutzt, um auffällige Ereignisse – hier Angriffe in einem Computernetzwerk – zu detektieren. Auch dabei ist eine Bewertung des Ereignisses durch einen Menschen vorgesehen. In ihrem Beitrag **Usability Testing of Visual Policy Evaluation for Network Security Event Detection** adressieren Volker Ahlers und Bastian Hellmann das Problem, dass ein Systemadministrator für die Bewertung eine Vielzahl von Daten und Regeln anschauen muss. Die Autoren stellen eine Visualisierung der Ereignisse im Netzwerk vor und zeigen in einer Usability-Studie, wie die Effi zienz dadurch gesteigert werden kann.

Im abschließenden Beitrag **Optimierte Diagnostik bei Nierentransplantationen** zeigen Matthias Katzensteiner, Nina Schewe, Maximilian Zubke, Kim Nörtemann, Bianca Frenzel und Oliver J. Bott wie eine umfangreiche Datenbasis in einem klinischen Kontext aufgebaut werden kann. Hierzu mussten nicht nur neue Schnittstellen zu verschiedenen Systemen implementiert werden, sondern es wurde außerdem ein semantisches Datenmodell entwickelt und es wurden Methoden aus der automatischen Sprachverarbeitung angewandt, um strukturierte Daten aus textuellen Quellen zu extrahieren. Auf der geschaffenen Datenbasis sollen Methoden der explorativen Datenanalyse sowie des maschinellen Lernens genutzt werden, um ein Expertensystem für die NTx-Abstoßungsdiagnostik zu implementieren.

Wir wünschen Ihnen viel Freude und wertvolle Erkenntnisse bei der Lektüre dieses Buches!

Ralf Bruns und Christian Wartena

Kurzporträts der laufenden Projekte

1

Skalierbarkeit mobiler mikro-Blockheizkraftwerke

Forschungsschwerpunkt

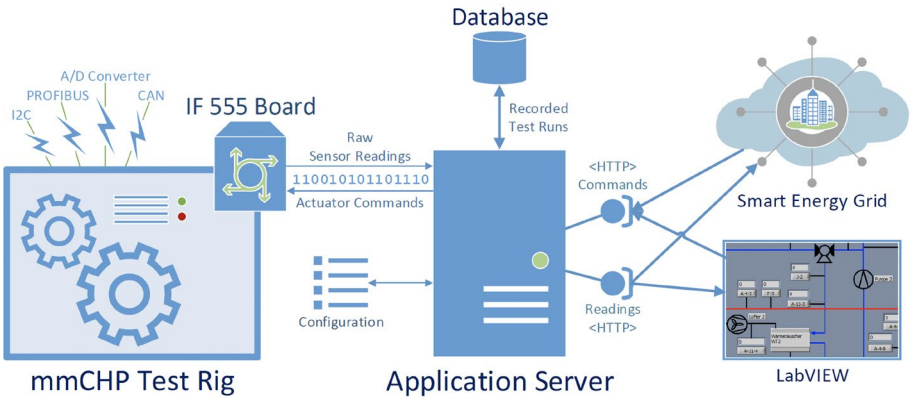
Projekttitel	Skalierbarkeit mobiler mikro-Blockheizkraftwerke
Projektlaufzeit	08.2013 bis 01.2019
Kurzbeschreibung	Untersuchung des Einsatzes portabler mikro-BHKW
Zielsetzung / Ergebnisse	Aufbau von BHKW-Prüfständen und -Prototypen, Messung von Kenngrößen und Wirkungsgraden, Simulation von Einsatzszenarien, Veröffentlichungen
Beteiligte Mitarbeiter*innen	Prof. Dr.-Ing. Lars Gusig (Fakultät II, Projektleitung), Prof. Dr.-Ing. Arne Koschel, Prof. Dr. Volker Ahlers, Richard Pump M.Sc., Dominik Schöner M.Sc., et al.
Beteiligte Institutionen	IAV GmbH, paragon AG, Kraftwerk GmbH, Enercity Contracting GmbH, Ostfalia Hochschule, University of South Wales, UK
Geldgeber / Förderlinie	MWK – Ministerium für Wissenschaft und Kultur, VolkswagenStiftung / Niedersächsisches Vorab, FSP-Pro (FKZ VWZN2891)
Finanzvolumen	983.595 € (Anteil HsH)
Projektwebseite	`http://www.ikme.wp.hs-hannover.de/projekte/skalierbarkeit-mobiler-mikro-bhkw/`

Die ressourcenschonende Bereitstellung von Mobilität und Energie ist für Menschen in unterschiedlichsten Lebenssituationen relevant. Seit Jahren wird intensiv an Lösungen zur Reichweitenproblematik von Elektrofahrzeugen gearbeitet. Jedoch gibt es auch in vielen anderen Bereichen Bedarf zur gekoppelten Bereitstellung von thermischen und elektrischen Energieströmen, beispielsweise beim Transport von Medikamenten und Lebensmitteln, für den dezentralen Betrieb von Arbeitsmaschinen und Kommunikationsanlagen oder im Outdoor- und Freizeitbereich. Heute verfügbare Lösungen sind oft zu groß und zu ineffizient.

Im Rahmen des Forschungsschwerpunktes wurden technische Möglichkeiten zur Skalierung von mobilen bzw. portablen mikro-Blockheizkraftwerken untersucht. Blockheizkraftwerke (BHKW, engl. *combined heat and power units*, CHP) erzeugen über das Prinzip der Kraft-Wärme-Kopplung thermische und elektrische Energie mit einem hohen Wirkungsgrad. Ausgehend von einem Power Conditioning Unit (PCU), einem Konzept der IAV GmbH, wurden auf Basis von Modellierungen Funktionsmuster in verschiedenen Leistungsbereichen aufgebaut und im Labor überprüft. Die Ergebnisse wurden auf eine neue Variante zunächst theoretisch, dann auch konstruktiv übertragen und die Praxistauglichkeit exemplarisch für die Automobilindustrie an einem Prototypen überprüft. Neben der Ermittlung der Einsatzgrenzen

unterschiedlicher Antriebstechnologien sollten insbesondere eine geeignete Abstimmung der Systemkomponenten gefunden werden, bei der in einem neuen Leistungsbereich eine optimale Energieeffizienz erreicht wird.

Die Arbeitspakete der beteiligten Mitglieder des Forschungsclusters Smart Data Analytics umfassten die Entwicklung eines Frontends zur Ansteuerung und Überwachung des BHKW-Prüfstandes sowie die Untersuchung von Konzepten zur Vernetzung portabler mikro-BHKW im Smart Energy Grid. Das auf LabVIEW basierende Frontend nutzt Standardschnittstellen wie CAN, Profibus und I^2C zur Ansteuerung des BHKW-Prüfstandes und erlaubt die Live-Visualisierung und Speicherung von Messdaten [6,7]. Die Vernetzung portabler mikro-BHKW im Smart Energy Grid wurde mit Hilfe eines Netzwerksimulators untersucht [1,4,5]. Ein besonderes Augenmerk wird dabei auf Sicherheitsaspekte gelegt, indem in laufenden Forschungsarbeiten der Einsatz künstlicher Immunsysteme für die Angriffserkennung untersucht wird [2,3].



Veröffentlichungen

1. Pump, R., Ahlers, V., Koschel, A.: Simulation des Kommunikationsnetzes im Smart Grid. In: Wittmann, J., Wolf, L., Ahlers, V., Koschel, A., Pump, R. (eds.) Simulation in Umwelt- und Geowissenschaften. Workshop Hannover 2018. ASIM Mitteilung, vol. AM 167, pp. 81–91. Shaker, Aachen (2018)
2. Pump, R., Ahlers, V., Koschel, A.: State of the art in artificial immune-based intrusion detection systems for smart grids. In: Yang, X.S., Dey, N., Joshi, A. (eds.) 2018 Second World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). pp. 119–126. IEEE (2018). https://doi.org/10.1109/WorldS4.2018.8611584
3. Pump, R., Ahlers, V., Koschel, A.: Evaluating artificial immune system algorithms for intrusion detection. In: Yang, X.S., Fong, S.J., Toapanta, M., Andronache, I., Phillips, N. (eds.) 2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4). pp. 92–97. IEEE (2020). https://doi.org/10.1109/WorldS450073.2020.9210342
4. Pump, R., Koschel, A., Ahlers, V.: On microservices in smart grid capable pmCHP. In: Koschel, A., Klink, J., Hausotter, A. (eds.) SERVICE COMPUTATION 2018: The Tenth International Conference on Advanced Service Computing. pp.

30–35. IARIA (2018), http://thinkmind.org/index.php?view=article&articleid=service_computation_2018_2_20_18002
5. Pump, R., Rüscher, H., Schöner, D., Saul, D., Ahlers, V., Koschel, A.: Simulation des Netzverkehrs der Informationstechnologie im Smart Grid. In: Wittmann, J. (ed.) Simulation in Umwelt- und Geowissenschaften. Workshop Berlin 2017. ASIM Mitteilung, vol. AM 163, pp. 85–96. Shaker, Aachen (2017)
6. Schöner, D., Pump, R., Rüscher, H., Koschel, A., Ahlers, V.: Providing a sustainable, adaptive IT infrastructure for portable micro-CHP test benches. In: Otjacques, B., Hitzelberger, P., Naumann, S., Wohlgemuth, V. (eds.) From Science to Society (Proceedings of EnviroInfo 2017). pp. 297–307. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65687-8_26
7. Schöner, D., Pump, R., Schmicke, C., Minnrich, J.P., Rüscher, H., Ahlers, V., Koschel, A.: IT-Unterstützung von BHKW-Prüfständen in der Angewandten Forschung. In: Mayr, H.C., Pinzger, M. (eds.) INFORMATIK 2016: Workshop Umweltinformatik zwischen Nachhaltigkeit und Wandel (UINW 2016). LNI, vol. P-259, pp. 1227–1238. GI, Bonn (2016), https://dl.gi.de/20.500.12116/1010

NOA
Nachnutzung von Open-Access-Abbildungen

Projekttitel	Entwicklung eines Verfahrens zur automatischen Sammlung, Erschließung und Bereitstellung von Open-Access-Abbildungen aus dem technischen Fächerspektrum mittels der Infrastruktur von Wikimedia Commons und Wikidata
Projektlaufzeit	1.12.2016 bis 1.4.2020
Kurzbeschreibung	Entwicklung einer Pipeline zum Sammeln, Archivieren und Erschließen von Bildern aus Open-Access Publikationen
Zielsetzung / Ergebnisse	Sammeln und Erschließen von frei-lizenzierten Bildern aus wissenschaftlichen Veröffentlichungen um die Wiederverwendung zu fördern
Beteiligte Mitarbeiter*innen	Prof. Dr. Christian Wartena, Prof. Dr. Ina Blümel, Jean Charbonnier M.A.
Beteiligte Institutionen	Hochschule Hannover, Leibniz-Informationszentrum Technik und Naturwissenschaften - TIB
Geldgeber / Förderlinie	DFG (Förderprogramm „Infrastruktur für elektronische Publikationen und digitale Wissenschaftskommunikation")
Finanzvolumen (HsH)	177.400 €
Projektwebseite	`http://noa.wp.hs-hannover.de/`

Forschungsergebnisse werden hauptsächlich als Textpublikationen veröffentlicht, wobei Bilder zur Visualisierung und zum Verständnis der Ergebnisse beitragen. Open Access vervielfacht das Potenzial der Nachnutzung dieser Abbildungen, zum Beispiel in der Forschungsliteratur, in Materialien zur Lehre und Wissensvermittlung, wie digitale Präsentationsfolien oder Unterrichtsblättern, für die eine Lizensierung von Bildmaterial oft eine Barriere darstellt. Während die Textpublikationen selbst durch standardisierte Indizes über Suchmaschinen und in bibliothekarischen Rechercheportalen gefunden und nachgenutzt werden können, fehlt ein vergleichbarer gezielter Zugriff auf die darin enthaltenen oder sie begleitenden Abbildungen, um sie besser zu verbreiten und nachnutzbar zu machen. Die Abbildungen, im NOA-Projekt überwiegend Grafiken, Abbildungen aus bildgebenden Verfahren und Fotos, sollen nun für weitere Forschungsarbeiten zur Verfügung gestellt werden.

Im ersten Schritt wurden Artikel führender Open-Access Verlage heruntergeladen, die für das Projekt passende Lizenzen und Formate hatten. Diese wurden aus verschiedenen Quellen heruntergeladen. Aus den Artikeln wurden Metadaten und Bilder extrahiert [6].

Die Metadaten der Artikel und Bilder wurden mit verschiedenen Mitteln angereichert. Einige Daten (Zitationsdaten, Fachgebiete) konnten von anderen Diensten abgefragt werden. Andere Daten wurden aus den Bildunterschriften

und den Textabschnitten, die auf die Bilder verweisen, extrahiert. Für jedes Bild werden Begriffe und Phrasen in der Bildunterschrift und in auf die Bilder verweisenden Sätzen erkannt und es wird überprüft, ob diese in Wikipedia als Artikelname vorkommen. Akronyme, die in den Bildunterschriften sehr häufig verwendet werden, werden detektiert, desambiguiert und aufgelöst [2]. Nicht alle Begriffe, die vorkommen sind geeignet um das referenzierte Bild zu beschreiben. Konkretheit und Bildhaftigkeit von Wörtern hat sich als wichtiges Merkmal herausgestellt um beschreibende Terme von anderen zu unterscheiden [3,4]. Schließlich werden die Begriffe nach verschiedenen Kriterien (Zentralität, Konkretheit, IDF) gerankt und die besten Begriffe benutzt, um über die korrespondierenden Wikipedia-Artikel die passende Wikipedia- und Wikimedia-Kategorien für das Bild zu bestimmen [5].

Die Suchmaschine wurde entwickelt, um die Projektergebnisse schnell sichtbar und durchsuchbar zu machen [1]. Sie unterstützt das Filtern nach Fachgebiet und Bildtyp, sowie das Suchen nach Wikipediakategorien. Sie verfügt außerdem über eine API, so dass automatisch Informationen abgefragt werden können. Die Suchmaschine ist öffentlich zugänglich auf der Seite `http://noa.wp.hs-hannover.de/`

Zusätzlich wurden zwei Funktionen in die Suchmaschine eingebaut, die ermöglichen, dass Benutzer über ihren Wikimedia Account Bilder hochladen. Metadaten zum Bild und zum Artikel, in dem das Bild veröffentlicht wurde, werden automatisch hochgeladen. Inhaltsbeschreibende Kategorien von Wikimedia Common werden vom Programm vorgeschlagen und können vom Nutzer übernommen werden. Die wissenschaftlichen Abbildungen werden somit leichter recherchierbar und können weltweit von Wikipedia-Nutzer*innen und der jeweiligen Wissenschaftscommunity verwendet werden.

Veröffentlichungen

1. Charbonnier, J., Sohmen, L., Rothman, J., Rohden, B., Wartena, C.: NOA: A Search Engine for Reusable Scientific Images Beyond the Life Sciences. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) Advances in Information Retrieval. Lecture Notes in Computer Science, vol. 10772, pp. 797–800. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_78, `http://link.springer.com/10.1007/978-3-319-76941-7_78`
2. Charbonnier, J., Wartena, C.: Using word embeddings for unsupervised acronym disambiguation. In: Proceedings of the 27th International Conference on Computational Linguistics. pp. 2610–2619. Association for Computational Linguistics, Santa Fe, New Mexico, USA (Aug 2018), `https://www.aclweb.org/anthology/C18-1221`
3. Charbonnier, J., Wartena, C.: Predicting Word Concreteness and Imagery. In: Proceedings of the 13th International Conference on Computational Semantics - Long Papers. pp. 176–187. Association for Computational Linguistics, Gothenburg, Sweden (2019), `https://www.aclweb.org/anthology/W19-0415`
4. Charbonnier, J., Wartena, C.: Predicting the Concreteness of German Words. In: Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS). vol. 2624. CEUR-WS, Zürich (2020), `http://ceur-ws.org/Vol-2624/paper8.pdf`

5. Josi, F., Wartena, C., Charbonnier, J.: Text-Based Annotation of Scientific Images Using Wikimedia Categories. In: Elloumi, M., Granitzer, M., Hameurlain, A., Seifert, C., Stein, B., Tjoa, A.M., Wagner, R. (eds.) Database and Expert Systems Applications, vol. 903, pp. 243–253. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-99133-7_20, http://link.springer.com/10.1007/978-3-319-99133-7_20
6. Sohmen, L., Charbonnier, J., Blümel, I., Wartena, C., Heller, L.: Figures in Scientific Open Access Publications. In: Méndez, E., Crestani, F., Ribeiro, C., David, G., Lopes, J.C. (eds.) Digital Libraries for Open Knowledge, vol. 11057, pp. 220–226. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-030-00066-0_19, http://link.springer.com/10.1007/978-3-030-00066-0_19

USEfUL

Untersuchungs-, Simulations- und Evaluations-Tool für Urbane Logistik

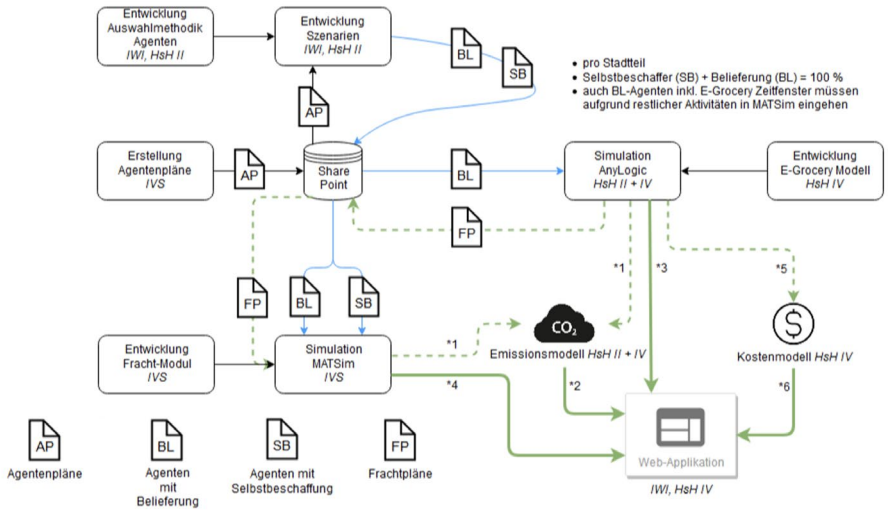
Projekttitel	Untersuchungs-, Simulations- und Evaluations-Tool für Urbane Logistik – USEfUL
Projektlaufzeit	09.2017 bis 12.2020
Kurzbeschreibung	Untersuchung neuer Konzepte für die urbane Logistik
Zielsetzung / Ergebnisse	Modellierung und Simulation von Logistikkonzepten, Analyse von Kenngrößen, Entwicklung eines Web-Tools zur Entscheidungsunterstützung, Veröffentlichungen
Beteiligte Mitarbeiter der HsH	Prof. Dr. Christoph von Viebahn (Projektleitung HsH), Prof. Dr.-Ing. Arne Koschel, Prof. Dr. Volker Ahlers, Prof. Dr.-Ing. Lars Gusig (Fakultät II), Marvin auf der Landwehr M.Sc., Hai Mi Ngo B.Sc., Richard Pump M.Sc., Maik Trott M.Sc., et al.
Beteiligte Institutionen	Landeshauptstadt Hannover (Gesamtprojektleitung), Leibniz Universität Hannover, Technische Universität Braunschweig, Volkswagen AG Nutzfahrzeuge
Geldgeber / Förderlinie	BMBF – Bundesministerium für Bildung und Forschung / Zukunftsstadt (FKZ 03SF0547D)
Finanzvolumen	643.723 € (Anteil HsH)
Projektwebseite	`https://www.hannover.de/Urbane-Logistik-Hannover/`

Bedingt durch eine steigende Anzahl von Paketsendungen, Digitalisierung, sich ändernde Mobilitätsformen und die damit verbundene Flächenkonkurrenz im urbanen Raum, nimmt die öffentliche Wahrnehmung bezüglich der urbanen Logistik und der Ausgestaltung der letzten Meile zu. Im Forschungsprojekt USEfUL wird eine Web-Applikation entwickelt, mit der Logistikkonzepte bewertet sowie deren Auswirkungen auf das Leben in der Stadt und den Gesamtverkehr aufgezeigt werden können. Insbesondere Politiker*innen und Kommunen sollen das zu entwickelnde Tool verwenden, um entsprechend ihrer Nachhaltigkeitsziele effektive Maßnahmen zu identifizieren.

Auf Grundlage der Modellierung sowie Simulation von Verkehrs- und Warenströmen sollen innovative Logistikkonzepte analysiert und bewertet werden. Mit der Simulation werden die jeweiligen Auswirkungen auf Verkehr und Umwelt analysiert und evaluiert – auch im Hinblick auf ihre Wirtschaftlichkeit und Kundenakzeptanz. Die erforschten Erkenntnisse werden in der Web-Applikation dargestellt. Somit können Entscheidungsträger*innen sich über innovative urbane Logistikkonzepte informieren und einen Einblick über die zu erwartenden Auswirkungen gewinnen, bevor Konzepte in der Realität umgesetzt werden.

Die Arbeitspakete der beteiligten Mitglieder des Forschungsclusters Smart Data Analytics umfassen die Entwicklung von Modellen für Logistikkonzepte,

die Simulation sowie die Entwicklung des Web-Tools zur Entscheidungsunterstützung. Für das agentenbasierte Simulationstool AnyLogic werden Modelle für verschiedene Logistikkonzepte entwickelt, wie zum Beispiel E-Grocery und Intralogistik [1,2,3,8,9]. Für die Simulation werden verschiedene Ansätze zur Co-Simulation zwischen AnyLogic (mikroskopische Modellierung) und dem von einem Projektpartner verwendeten Simulationstool MATSim (makroskopische Modellierung) untersucht [4,5,6,7].



Literatur

1. Auf der Landwehr, M., Trott, M., von Viebahn, C.: E-grocery in terms of sustainability – simulating the environmental impact of grocery shopping for an urban area in hanover. In: Putz, M., Schlegel, A. (eds.) Simulation in Produktion und Logistik 2019. pp. 87–96. Verlag Wissenschaftliche Scripten, Auerbach (2019)
2. Auf der Landwehr, M., Trott, M., von Viebahn, C.: Computer simulation as evaluation tool of information systems: Identifying quality factors of simulation modeling. In: Aier, S., Gordijn, J., Proper, H.A., Verelst, J. (eds.) 22nd IEEE Conference on Business Informatics (CBI 2020), Volume 1. pp. 211–220. IEEE, Los Alamitos (2020). https://doi.org/10.1109/CBI49978.2020.00030
3. Jordan, Y., von Viebahn, C., Haller, S.: Konzeptionierung von Use Cases für die urbane Logistik in einer Schweizer Smart City mittels Design Thinking. In: Tokarski, K.O., Schellinger, J., Berchtold, P. (eds.) Nachhaltige Unternehmensführung, pp. 391–420. Springer Fachmedien, Wiesbaden (2019)
4. Meinhardt, M., Meyran, T., Schoolmann, K., Pump, R., Ahlers, V., Koschel, A.: Co-Simulation zwischen AnyLogic und MATSim. In: Wittmann, J. (ed.) Simulation in Umwelt- und Geowissenschaften. Workshop 2020. ASIM Mitteilung, vol. AM 173, pp. 117–130. Shaker, Düren (2020)
5. Pump, R., Baumann, M., Bellok, J.T., Ahlers, V., Koschel, A.: Kombinierte Simulation logistikrelevanter Verkehrszusammenhänge. In: Wittmann, J. (ed.) Simulation in

Umwelt- und Geowissenschaften. Workshop Kassel 2019. ASIM Mitteilung, vol. AM 171, pp. 55–66. Shaker, Düren (2019)
6. Pump, R., Koschel, A., Ahlers, V.: Applying microservice principles to simulation tools. In: Hausotter, A. (ed.) SERVICE COMPUTATION 2019: The Eleventh International Conference on Advanced Service Computing. pp. 6–9. IARIA (2019), https://www.thinkmind.org/index.php?view=article&articleid=service_computation_2019_1_20_18002
7. Pump, R., van Perbandt, C., Ahlers, V., Koschel, A.: Creating cloud simulations for urban logistics. In: Deatcu, C., Lückerath, D., Ullrich, O., Durak, U. (eds.) Proceedings ASIM SST 2020. 25. ASIM Symposium Simulationstechnik. ARGESIM Report/ASIM Mitteilung, vol. 59/AM 174, pp. 427–432. ARGESIM, Vienna (2020). https://doi.org/10.11128/arep.59.a59059
8. Trott, M., Auf der Landwehr, M., von Viebahn, C.: A new simulation approach for scheduling consolidation activities in intralogistics – optimising material flow processes in industrial practice. In: Putz, M., Schlegel, A. (eds.) Simulation in Produktion und Logistik 2019. pp. 325–334. Verlag Wissenschaftliche Scripten, Auerbach (2019)
9. von Viebahn, C., Auf der Landwehr, M., Trott, M.: The future of grocery shopping? A taxonomy-based approach to classify e-grocery fulfillment concepts. In: Gronau, N., Heine, M., Krasnova, H., Poustcchi, K. (eds.) Entwicklungen, Chancen und Herausforderungen der Digitalisierung: Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik (WI 2020). pp. 955–971. GITO, Berlin (2020). https://doi.org/10.30844/wi_2020_j2-viebahn

HiGHmed-Medizininformatik-Konsortium

Bessere Patientenversorgung und Forschung durch innovativen Informations- und Datenaustausch

Projekttitel	HiGHmed – Medizininformatik – Bessere Patientenversorgung und Forschung durch innovativen Informations- und Datenaustausch Beitrag der HsH im AP05: Lehre und Ausbildung / Stärkung der Medizininformatik
Projektlaufzeit	1.1.2018 bis 31.12.2022
Kurzbeschreibung	Im Arbeitspaket Lehre und Training des HiGHmed-Projektes werden im Verbund verschiedener akademischer Partner digital-gestützte Online-Lernmodule zu unterschiedlichen Themenschwerpunkten der Medizinischen Informatik entwickelt.
Zielsetzung / Ergebnisse	Entwickelt und umgesetzt werden sollen verschiedene Maßnahmen zur Stärkung der Medizininformatik (MI): Weiterentwicklung der existierenden Curricula, Stärkung der MI in der medizinischen Ausbildung, Stärkung von Frauen in der MI, Entwicklung von digitalen Lernformaten und einer zentralen HiGHmed-Lehrnplattform, Entwicklung neuer Ausbildungsformate für Patientenpartizipation in Forschung und Versorgung, Kooperation mit Industriepartnern, Etablierung von HiGHmed-Zertifikaten.
Beteiligte Mitarbeiter*innen	Marie-Louise Witte, Prof. Dr. Oliver J. Bott, Prof. Dr. Peter Wübbelt
Beteiligte Institutionen	Akademische Partner: Charité Universitätsmedizin Berlin; Helmholtz Zentrum für Infektionsforschung; Hochschule Heilbronn; Hochschule Hildesheim/Holzminden/Göttingen; Medizinische Hochschule Hannover; Technische Universität Braunschweig; Uniklinikum Würzburg; Universität Münster; Universität zu Köln; Universitätsklinikum Heidelberg; Universitätsmedizin Göttingen Industriepartner: Ada Health GmbH
Geldgeber / Förderlinie	BMBF – Bundesministerium für Bildung und Forschung (FKZ: 01ZZ1802P) im Rahmen der Medizininformatik-Initiative
Finanzvolumen	280.000,80 €
Projektwebseite	`https://education.highmed.org/de`

Im Rahmen der Medizininformatik-Initiative (MII) fördert das Bundesministerium für Bildung und Forschung vier Konsortien. Neben der Entwicklung von Strukturen und Technologien zum Austausch und zur Analyse

von Daten aus klinischer Forschung und Gesundheitsversorgung, ist ein weiterer Schwerpunkt der MII die Verbesserung der Aus- und Weiterbildung u.a. auch, um dem Fachkräftemangel in der Medizinischen Informatik (MI) entgegenzuwirken. In HiGHmed, einem der MII-Konsortien, werden innerhalb des HiGHmeducation-Programms derzeit an 12 Mitgliedsstandorten Online-Lernmodule zu unterschiedlichen Themenschwerpunkten der MI angeboten. Die Hochschule Hannover (HsH) bietet das Modul „Fortgeschrittene Technologien und Methoden der Datenanalyse und des Datenqualitätsmanagements" an.

Hauptziele bei der Gestaltung bilden die Kombination aller Module in dem sogenannten HiGHmeducation-Zertifizierungsprogramm sowie deren standortübergreifende Belegung. Zur Berücksichtigung der besonderen Herausforderungen der Online-Lehre und als Leitfaden für die Gestaltung der Kurse wurde ein phasenbasiertes didaktisches Konzept entwickelt, das insbesondere zu einer einheitlichen Struktur der Module führen soll [1,3].

Das didaktische Konzept kann in verschiedenen Open-Source Lernmanagementsystemen (LMS), wie ILIAS, Moodle oder Stud.IP implementiert werden. Ein zentrales Element des didaktischen Konzepts für ein aktivierendes Online-Lernen sind sogenannte E-tivities. E-tivities sind auf Zusammenarbeit und Interaktion ausgerichtete und nach einem definierten Schema strukturierte Aufgaben, die während des Kurses von den Lernenden gelöst werden. Das Modul der HsH ist in zwei Teilmodule aufgeteilt, wobei zunächst nur das erste rein online-gestützt durchgeführt wird. Das erste Teilmodul „Data Warehouses (DTW) und Business Intelligence (BI)" führt in die Anwendung und den Aufbau von Data Warehouses und die damit verknüpften Analysemethoden des Online Analytical Processing (OLAP) ein. Das zweite Teilmodul „Einführung in das Maschinelle/Statistische Lernen" adressiert das hochaktuelle Feld explorativer Methoden der Datenanalyse und gibt auch eine Einführung in Künstliche Neuronale Netzwerke.

Zu Beginn des Projekts lag der Fokus zunächst auf der Umsetzung des Kurses anhand des didaktischen Konzepts und der Produktion hochwertiger Lernvideos u.a. als Bestandteile der E-tivities. Die erste Durchführung des Moduls im Wintersemester 20/21 erfolgte zunächst ausschließlich mit Studierenden der HsH und zeigte Verbesserungsbedarf auf in Bezug auf die für rein digital durchgeführte Kurse wichtige Phase der Online-Sozialisierung. Dass in der ersten Durchführungsphase den Teilnehmern erlaubt war, die Ergebnisse der E-tivities bis spätestens zum Ende des Semesters abzugeben, führte dazu, dass die für eine effektive Zusammenarbeit und gegenseitige Reflektion der Arbeitsergebnisse erforderliche zeitnahe Bearbeitung der E-tivities nicht im ausreichendem Umfang gelang. In einer zweiten Durchführungsphase im Sommersemester 2020 wurde das Konzept daher überarbeitet und weitere Anreize zur gemeinsamen Lösung der E-tivities geschaffen. Zudem wurde das Modul auch für HsH-externe Teilnehmer geöffnet. Darüber hinaus haben die Lehrenden den Umfang der Feedbackrunden intensiviert und Diskussionsrunden eröffnet, um die Motivation aufrecht zu erhalten. Eine

Evaluation des zweiten Durchlaufs steht zum Zeitpunkt dieser Veröffentlichung noch aus. Im weiteren Projektverlauf wird das Online-Angebot weiter verbessert und ausgebaut.

Ein weiteres Ziel von HiGHmed ist die Stärkung von Frauen an der Schnittstelle zwischen Medizin und Informatik. Ein Konzept für eine Sommerschule für Frauen mit Workshops und Podcasts wurde von der HiGHmeducation-Arbeitsgruppe „Frauen in Gesundheit und Informatik“ (WinHaCS), in der die HsH tatkräftige Unterstützung leistet, entwickelt. Hierdurch sollen mittelfristig mehr Frauen für Forschung und Wissenschaft in diesem Bereich gewonnen werden. Der Podcast "DigitalisiergDerMedizin.de"richtet sich an ein breites Publikum, aber - anders als die meisten Podcasts im Kontext der Gesundheitsinformatik - sind nur Frauen Interviewpartner. Die Folgen sollen das Interesse an Themen zwischen Medizin und Informatik wecken sowie spannende Berufsfelder und Karrierewege aufzeigen. Ein Beitrag zur Jahrestagung der GMDS im September 2020, stellt die Konzepte der bisherigen Aktivitäten vor und präsentiert Ergebnisse der Evaluation der Veranstaltungen der Summer School (zugängliche Veröffentlichung steht noch aus) [2].

Veröffentlichungen

1. Behrends, M., Benning, N.H., Witte, M.L., Hoffmann, I., Bott, O.J.: Didactical framework for cross-location online learning modules on medical informatics. https://doi.org/10.3205/19gmds120
2. Hoffmann, I., Warnecke, J., Witte, M.L., Behrends, M., Consortium, H.: Highmed summer school 2019 for women – empowering women in health and computer science.
3. Witte, M.L., Behrends, M., Benning, N.H., Hoffmann, I., Bott, O.J., et al.: The highmed didactical framework for online learning modules on medical informatics: First experiences. Studies in Health Technology and Informatics **272**, 163–166 (2020). https://doi.org/10.3233/SHTI200519

ZURESO

Zufriedenheit von Rehabilitanden in Erfahrungsberichten der sozialen Medien

Projekttitel	ZURESO - Zufriedenheit von Rehabilitanden in Erfahrungsberichten der sozialen Medien
Projektlaufzeit	01.02.2018 bis 31.10.2019
Kurzbeschreibung	Rehabilitanden nutzen soziale Medien, um ihren medizinischen Rehabilitationsaufenthalt zu bewerten und ihre Erfahrungen anderen Nutzern mitzuteilen. Es ist bisher nicht bekannt, wie die Bewertungen ausfallen und anhand welcher Kriterien die Bewertungen erfolgen.
Zielsetzung / Ergebnisse	Ziele des Projektes sind a) eine erstmalige systematische Analyse der Inhalte der Onlinebewertungen am Beispiel des führenden Bewertungsportals Klinikbewertungen.de und des sozialen Netzwerkes Facebook, b) ein Vergleich der Ergebnisse und der Themen der Erfahrungsberichte mit der Rehabilitandenbefragung der Rentenversicherung und c) die Entwicklung von Handlungsempfehlungen für Rentenversicherungsträger und Rehakliniken für den Umgang mit online abgegebenen Patientenkommentaren unter Einbindung von Experten
Beteiligte Mitarbeiter*innen	Prof. Dr. Uwe Sander, Dr. Christiane Patzelt, Susann Kiss, Felix Hopf
Beteiligte Institutionen	Friedrich-Alexander-Universität Erlangen-Nürnberg, Fachbereich Wirtschaftswissenschaften, Lehrstuhl für Gesundheitsmanagement (PD Dr. Martin Emmert)
Geldgeber / Förderlinie	Deutsche Rentenversicherung Bund
Finanzvolumen	108.789 € (mit Eigenanteil)
Projektwebseite	`http://www.public-reporting.wp.hs-hannover.de/projekte/`

Ergänzenden Information sind bei dem nachfolgenden und methodisch zusammenhängendem Projekt PasoMe vermerkt.

Veröffentlichungen

1. Hopf, F., Patzelt, C., Kiss, S., Sander, U.: Teilen, liken, chatten. rehabilitandenzufriedenheit auf facebook. In: 29. Rehabilitationswissenschaftliches Kolloquium, 2.-4. März 2020. DRV-Schriften, vol. 120, pp. 474–476. Deutsche Rentenversicherung Bund, Hannover (2020)
2. Hopf, F., Patzelt, C., Emmert, M., Sander, U.: Zufriedenheit von rehabilitanden in der onkologie. erfahrungsberichte in den sozialen medien. In: 28. Rehabilitationswissenschaftliches Kolloquium, 15th Congress of EFRR, 15.-17. April 2019. DRV-Schriften, vol. 117, pp. 350–352. Deutsche Rentenversicherung Bund, Berlin (2019)

3. Kiss, S., Hopf, F., Patzelt, C., Emmert, M., Sander, U.: Rehabilitanden- und angehörigenzufriedenheit auf klinikbewertungen.de. eine quantitative analyse von onlinebewertungen zur medizinischen rehabilitation. In: 29. Rehabilitationswissenschaftliches Kolloquium, 2.-4. März 2020. DRV-Schriften, vol. 120, pp. 207–208. Deutsche Rentenversicherung Bund, Hannover (2020)
4. Kiss, S., Hopf, F., Patzelt, C., Emmert, M., Sander, U.: Was berichten rehabilitanden auf klinikbewertungen.de über ihre neurologische rehabilitation? In: 19. Deutscher Kongress für Versorgungsforschung (DKVF). German Medical Science GMS Publishing House, Düsseldorf (2020)
5. Kiss, S., Hopf, F., Patzelt, C., Emmert, M., Sander, U.: Zufriedenheit mit der neurologischen rehabilitation. was berichten rehabilitanden in den sozialen medien? In: 29. Rehabilitationswissenschaftliches Kolloquium, 2.-4. März 2020. DRV-Schriften, vol. 120, pp. 173–175. Deutsche Rentenversicherung Bund, Hannover (2020)
6. Patzelt, C., Kiss, S., Hopf, F., Biedermann, I., Schindler, A., Emmert, M., Sander, U.: Zufrieden mit der reha? rehabilitanden bewerten online. (2020)
7. Patzelt, C.: Rehabilitandenzufriedenheit in den sozialen medien. RehaStimme. Das Verbandsmagazin der DEGEMED p. 13 (2019)
8. Sander, U., Biedermann, I., Hofmann, A.L., Hopf, F., Kiss, S., Patzelt, C., Schindler, A.: Regressionsanalyse zur zufriedenheit onkologischer rehabilitanden in den sozialen medien. In: 29. Rehabilitationswissenschaftliches Kolloquium, 2.-4. März 2020. DRV-Schriften, vol. 120, pp. 181–183. Deutsche Rentenversicherung Bund, Hannover (2020)
9. Sander, U., Emmert, M., Hofmann, A.L., Hopf, F., Kiss, S., Schindler, A., Patzelt, C.: Klinik-weiterempfehlung bei rehaerfolg: Erfahrungsberichte von rehabilitanden auf einer krankenhaus-bewertungsplattform. In: 19. Deutscher Kongress für Versorgungsforschung (DKVF). German Medical Science GMS Publishing House, Düsseldorf (2020)

Analyse von Erneuerungsverträgen

Projekttitel	Analyse von Erneuerungsverträgen
Projektlaufzeit	01.05.2018 bis 30.09.2022
Kurzbeschreibung	Entwicklung von Methoden zur Unterstützung im Underwriting-Prozess
Zielsetzung / Ergebnisse	Implementieren von Methoden aus der Computerlinguistik und dem Maschinellem Lernen für die Analyse der Vertragsdokumente.
Beteiligte Mitarbeiter*innen	Prof. Dr. Christian Wartena, Prof. Dr. Fabian Schmieder, Frieda Josi M.A.
Beteiligte Institutionen	Hannover Rück SE (Dr. Julia Perl)
Geldgeber / Förderlinie	Hannover Rück SE
Finanzvolumen	164.972 €
Projektwebseite	`textmining.wp.hs-hannover.de/projects.html`

In dem Kooperationsprojekt zwischen der Hannover Rück SE und der Hochschule Hannover wird untersucht, wie unternehmensinterne Underwriter beim Prüfen und Bewerten der Risiken in Vertragstexten und insbesondere beim Vergleich von aktualisierten Vertragstexten mit den früheren Versionen mit Methoden aus dem Bereich der automatischen Sprachverarbeitung unterstützt werden können.

Das Projekt ist eng an dem Projekt JuVer angebunden. Für Ergebnisse und Veröffentlichungen verweisen wir auf die Beschreibung von JuVer.

PasoMe

Patientenzufriedenheit in den sozialen Medien. Handlungsempfehlungen für niedersächsische Kliniken.

Projekttitel	PasoMe - Patientenzufriedenheit in den sozialen Medien. Handlungsempfehlungen für niedersächsische Kliniken.
Projektlaufzeit	01.09.2018 bis 31.08.2021
Kurzbeschreibung	Soziale Medien und Vergleichsportale sollen erstmalig für die Qualitätsverbesserungen und das Reputationsmarketing niedersächsischer Krankenhäuser systematisch nutzbar gemacht werden.
Zielsetzung / Ergebnisse	Patientenerfahrungsberichte und Bewertungen in den sozialen Medien und Krankenhausvergleichsportale werden mit quantitativen und qualitativen Methoden im Sinne eines Social Media Monitoring analysiert. Ursachen für die Patientenunzufriedenheit werden systematisch erarbeitet und die Bewertungen und Themen in den sozialen Medien und Vergleichsportalen mit den Befragungen der Krankenkassen verglichen. Die Erkenntnisse über die Potenziale der Bewertungen und Themen in den sozialen Medien und Vergleichsportalen werden ergänzt durch eine Befragung von Krankenhäusern und in einem Expertenworkshop im Sinne eines Innovationstransfers zu Handlungsempfehlungen für niedersächsischer Krankenhäusern weiterentwickelt. Eine automatische textmining-basierte Identifikation relevanter Patientenberichte wird entwickelt.
Beteiligte Mitarbeiter*innen	Prof. Dr. Uwe Sander, Felix Hopf, Dr. Christiane Patzelt, Prof. Dr. Christian Wartena, Prof. Gunnar Spellmeyer
Beteiligte Institutionen	– Friedrich-Alexander-Universität Erlangen-Nürnberg, Fachbereich Wirtschaftswissenschaften, Lehrstuhl für Gesundheitsmanagement (PD Dr. Martin Emmert) – Metropolregion Hannover, Braunschweig, Göttingen, Wolfsburg GmbH – Qualitätsinitiative – Niedersächsischer Verein zur Förderung der Qualität im Gesundheitswesen e.V. – Zentrum für Qualität und Management im Gesundheitswesen (ZQ) der Ärztekammer Niedersachsen – Techniker Krankenkasse, Landesvertretung Niedersachsen

Geldgeber / Förderlinie	EFRE - Europäischer Fonds für regionale Entwicklung, Innovation durch Hochschulen und Forschungseinrichtungen, Anwendungsorientierte Forschung an Fachhochschulen
Finanzvolumen	340.712 EUR € (mit Eigenanteil)
Projektwebseite	`http://www.public-reporting.wp.hs-hannover.de/projekte/`

In den Projekten PasoMe und ZURESO wurden verschiedene Verfahren der Datenanalyse (Mixed Methods Ansatz) erprobt und kombiniert. Die qualitative Inhaltsanalyse nach Kuckartz von Texten in den sozialen Medien ermöglicht eine sehr differenzierte, aber zeitaufwändige Kategorisierung von Inhalten und bildet den Goldstandard, mit dem andere Verfahren verglichen werden. Eine quantitative automatisierte Inhaltsanalyse mithilfe von Wortlisten mit Synonymen von Begriffen ist für bestimmte Themen (etwa Berufsgruppen, die bewertet werden, oder Krankheitsbilder, über die Patienten berichten) erfolgversprechend. Auch Textmining Verfahren wurden erprobt, etwa Machine Learning Verfahren zur Analyse der Texte mit einem Joint topic-sentiment model. Eine wesentliche Forschungsfragestellung ist dabei, Erkenntnisse über die Ursachen für die Patientenunzufriedenheit oder Zufriedenheit herauszuarbeiten. Dafür untersuchen wir den Zusammenhang zwischen positiven (negativen) Aussagen mit der (Nicht-)Weiterempfehlung oder mit Sterneratings mittels logistischer Regression. So wurde eine multivariate Analyse der Assoziation des Vorhandenseins positiver Aussagen zu Aspekten der Versorgung mit der Weiterempfehlung und negativer Aussagen mit der Nicht-Weiterempfehlung durchgeführt. Die Ergebnisse dieser Projekte wurden auf mehreren Kongressen vorgestellt. Drei Veröffentlichungen in peer reviewed Fachzeitschriften wurden eingereicht.

Veröffentlichungen

1. Holpf, F., Patzelt, C., Emmert, Martin Sander, U.: Sentiment independent topic detection in rated hospital reviews. In: 19. Deutscher Kongress für Versorgungsforschung (DKVF). German Medical Science GMS, Düsseldorf (2020)
2. Wartena, C., Sander, U., Patzelt, C.: Sentiment independent topic detection in rated hospital reviews. In: Proceedings of the 13th International Conference on Computational Semantics - Short Papers. pp. 59–64. Association for Computational Linguistics, Gothenburg, Sweden (May 2019). https://doi.org/10.18653/v1/W19-0509, `https://www.aclweb.org/anthology/W19-0509`

ViTAWiN

Virtuell-augmentiertes Training für die Aus- und Weiterbildung in der interprofessionellen Notfallversorgung

Projekttitel	Virtuell-augmentiertes Training für die Aus- und Weiterbildung in der interprofessionellen Notfallversorgung - ViTAWiN
Projektlaufzeit	03.2019 bis 02.2022
Kurzbeschreibung	Konsortialleitung und Medientechnische Erforschung
Zielsetzung / Ergebnisse	Entwicklung und Erprobung einer Mehrbenutzer-Mixed Reality (MXR)- Simulationsumgebung, Konzeption, Entwicklung und Evaluation neuer Interaktionskonzepte für die haptische Augmentierung virtueller Szenarien, für die soziale Interaktion mit virtuellen Charakteren, Steigerung der Präsenz in vernetzten virtuellen Trainingswelten
Beteiligte Mitarbeiter*innen	Prof. Dr. Jonas Schild, Dr. Guillermo Carbonell, David Scherfgen, M.Sc, Sebastian Misztal, M.Sc., Leonard Flock, M.A.
Beteiligte Institutionen	Fraunhofer Institut für Experimentelles Software Engineering IESE, TriCAT GmbH, Malteser Hilfsdienst gGmbH, Bildungszentrum der Region Hessen/Rheinland-Pfalz/Saarland (HRS), Johanniter-Unfall-Hilfe e. V., Hanse Institut Oldenburg – Bildung und Gesundheit gGmbH
Geldgeber / Förderlinie	BMBF – Bundesministerium für Bildung und Forschung / Virtuelle und Erweiterte Realität (VR/AR) in der beruflichen Bildung (VRARBB) (FKZ 01PV18006)
Finanzvolumen	875.661 € (Anteil HsH), 2.697.559 € (Gesamtvolumen)
Projektwebseite	`https://www.vitawin.info/`

Das Forschungsprojekt ViTAWiN (Virtuell-augmentiertes Training für die Aus- und Weiterbildung in der interprofessionellen Notfallversorgung) wird im Rahmen des Programms Virtuelle und Erweiterte Realität (VR/AR) in der beruflichen Bildung (VRARBB) durch das Bundesministerium für Bildung und Forschung gefördert (Förderkennzeichen 01PV18006).

Die Forschungsgruppe IREX der Abteilung Informatik an der Fakultät IV leitet ein interdisziplinäres Konsortium mit Partnern aus der Wirtschaft und aus Bildungs- und Forschungsinstituten. Das Projekt zielt darauf ab, eine Mehrbenutzer-Mixed Reality (MX)- Simulationsumgebung zu entwickeln, zu erproben und zu evaluieren. Die Simulationsumgebung adressiert die Bildungsbedarfe der Lehrenden und der Lernenden in der Ausbildung zum Notfallsanitäter bzw. zur Notfallsanitäterin und der Weiterbildung zur Notfallpflege. Dabei stützt sich das Projekt ViTAWiN auf eine im BMBF/ESF-Projekt EPICSAVE (FKZ 01PD15004)

bereits entwickelte VR-Trainingsumgebung für die Notfallsanitäterausbildung ([11,12,4]).

Von einer technischen Perspektive her steht das ViTAWiN-Konsortium vor der Herausforderung, eine haptisch augmentierte, virtuelle Mehrbenutzer-Lern- und Trainingsumgebung zu entwickeln. Dafür ist eine innovative Integration von Simulationspuppen mit einer realistischen und dynamisch veränderlichen, visuellen Darstellung aller Leitsymptome erforderlich. Dieser moderne und innovativere Ansatz geht mit der Erforschung der Interaktion zwischen Mensch und VR-Technologie einher. Aus diesem Grund sind Themen der Mensch-Computer Interaktion, wie der Einfluss des Präsenzerlebens ([11,4]), des sozialen und kollaborativen Arbeitens ([2,3,12]) und spielerischer Elemente ([13,10,8]) gegenüber negativen Faktoren wie insb. Simulator Sickness ([1]) und Grenzen der Navigation ([5]) auf die Nutzerinnen und Nutzer der VR-Anwendung, von größter Relevanz für das Projekt.

Die realitätsnahen und verräumlichten 3D-Darstellungen in VR-Umgebungen mit einer haptischen Augmentierung knüpfen unmittelbar an berufliche Alltagserfahrungen an und sind so in hohem Maße authentisch und kontextualisiert. Durch den Technologieeinsatz werden – entsprechend der Bildungsbedarfe – praxisnahe Lern- und Arbeitserfahrungen für unterschiedliche, curriculare Lernfelder ermöglicht. Hierfür ist eine iterative Analyse der Bedarfe der Zielgruppen nötig. Seit Beginn des Projekts haben verschiedene Design-Thinking-Workshops mit Lernenden und Lehrenden stattgefunden, die darauf abzielten, Bedarfe zu identifizieren. Auch wurden Train-the-Trainer-Workshops und ein Beiratstreffen durchgeführt, um den aktuellen Stand der technischen Entwicklung zu evaluieren.

Bisher sind alle vorgesehenen Ziele des Projekts, wie die didaktische und medientechnische Konzeption und Evaluation eines komplexen VR-Trainingsszenarios auf multiprofessioneller Ebene, erreicht. Im Rahmen des Projekts müssen künftig verschiedene Forschungsthemen erörtert und Ziele erreicht werden, wie z.B. die didaktische Integration der VR-Anwendung in die Bildungspraxis, die kontinuierliche Förderung der interprofessionellen Arbeit und die Schaffung der Bedingungen, um die Simulationspuppe in das VR-Szenario zu integrieren. Ein wichtiger dabei ist die Verlagerung von nicht-visuellen Wahrnehmungen auf den visuellen Kanal ([6]), insb. zur Erzeugung von Empathie mit Patienten ([9,7]) und intensivierung von Trainingserlebnissen.

Veröffentlichungen

1. Buhler, H., Misztal, S., Schild, J.: Reducing vr sickness through peripheral visual effects. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 517–9 (March 2018). https://doi.org/10.1109/VR.2018.8446346
2. Buhler, H., Schild, J., Seele, S., Herpers, R.: Integration von panorama-bilddaten in eine echtzeit-game engine für virtual reality szenen. In: Virtuelle und Erweiterte Realität 13. Workshop der GI-Fachgruppe VR/AR. pp. 129 – 136. Shaker (2016)
3. Lerner, D., Luiz, T., Franke, A., Pranghofer, J., Blum, E., Gorodilova, L., Kemp, C., Runggaldier, K., Neuberger, M., Schild, J., Seele, S., Herkersdorf, M.: Serious games und virtual reality in der Ausbildung? das projekt „epicsave“! Deutscher Interdisziplinärer Notfallmedizin Kongress DINK 2016 pp. 234–237 (2016)
4. Lerner, D., Mohr, S., Schild, J., Göring, M., Luiz, T.: An immersive multi-user virtual reality for emergency simulation training: Usability study. JMIR Serious Games **8**(3), e18822 (2020)
5. Misztal, S., Carbonell, G., Ganther, N., Schild, J.: Portals with a twist: Cable twist-free natural walking in room-scaled virtual reality. In: 26th ACM Symposium on Virtual Reality Software and Technology. VRST '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3385956.3422109, `https://doi.org/10.1145/3385956.3422109`
6. Misztal, S., Carbonell, G., Schild, J.: Visual delegates – enhancing player perception by visually delegating player character sensation. In: Proceedings of the Annual Symposium on Computer-Human Interaction in Play. CHI PLAY '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3410404.3414238, `https://doi.org/10.1145/3410404.3414238`
7. Misztal, S., Carbonell, G., Zander, L., Schild, J.: Intensifying stress perception using visual effects in vr games. FDG '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3402942.3409781, `https://doi.org/10.1145/3402942.3409781`
8. Misztal, S., Carbonell, G., Zander, L., Schild, J.: Jumphair: Improving jumping performance in first-person video games through visual assistance. In: International Conference on the Foundations of Digital Games. FDG '20, Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3402942.3403020, `https://doi.org/10.1145/3402942.3403020`
9. Misztal, S., Carbonell, G., Zander, L., Schild, J.: Simulating illness: Experiencing visual migraine impairments in virtual reality. In: 2020 IEEE 8th International Conference on Serious Games and Applications for Health (SeGAH). IEEE, Piscataway, New Jersey, USA (2020)
10. Schild, J., Flock, L., Martens, P., Roth, B., Schünemann, N., Heller, E., Misztal, S.: Epicsave lifesaving decisions – a collaborative vr training game sketch for paramedics. In: 2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 1389–1389 (March 2019). https://doi.org/10.1109/VR.2019.8798365
11. Schild, J., Lerner, D., Misztal, S., Luiz, T.: Epicsave — enhancing vocational training for paramedics with multi-user virtual reality. In: 2018 IEEE 6th International Conference on Serious Games and Applications for Health (SeGAH). pp. 1–8 (May 2018). https://doi.org/10.1109/SeGAH.2018.8401353
12. Schild, J., Misztal, S., Roth, B., Flock, L., Luiz, T., Lerner, D., Herkersdorf, M., Wegner, K., Neuberger, M., Franke, A., Kemp, C., Pranqhofer, J., Seele, S., Buhler, H., Herpers, R.: Applying multi-user virtual reality to collaborative medical training. In: 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). pp. 775–776 (March 2018). https://doi.org/10.1109/VR.2018.8446160

13. Wegner, K., Seele, S., Buhler, H., Misztal, S., Herpers, R., Schild, J.: Comparison of two inventory design concepts in a collaborative virtual reality serious game. In: Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play. p. 323–329. CHI PLAY '17 Extended Abstracts, Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3130859.3131300

GLACIER

Angriffserkennung durch multidimensionale Analyse sicherheitsrelevanter Datenströme

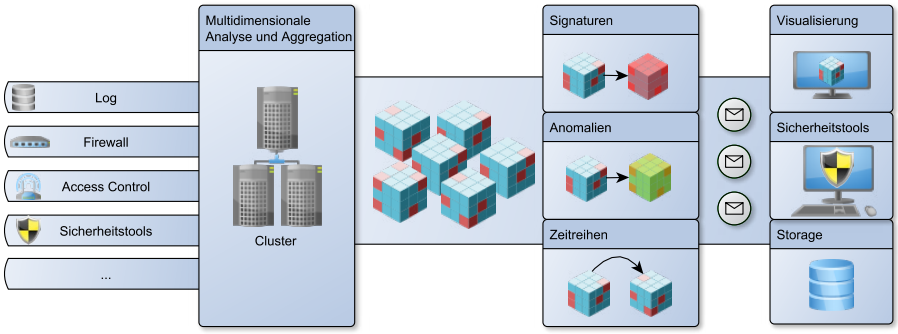
Projekttitel	Angriffserkennung durch multidimensionale Analyse sicherheitsrelevanter Datenströme – GLACIER
Projektlaufzeit	04.2019 bis 09.2021
Kurzbeschreibung	Erforschung neuer Konzepte für die Analyse von SIEM Datenströmen
Zielsetzung / Ergebnisse	Ausarbeitung einer SIEM-System Architektur, Implementierung neuartiger Algorithmen zur multidimensionalen Datenanalyse, Entwicklung einer GUI zur Visualisierung der Analyseergebnisse, Entwicklung einer simulierten Netzwerkumgebung zur Erzeugung von Trainingsdaten, Veröffentlichungen
Beteiligte Mitarbeiter*innen	Prof. Dr. Volker Ahlers, Prof. Dr. Felix Heine, Prof. Dr. Carsten Kleiner, Tim Laue M.Sc., Philip Klostermeyer B.Sc., Nils Wellermann B.Sc.
Beteiligte Institutionen	DECOIT GmbH, rt-solutions.de GmbH
Geldgeber / Förderlinie	BMBF – Bundesministerium für Bildung und Forschung / KMU-innovativ (FKZ 16KIS0950)
Finanzvolumen	440.268 € (Anteil HsH)
Projektwebseite	`https://www.glacier-project.de/`, `http://trust.f4.hs-hannover.de/`

Gegenstand des Vorhabens ist die Entwicklung von fortgeschrittenen Konzepten zur automatischen Aggregation und Analyse sicherheitsrelevanter Netzwerkdaten. Durch die automatisierte Aggregation und Analyse werden nicht nur Anomalien erkannt, es wird auch die Sicht auf die Daten (Aggregationsebene), die deren Fehlverhalten beschreibt, direkt aufgezeigt. Eine manuelle Beschreibung der Aggregation für jede Dimensionskombination entfällt, wodurch die Konfiguration des Systems entscheidend erleichtert wird. Das Forschungsvorhaben GLACIER baut auf den früheren Drittmittelprojekten CLEARER und VisITMeta sowie weiteren Vorarbeiten der Arbeitsgruppe Trust@HsH auf.

Besondere Bedeutung kommt der automatischen und effizienten Verarbeitung zu, die auch mit regulärer Hardware zu bewältigen ist, bzw. mittels horizontaler Skalierung ermöglicht werden muss. Speziell im Rahmen dieser gesteigerten Automatisierung der Auswertungsprozesse bekommt auch die Interpretation der Ergebnisse eine erhöhte Relevanz. Zum einen sollten Ergebnisse und deren Entscheidungsgrundlage auf Basis einer formalisierten Beschreibung (Indicator of Compromise) mit anderen Monitoring- und SIEM-Systemen (Security Information and Event Management [3]) geteilt werden.

Zum anderen muss für die Unterstützung manueller Analyseprozesse die Nachvollziehbarkeit [6], [7] durch einen menschlichen Security-Analysten gewährleistet werden, um angemessene Maßnahmen ergreifen zu können. Hierfür sind Konzepte einer geeigneten Visualisierung relevanter Informationen von zentraler Bedeutung [1], [2], [8].

Im Kern wird daher das Projekt GLACIER effiziente Verfahren zur automatisierten multidimensionalen Datenanalyse entwickeln. Eine komplexe Aufgabe stellt dabei die Generierung der unterschiedlichen Dimensionskombinationen mit einem hohen Maß an zu verarbeitenden Daten dar (OLAP-Cubing [4], [5]). Aus diesem Grund ist insbesondere die möglichst effiziente sowie parallele Verarbeitung ein Hauptaugenmerk des Forschungsvorhabens. Eine vielversprechende Möglichkeit stellen Verfahren dar, die frühzeitig irrelevante Informationen erkennen und somit verwerfen können bzw. nicht weiter betrachten.



Veröffentlichungen

1. Ahlers, V., Hellmann, B., Dreo Rodosek, G.: A user study of the visualization-assisted evaluation and management of network security detection events and policies. In: 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). vol. 2, pp. 668–673 (2019)
2. Ahlers, V., Hellmann, B.: Visual analytics of network security metadata (invited paper). In: Proceedings of IWEIC 2019 Hiroshima: International Workshop on Electronics, Information and Communication. pp. 1–4. Hiroshima City University, Hiroshima (2019)
3. Detken, K.O., Jahnke, M., Kleiner, C., Rohde, M.: Combining network access control (NAC) and SIEM functionality based on open source. In: Proceedings of the 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS): Technology and Applications. University Politehnica of Bucharest (Rumania), Bucharest (September 2017)
4. Heine, F.: Outlier detection in data streams using OLAP cubes. In: Kirikova, M., Nørvåg, K., Papadopoulos, G.A., Gamper, J., Wrembel, R., Darmont, J., Rizzi, S. (eds.) New Trends in Databases and Information Systems. pp. 29–36. Springer International Publishing, Cham (2017)
5. Heine, F., Rohde, M.: Popup-cubing: An algorithm to efficiently use iceberg cubes in data streams. In: Proceedings of the Fourth IEEE/ACM International Conference on

Big Data Computing, Applications and Technologies. p. 11–20. BDCAT '17, Association for Computing Machinery, New York, NY, USA (2017), https://doi.org/10.1145/3148055.3148061
6. Hellmann, B., Ahlers, V., Dreo Rodosek, G.: Integrating visual analysis of network security and management of detection system configurations. In: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). vol. 2, pp. 1020–1025 (2017)
7. Renners, L., Heine, F., Kleiner, C., Dreo Rodosek, G.: A feedback-based evaluation approach for the continuous adjustment of incident prioritization. In: 2018 1st International Conference on Data Intelligence and Security (ICDIS). pp. 176–183 (2018)
8. Renners, L., Heine, F., Kleiner, C., Dreo Rodosek, G.: Design and evaluation of an approach for feedback-based adaptation of incident prioritization. In: 2019 2nd International Conference on Data Intelligence and Security (ICDIS). pp. 28–35 (2019)

QualiPRO

Qualitätstransparenz in der Hüftendoprothetik durch Patient Reported Outcomes

Projekttitel	QualiPRO - Qualitätstransparenz in der Hüftendoprothetik durch Patient Reported Outcomes.
Projektlaufzeit	01.05.2019 bis 30.04.2022
Kurzbeschreibung	Die Qualitätssicherung (QS) in der Hüftendoprothetik umfasst zwar unterschiedliche QS-Systeme, aber derzeit nicht die Ergebnisqualität aus der Patientenperspektive (Patient Reported Outcomes; PROs). Damit ist sowohl die Messung des langfristigen Behandlungserfolges als auch die Beurteilung der Qualität von Kliniken nur eingeschränkt möglich. Daher wird im Rahmen von QualiPRO einerseits evaluiert, inwiefern eine sektorenübergreifende Erhebung von PROs zur Qualitätsmessung von Kliniken möglich ist sowie anderseits untersucht, ob PROs für die Klinikwahl sowie das interne Qualitätsmanagement akzeptiert werden. Die PROs werden im Rahmen einer Befragung AOK-versicherte Patienten in Niedersachsen zu vier unterschiedlichen Zeitpunkten erhoben..
Zielsetzung / Ergebnisse	Es wird untersucht, ob die erhobenen PROs einen Zusammenhang mit klinischen Versorgungsdaten zulassen, einen Rückschluss auf die Qualität von Kliniken zulassen sowie sich als Qualitätsindikator eignen. Unter Anwendung des Technologieakzeptanzmodells UTAUT und der Discreet Choice Analyse (DCE) wird zudem untersucht, welche Rolle PROs für die Klinikwahl und das interne Qualitätsmanagement spielen.
Beteiligte Mitarbeiter*innen	Prof. Dr. Uwe Sander, Prof. Dr. Cornelia Frömke, Susann Kiss, Dr. Christiane Patzelt, Dr. Anja Schindler
Beteiligte Institutionen	– Friedrich-Alexander-Universität Erlangen-Nürnberg, Fachbereich Wirtschaftswissenschaften, Lehrstuhl für Gesundheitsmanagement (PD Dr. Martin Emmert) – AOK Niedersachsen – DIAKOVERE Annastift gGmbH, Orthopädische Klinik der Medizinischen Hochschule Hannover – Herzogin Elisabeth Hospital Braunschweig, Orthopädische Klinik

Geldgeber / Förderlinie	Gemeinsamer Bundesausschuss –Innovationsausschuss, themenspezifische Förderung von Versorgungsforschung gemäß § 92a Abs. 2 Satz 1 SGB V
Finanzvolumen	540.652 € (Anteil HsH), 1.421.860 € (Gesamtvolumen)
Projektwebseite	`http://www.public-reporting.wp.hs-hannover.de/projekte/`

Die Anforderungen an die Datenanalyse in diesem Projekt sind vielfältig. Die zugrundeliegenden Methoden sind der Versorgungsforschung zuzuordnen. Es erfolgen primär- und Sekundärdatenerhebungen und eine Data-Linkage. So werden Patienten zur Ergebnisqualität aus der Patientenperspektive (Patient Reported Outcomes; PROs) befragt und Routinedaten der Krankenkasse AOKN (Abrechnungsdaten) damit verknüpft. Die Datenschutzanforderungen hierbei sind umfangreich. Eine Vertrauensstelle wurde eingerichtet. Zudem erheben wir Präferenzen von Patienten sowie niedergelassenen und klinisch tätigen Orthopäden und Unfallchirurgen durch Befragungen mit vorgeschalteten Interviews und Fokusgruppen. Als Befragungsinstrumente werden das Technologieakzeptanzmodel UTAUT und die Discreet Choice Analyse (DCE) verwendet. Die Auswertungen erfolgen mithilfe von Strukturgleichungsmodellen und Regressionsanalysen.

ZLG
Zukunftslabor Gesundheit

Projekttitel	Zukunftslabor Gesundheit des Zentrums für digitale Innovationen Niedersachsen (ZDIN)
Projektlaufzeit	10.2019 bis 09.2024
Kurzbeschreibung	Forschungsverbund zur Digitalisierung der Gesundheitsversorgung in Niedersachsen
Zielsetzung / Ergebnisse	Vernetzung zur Digitalisierung im Gesundheitswesen forschender Einrichtungen in Niedersachsen; Aufbau einer transsektoralen Forschungsdatenplattform zur kollaborativen Forschung; Innovative Konzepte für die Nutzung von Sensordaten in Forschung und Versorgung; Entwicklung von Online-Lehrangeboten für die Aus- Fort- und Weiterbildung zu Themen der Digitalisierung des Gesundheitswesens.
Beteiligte Mitarbeiter*innen	Prof. Dr. Oliver J. Bott, Matthias Katzensteiner
Beteiligte Institutionen	Universitätsmedizin Göttingen, Institut für Medizinische Informatik (UMG) Hochschule Hannover Fakultät III - Medien, Information und Design (HsH) Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der MHH, Standort Hannover (PLRI H) Peter L. Reichertz Institut für Medizinische Informatik der TU Braunschweig und der MHH, Standort Braunschweig (PLRI BS) Gesundheitscampus Göttingen (HAWK) Leibniz Universität, Hannover L3S Forschungszentrum (L3S LUH) Carl von Ossietzky Universität Oldenburg (OL) Jade Hochschule Oldenburg (JOL) Hochschule Osnabrück (HOS) Leibniz Universität Hannover (LUH)
Geldgeber / Förderlinie	Volkswagen Stiftung und Niedersächsisches Ministerium für Wissenschaft und Kultur (Niedersächsisches Vorab)
Finanzvolumen	3.947.635 €(mit Eigenanteil), davon 364.040 €für die Hochschule Hannover
Projektwebseite	`https://www.zdin.de/zukunftslabore/gesundheit`

Das Zukunftslabor Gesundheit (ZLG) ist ein Transferverbundprojekt im Rahmen des Zentrums für Innovation Niedersachsen (ZDIN). Neben der HsH sind folgende Partner beteiligt: UMG, PLRI, HAWK, L3S, LUH, OL, JOL, HOS.

Als übergreifendes Ziel der verschiedenen Zukunftslabore des ZDIN im Allgemeinen steht die Vernetzung der Forschungseinrichtungen untereinander und die Kooperation mit der freien Wirtschaft. Die Ziele des ZLG im Besonderen sind die Etablierung einer gemeinsamen Forschungsplattform, die Entwicklung von Systemen und Algorithmen zur Erschließung des Lebensumfeldes von Patienten durch eine Auswahl hierfür geeigneter Sensorik und die Entwicklung von Online-Lehrangeboten für die Aus- Fort- und Weiterbildung zu Themen des ZLG.

Die HsH ist in Teilprojekten 1 („Datenanalyse und –austausch") und 3 („Aus-, Fort- und Weiterbildung") beteiligt und forscht in Teilprojekt 1 unter anderem an der Entwicklung interoperabler Datenmodelle, privatheitsbewahrender Datenanalysemethoden und einrichtungsübergreifender Datenintegrationsmethoden. Weiter ist die HsH an der Erstellung eines didaktischen Konzeptes für die Online-Wissensvermittlung im Teilprojekt 3 befasst und unterstützt dessen Nutzung im Rahmen der Entwicklung von Online-Kursangeboten. Zielgruppen für das Kursangebot sind neben Medizininformatiker*innen und Vertreter*innen verwandter Berufe auch Patient*innen und Betroffene sowie Bürger*innen und die interessierte Öffentlichkeit.

Aktueller Arbeitsstand ist die Konzeptionierung einer interoperablen Austauschplattform zur einrichtungsübergreifenden Datenaggregation und -analyse. In diesem Zusammenhang wird zum Zeitpunkt dieser Veröffentlichung an ersten Konzepten zur Pseudonymisierung oder Synthetisierung personenbezogener klinischer Daten gearbeitet. In Anlehnung an das didaktische Konzept des HiGHmed-Lehre Projekts [1] wird ein auf die spezifischen Zielgruppen des ZLG abgestimmtes didaktisches Konzept für die Online-Lehre erarbeitet und voraussichtlich zum Jahresende fertig gestellt.

In den nächsten Schritten soll die Assoziation vom Konsortium HiGHmed der Medizininformatikinitiative für Teilprojekt 1 angestoßen werden, um eine vernetzte Zusammenarbeit in der Entwicklung interoperabler Datenmodelle und der angestrebten Forschungsplattform zur Datenanalyse zu ermöglichen. Zudem sollen die ersten Ergebnisse der Teilprojekte 1 und 2 als Ausgangspunkt der Entwicklung erster Fortbildungsangebote im Rahmen des Teilprojektes 3 genutzt werden.

Gefördert vom Niedersächsischen Ministerium für Wissenschaft und Kultur unter Fördernummer ZN3491 im Niedersächsischen Vorab der Volkswagen Stiftung und betreut vom Zentrum für digitale Innovationen (ZDIN).

Veröffentlichungen

1. Witte, M.L., Behrends, M., Benning, N.H., Hoffmann, I., Bott, O.J.: The highmed didactical framework for online learning modules on medical informatics: First experiences. Studies in health technology and informatics **272**, 163–166 (2020). https://doi.org/10.3233/SHTI200519

JuVer

Extraktion von juristischen Verweisstrukturen: Intertextualität in Gesetzen, Kommentaren und Gerichtsurteilen

Projekttitel	Extraktion von juristischen Verweisstrukturen – JuVer
Projektlaufzeit	10.2019 bis 09.2022
Kurzbeschreibung	Analyse der Beziehungen zwischen juristischen Dokumenten
Zielsetzung / Ergebnisse	Aufbau von Analyse-Prototypen, Erstellung und Veröffentlichung von Toolkits
Beteiligte Mitarbeiter der HsH	Prof. Dr. Christian Wartena, Prof. Dr. Fabian Schmieder, Jean Charbonnier M.A., Frieda Josi M.A.
Beteiligte Institutionen	
Geldgeber / Förderlinie	MKW – Ministerium für Wissenschaft und Kultur, Niedersächsisches Vorab, VolkswagenStiftung - Wissenschaftskommunikation und Wissenstransfer
Finanzvolumen	320.200 €
Projektwebseite	`textmining.wp.hs-hannover.de/juver.html`

In der Rechtswissenschaft spielen Texte eine zentrale Rolle und werden im Kontext weiterer Texte interpretiert. In diesem Projekt sollen Verfahren entwickelt werden, um diese komplexen Zusammenhänge innerhalb und zwischen den Texten explizit zu machen. Das Projekt beinhaltet die Entwicklung einer Pilotanwendung, den Aufbau von Datensammlungen bestehend aus juristischen Texten und die Analyse der Bezüge zwischen den Verträgen. Dabei werden Methoden aus der Computerlinguistik, dem Maschinellen Lernen und Semantic Web eingesetzt um Verweise zwischen den Texten systematisch zu erfassen. Dabei ist die Zielsetzungen einen Wissensgraph aufzubauen, der die Beziehungen zwischen Gesetzestexten, juristischer Fachliteratur wie Kommentaren und Fachartikeln sowie der Rechtsprechung, aufzeigt.

Die Arbeit im Projekt konzentrierte sich bisher auf den Aufbau von Textkorpora und die Identifikation von Standardtexten. Viele Texte liegen in dem juristischen Bereich nur als PDF mit komplexer Formatierung vor. Die Extraktion der Texte stellt hier eine Herausforderung dar, die mit Hilfe von maschinellen Lernverfahren angegangen wurde [1]. Durch die hier vorgenommene Analyse von Layout-Merkmalen kann der Textfluss größtenteils erkannt und die Funktion einzelner Textteile bestimmt werden. Um Überschriften in Texte noch besser zu erkennen, ist ein Abgleich mit bekannten Überschriften sehr hilfreich. Da oft kleine textuelle Änderungen auftreten, haben wir in [2] verschiedene Methoden für die Ähnlichkeitsbestimmung von kurzen Textabschnitten untersucht.

Im weiteren Kontext des Projektes wurde ein Programm zur Lemmatisierung und Wortartbestimmung deutscher Texte entwickelt [3]. Das Tool gibt leicht

bessere Ergebnisse für die Lemmatisierung als bisher verfügbare Programme, es sichert aber vor allem die Unabhängigkeit von Software, die nicht Open-Source und für alle Zwecke frei verfügbar ist.

Veröffentlichungen

1. Josi, F., Wartena, C.: Structural Analysis of Contract Renewals. In: Proceedings of the CIKM 2018 Workshops. vol. 2482. CEUR-WS, Torino (Nov 2019), http://ceur-ws.org/Vol-2482/paper31.pdf
2. Josi, F., Wartena, C., Heid, U.: Detecting Paraphrases of Standard Clause Titles in Insurance Contracts. In: RELATIONS - Workshop on meaning relations between phrases and sentences. Association for Computational Linguistics, Gothenburg, Sweden (2019), https://www.aclweb.org/anthology/W19-0803
3. Wartena, C.: A Probabilistic Morphology Model for German Lemmatization. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers. pp. 40–49. German Society for Computational Linguistics & Language Technology, Erlangen (Oct 2019), https://corpora.linguistik.uni-erlangen.de/data/konvens/proceedings/papers/KONVENS2019_paper_10.pdf

HsH-PROFIL

Kooperation HsH-Forschungscluster 'Smart Data Analytics' mit Forschungsinstitut CETINIA in Madrid

Projekttitel	Kooperation HsH-Forschungscluster 'Smart Data Analytics' mit Forschungsinstitut CETINIA in Madrid
Projektlaufzeit	01.06.2020 bis 30.06.2021
Kurzbeschreibung	Institutionalisierung der wissenschaftlichen Zusammenarbeit zwischen der Hochschule Hannover (HsH) und der Universidad Rey Juan Carlos (URJC) in Madrid
Zielsetzung / Ergebnisse	Ziel ist die Anbahnung einer Kooperationsvereinbarung zwischen dem HsH-Forschungscluster ‚Smart Data Analytics' und dem URJC-Forschungsinstitut CETINIA (Centre for Intelligent Information Technologies)
Beteiligte Mitarbeiter*innen	Prof. Dr. Ralf Bruns, Prof. Dr. Jürgen Dunkel
Beteiligte Institutionen	Universidad Rey Juan Carlos (URJC) Madrid, Spanien
Geldgeber / Förderlinie	Niedersächsisches Ministerium für Wissenschaft und Kultur, Förderformat Profil-INTERNATIONAL
Finanzvolumen	13.000 €
Projektwebseite	`http://sw-architecture.inform.hs-hannover.de/`

Das Ziel des Projektvorhabens ist die Institutionalisierung der wissenschaftlichen Zusammenarbeit zwischen der Hochschule Hannover (HsH) und der Universidad Rey Juan Carlos (URJC) in Madrid. Als Kernpunkt soll eine Kooperationsvereinbarung zwischen dem HsH-Forschungscluster ‚Smart Data Analytics' und dem URJC-Forschungsinstitut CETINIA (Centre for Intelligent Information Technologies) getroffen und flankierend in Form eines gemeinsamen Forschungsprojekts mit Leben gefüllt werden. Mit dem hier vorgestellten Projektvorhaben soll die Zusammenarbeit zwischen der HsH und der URJC in Forschung, Studium und Lehre auf eine noch breitere Basis gestellt und verstetigt werden.

Die Umsetzung dieses Projektvorhaben erhöht die Sichtbarkeit der internationalen Forschung an der HsH, stärkt deren Drittmittelfähigkeit und hat positive Auswirkungen auf die Lehre in den Informatikstudiengängen. Darüber hinaus schärft das Vorhaben das internationale Profil der HsH im Bereich Forschung und besitzt Modellcharakter für weitere transnationale Kooperationen.

Zwischen der HsH und der URJC bestehen seit dem Jahre 2000 vielfältige Beziehungen, die bisher durch einen fortwährenden Austausch einzelner Studierender und Wissenschaftler der Informatik-Fachbereiche geprägt waren.

SenAD

Sensorintegration in Asphalt für ein Daten-basiertes Degradationsmonitoring

Projekttitel	Sensorintegration in Asphalt für ein Daten-basiertes Degradationsmonitoring
Projektlaufzeit	01.10.2020 bis 30.06.2021
Kurzbeschreibung	Entwicklung eines Messsystems für den strukturellen Zustand einer Asphalttragschicht
Zielsetzung / Ergebnisse	Entwicklung eines Messsystems in Form eines funktionalisierten Gewebes, welches eingebaut in die Asphalttragschicht eine zerstörungsfreie, kontinuierliche Messung der strukturellen Substanz (Degradationsgrad) ermöglicht
Beteiligte Mitarbeiter*innen	Prof. Dr. Ralf Bruns, Prof. Dr. Jürgen Dunkel
Beteiligte Institutionen	Hochschule Magdeburg-Stendal, Fraunhofer WKI, Fraunhofer IFAM
Geldgeber / Förderlinie	Bundesministerium für Verkehr und digitale Infrastruktur BMVI / mFUND
Finanzvolumen	12.000 €
Projektwebseite	`http://sw-architecture.inform.hs-hannover.de/`

Die Asphalttragschicht stellt einen wesentlichen Teil des Fundaments der Asphaltstraßenbefestigung dar. Der strukturelle Zustand dieser Schicht hat wesentlichen Einfluss auf die gesamte Nutzungsdauer der Befestigung. Der Beurteilung des strukturellen Zustands von Asphalttragschichten kommt deshalb besondere Bedeutung zu. Momentan kommen „zerstörende“ Messverfahren zum Einsatz, bei denen die zu untersuchende Straße zusätzlich geschädigt wird. Eine flächendeckende kontinuierliche Bewertung der strukturellen Substanz ist aufgrund des hohen Aufwandes und der zusätzlich eingetragenen Schädigung nicht realisierbar.

Ziel dieses Vorhabens ist die Entwicklung eines Messsystems in Form eines funktionalisierten Gewebes, welches eingebaut in die Asphalttragschicht eine zerstörungsfreie, kontinuierliche Messung der strukturellen Substanz (Degradationsgrad) ermöglicht. So kann der Zustand der Tragschicht zu jedem Zeitpunkt ohne nennenswerten technischen Messaufwand erfasst werden. Durch eine großflächige Integration des Messsystems in die Asphalttragschicht sind flächenübergreifende Aussagen zum Tragschichtzustand möglich. Die integrierte Messung ermöglicht somit genauere und frühzeitigere Aussagen über den Straßenzustand.

Es wird ein Hybridgewebe konstruiert, welches durch integrierte Sensorik Zustandsänderungen der Umgebungskomponenten wahrnehmen kann. Kernstück des Gewebes ist die Sensorik, welche durch Belastungen so geschädigt

und in ihren Eigenschaften verändert wird, dass dies anhand entsprechender Kennwerte ausgelesen werden kann. Die Weiterleitung und Speicherung der Daten wird ebenso untersucht wie der sichere Datentransport. Die Verarbeitung und Auswertung der Messergebnisse wird durch Adaption einer Software möglich. Zur Interpretation künftiger Messungen wird der Zusammenhang zwischen Widerstandsänderung und Degradation der Straße dargestellt.

Kombinom

Datenmodellierung für den Einsatz von autonomen Kleinbussen im ländlichen Raum zum kombinierten Transport von Personen und Gütern

Projekttitel	Datenmodellierung für den Einsatz von autonomen Kleinbussen im ländlichen Raum zum kombinierten Transport von Personen und Gütern
Projektlaufzeit	10.2020 bis 9.2021
Kurzbeschreibung	Potentialanalyse zur Umsetzung eines hybriden Transportkonzeptes für Personen und Güter im ländlichen Raum mittels autonomer Kleinbussysteme.
Zielsetzung / Ergebnisse	Analyse von Datenbeständen, Analyse von Kenngrößen, Potentialbewertung.
Beteiligte Mitarbeiter*innen	Prof. Dr. Christoph von Viebahn (Projektleitung HsH), Marvin Auf der Landwehr M.Sc.
Beteiligte Institutionen	Frankfurt University of Applied Sciences (Gesamtprojektleitung), Fachzentrum Mobilität im ländlichen Raum
Geldgeber / Förderlinie	BMVI – Bundesministerium für Verkehr und digitale Infrastruktur (FKZ 19F1072B)
Finanzvolumen	47.984 € (Anteil HsH)
Projektwebseite	n.a.

Im Rahmen von „Kombinom“ wird eine Potenzialanalyse für den Einsatz von autonomen Kleinbussen im ländlichen Raum zum kombinierten Transport von Personen und Gütern erstellt. Der Fokus liegt auf dem Zugang zu sowie der potentiellen Nutzung von adäquaten Daten. Dabei wird die Verfügbarkeit von existierenden Datenquellen im Bereich des ÖPNV überprüft und der logistische Güterverkehr im ländlichen Raum evaluiert. Fernern werden Erfolgsfaktoren von nachfrageorientierten Bedienungsformen identifiziert und die Nutzbarkeit existierender Datenquellen bzgl. der Auslastung autonomer Kleinbusse im ländlichen Raum verbessert. Auf Basis der geprüften Daten werden bestehende Datenlücken identifiziert und entsprechende Konzepte zur Erhebung fehlender Daten erstellt bzw. Expertenschätzungen zur Komplementierung von Datenlücken herangezogen. Abschließend wird eine Potenzialanalyse einer datenbasierten Anwendung (Simulation) erstellt, um die zuvor genannten Daten und Informationen zu kombinieren und entsprechende Potential systematisch zu identifizieren. Im Rahmen der Potentialanalyse werden die zusammengestellten Daten analysiert, interpretiert und evaluiert, um zu erfassen, ob und in wie weit eine umfassende Simulationsstudie die möglichen Einflüsse eines Systems, bei dem der Transport von Gütern und Personen im ländlichen Raum mittels autonomer Kleinbusse kombiniert wird,

im Hinblick auf räumliche, zeitliche und wirtschaftliche Potenziale abbilden, quantifizieren und sowohl auf konzeptioneller als auch auf struktureller Ebene bewerten kann. Die simulationsgestützte Konzeption, Analyse und Bewertung des Konzeptes soll mittels der akquirierten Daten im Nachgang im Rahmen eines nachgelagerten Verstetigungsprojektes durchgeführt werden und stützt sich auf die Potentialanalyse sowie im Vorfeld durchgeführte Simulationsstudien zu logistischen Auswirkungen in verschiedenen Kontexten.

Veröffentlichungen

1. Auf der Landwehr, M., Trott, M., Von Viebahn, C.: E-Grocery in Terms of Sustainability–Simulating the Environmental Impact of Grocery Shopping for an Urban Area in Hanover. In: Putz, M., Schlegel, A. (eds.) Simulation in Produktion und Logistik, pp. 87–96. Verlag Wissenschaftliche Scripten, Auerbach (2019)
2. Auf der Landwehr, M., Trott, M., Von Viebahn, C.: Computer simulation as evaluation tool of information systems: Identifying quality factors of simulation modeling. In: Aier, S., Gordijn, J., Proper, H., Verelst, J. (eds.) 2020 IEEE 22nd Conference on Business Informatics (CBI). vol. 1, pp. 211–220. IEEE, Los Alamitos (2020). https://doi.org/10.1109/CBI49978.2020.00030
3. Auf der Landwehr, M., Trott, M., Von Viebahn, C.: Simulation-based assessment of grocery shopping in urban areas. Simulation News Europe **30**(4) (2020), in press
4. Jordan, Y., Von Viebahn, C., Haller, S.: Konzeptionierung von Use Cases für die urbane Logistik in einer Schweizer Smart City mittels Design Thinking, pp. 391–420. Springer Fachmedien, Wiesbaden (2019). https://doi.org/10.1007/978-3-658-22101-0_16
5. Trott, M., Auf der Landwehr, M., Von Viebahn, C.: A new simulation approach for scheduling consolidation activities in intralogistics – optimising material flow processes in industrial practice pp. 325–334 (2019)
6. Trott, M., Auf der Landwehr, M., Von Viebahn, C.: E-grocery of tomorrow - home delivery of food between profitability, customer acceptance and ecological footprint. World Review of Intermodal Transportation Research (2020), in press
7. Trott, M., Von Viebahn, C., Auf der Landwehr, M. (eds.): Towards a More Sustainable Future? Simulation the Environmental Impact of Online and Offline Grocery Supply Chains (2020)
8. Von Viebahn, C., Auf der Landwehr, M., Trott, M.: The future of grocery shopping? a taxonomy-based approach to classify e-grocery fulfillment concepts. In: Gronau, N., Heine, M., Krasnova, H., Poustcchi, K. (eds.) Entwicklungen, Chancen und Herausforderungen der Digitalisierung: Proceedings der 15. Internationalen Tagung Wirtschaftsinformatik (WI 2020). pp. 955—-971. GITO, Berlin (2020). https://doi.org/10.30844/wi_2020_j2-viebahn

LernMINT

Datengestützter Unterricht in den MINT-Fächern

Projekttitel	Datengestützter Unterricht in den MINT-Fächern – LernMINT
Projektlaufzeit	2020 bis 2024
Kurzbeschreibung	Graduiertenkolleg mit dem Schwerpunkt datengestütztes Lernen und Lehren im MINT-Unterricht
Zielsetzung / Ergebnisse	Das Promotionsprogramm zielt darauf ab, eine Gruppe von hochqualifizierten Expert*innen für das adressierte Teilgebiet der Digitalen Bildung, dem datengestützten Lernen und Lehren im MINT-Unterricht, interdisziplinär auszubilden.
Beteiligte Mitarbeiter*innen	Prof. Dr. Christian Wartena, Vitor L. Fontanella
Beteiligte Institutionen	Institut für Didaktik der Naturwissenschaften der Leibniz Universität Hannover (LUH), Institut für Didaktik der Mathematik und Physik der LUH, Fakultät für Elektrotechnik und Informatik der LUH Technische Informationsbibliothek (TIB) – Hannover, Ostfalia Hochschule: Zentrum für erfolgreiches Lehren und Lernen (ZeLL), Hochschule Hannover
Geldgeber / Förderlinie	MKW – Ministerium für Wissenschaft und Kultur, Niedersächsisches Promotionsprogramm
Finanzvolumen	1 Georg-Christoph-Lichtenberg-Stipendium
Projektwebseite	`http://lernmint.org`

Das Ziel des interdisziplinären Promotionsprogramms LernMINT ist es, die Chancen, Begrenzungen und Risiken des datengestützten Lehrens und Lernens zu erforschen: Welche Möglichkeiten bietet die Digitale Transformation zur Qualitäts- und Effizienzsteigerung in Schule und Hochschule im Spannungsfeld von Fachdidaktik, Learning Analytics und Datenschutz? Bislang weitgehend ungenutzte Potenziale für den Einsatz von daten- und computergestützten, intelligenten Methoden aus den Bereichen Learning Analytics und Educational Data Mining im Schulunterricht und im Übergang zur Hochschule sollen zur Unterstützung der Lehrenden und des Lernens erschlossen und genutzt werden. Im Fokus des interdisziplinären Promotionsprogramms werden die gezielte Entwicklung und fachdidaktische Evaluierung von datengestützten, intelligenten Methoden und deren sinnvolle Einbeziehung in den Unterricht der MINT-Fächer stehen.

Das Programm zielt darauf ab, eine Gruppe von hochqualifizierten Expert*innen für das adressierte Teilgebiet der Digitalen Bildung interdisziplinär auszubilden. Dabei sind die Promotionsprojekte in fünf Themenkomplexe gegliedert:

1. Learning-Analytics-Methoden für den MINT-Unterricht
2. Datengestützte Nutzung von Lernpotenzialen
3. Informelles Lernen in Schule und Hochschule
4. Übergang von Schule zu Hochschule
5. Datenschutz, Fairness und Akzeptanz von Learning Analytics.

Verantwortungsvolle Künstliche Intelligenz in der Digitalen Gesellschaft

Projekttitel	Verantwortungsvolle Künstliche Intelligenz in der Digitalen Gesellschaft
Projektlaufzeit	2020 bis 2024
Kurzbeschreibung	Graduiertenkolleg mit dem Schwerpunkt Künstliche Intelligenz
Zielsetzung / Ergebnisse	Forschung in den Bereichen Qualität, Haftung, Interpretierbarkeit sowie verantwortungsvollem Umgang mit Informationen und der Anwendung von KI
Beteiligte Mitarbeiter*innen	Prof. Dr. Ina Blümel, Prof. Dr. Fabian Schmieder, Prof. Dr. Christian Wartena
Beteiligte Institutionen	Forschungszentrums L3S der Leibniz Universität Hannover und der Technischen Universität Braunschweig, Technische Informationsbibliothek Hannover, Hochschule Hannover.
Geldgeber / Förderlinie	MKW – Ministerium für Wissenschaft und Kultur, Niedersächsisches Promotionsprogramm
Finanzvolumen	3 Georg-Christoph-Lichtenberg-Stipendien
Projektwebseite	`https://verantwortungsvolle-ki.de/`

Im Rahmen der Digitalisierung wirken Technologien der Künstlichen Intelligenz (KI) als treibende Kraft. Aufgrund ihrer enormen gesellschaftlichen Relevanz ist ein verantwortungsvoller Umgang mit KI von besonderer Bedeutung. Die Erforschung und Anwendung verantwortungsvoller KI ist eine sehr junge Disziplin und erfordert die Bündelung von Forschungsaktivitäten aus unterschiedlichsten Disziplinen, um KI Systeme zuverlässig, transparent, sicher und rechtlich akzeptabel zu gestalten und anzuwenden.

Das Promotionsprogramm adressiert die interdisziplinären Forschungsherausforderungen im Rahmen von 14 transdisziplinären Promotionen. Organisiert in vier Clustern, erforschen Stipendiaten*innen die drängendsten Forschungsfragen in den Bereichen Qualität, Haftung, Interpretierbarkeit, verantwortungsvollem Umgang mit Informationen und der Anwendung von KI. Das innovative, zielorientierte und international ausgerichtete Betreuungskonzept und erfahrene PI-Team unterstützt die Stipendiaten*innen bei exzellenter Forschung.

ODG

Orte des Gestapoterrors im heutigen Niedersachsen

Projekttitel	Orte des Gestapoterrors im heutigen Niedersachsen
Projektlaufzeit	1.11.2020 - 31.10.2022
Kurzbeschreibung	Aufbereitung und Bereitstellung der Daten zu Orten des Gestapoterrors in Niedersachsen
Zielsetzung / Ergebnisse	Offene, interaktive digitale Landkarte zur Visualisierung und Anreicherung der Daten für Wissenschaft und Bürgerwissenschaft
Beteiligte Mitarbeiter*innen	Prof. Dr. Ina Blümel
Beteiligte Institutionen	Technische Informationsbibliothek (TIB), Stiftung niedersächsische Gedenkstätten
Geldgeber / Förderlinie	MWK (Förderprogramm „Pro*Niedersachsen – Forschungsprojekte der Geistes-, Kultur- und Sozialwissenschaften")
Finanzvolumen	246.000 €

Im Mittelpunkt des Forschungsprojektes stehen die Orte des Gestapoterrors im Bereich des heutigen Landes Niedersachsen. Die damaligen Dienstsitze und Haftstätten sollen erstmals flächendeckend als grundlegende Faktoren der Gestapotätigkeit untersucht werden.

Anhand von Standorten und Infrastruktur soll nicht nur die Tätigkeit der Gestapo erforscht, sondern auch das Wissen über deren Verbrechen und die spätere Erinnerung daran digital zugänglich gemacht werden. Als kulturelles Erbe Niedersachsens werden dabei insbesondere Gedenkstättensammlungen, themenspezifische Archivbestände und Baudenkmale erschlossen. Die während der NS-Zeit mehrfach wechselnden Standorte der Gestapo und die heutigen Erinnerungsorte sollen in einer vierdimensionalen digitalen Landkarte – mit orts- und zeitbezogenen Daten – dargestellt werden.

Bei der digitalen Erschließung und Aufbereitung der Daten wird dafür gesorgt, dass diese sowohl von Menschen als auch Maschinen interpretierbar sowie frei zugänglich sind. Sie sollen beispielsweise für die Bildungsarbeit oder für weitere Forschung verwendet werden können, insbesondere im Digital Humanities bzw. Open-GLAM-Kontext. GLAM steht für „Galleries, Libraries, Archives, Museums", also kulturelle Institutionen und Gedächtnisorganisationen, die unter dem Begriff „Open GLAM" ihre Werke öffentlich frei zugänglich machen.

Die im Projekt entwickelte digitale Landkarte bezieht neben den neuen Forschungsergebnissen auch Informationen zu digitalen Objekten ein, die bereits an anderer Stelle existieren. Insbesondere sollen Datensätze zu

historischen Stätten, Institutionen und Ereignissen verlinkt werden, die in frei zugänglichen Datenbanken wie dem Digitalen Denkmalatlas Niedersachsen, Wikidata, Wikimedia Commons, Pelagios Commons und Open Street Map hinterlegt sind.

Nach dem erfolgreichen Aufbau der Plattform können Nutzerinnen und Nutzer kontinuierlich weitere Informationen einpflegen. Damit wird Partizipation im Sinne von „Citizen Science“ – oder Bürgerwissenschaft – erreicht, wo Laien die Möglichkeit haben, sich in wissenschaftliche Projekte einzubringen.

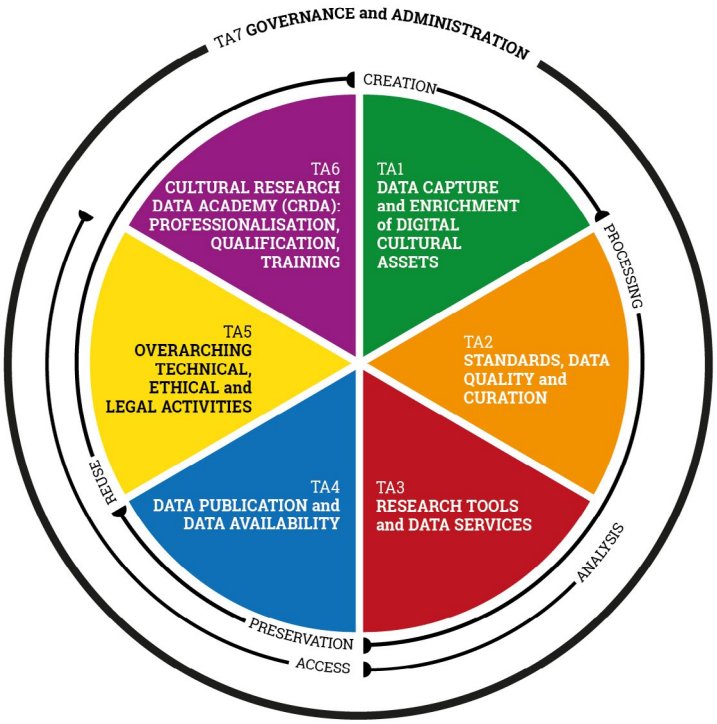
NFDI4Culture

Consortium for research data on material and immaterial cultural heritage

Projekttitel	Nationale Forschungsdateninfrastruktur zu materiellen und immateriellen Kulturgütern
Projektlaufzeit	1.10.2020 - 30.9.2025
Kurzbeschreibung	Aufbau einer bedarfsorientierten Infrastruktur für Forschungsdaten zu materiellen und immateriellen Kulturgütern
Zielsetzung / Ergebnisse	Vorantreiben von Qualifizierung und Professionalisierung in der datengestützten Forschung im Bereich des Kulturerbes. Schaffung einer nutzerzentrierten, forschungsgetriebenen Infrastruktur unter Berücksichtigung der FAIR-Prinzipien.
Beteiligte Mitarbeiter*innen	Prof. Dr. Ina Blümel, William Attatsitsey M.Sc.
Beteiligte Institutionen	Konsortialführer: Akademie der Wissenschaften Mainz
Geldgeber / Förderlinie	Deutsche Forschungsgemeinschaft (DFG)
Finanzvolumen (Anteil Blümel)	1.168.000 €
Projektwebseite	`https://nfdi4culture.de/`

Digitale Daten über materielle und immaterielle Kulturgüter sind ein wesentlicher Bestandteil des täglichen Lebens, der Kommunikation und des Erlebnisses. Sie haben einen nachhaltigen Einfluss auf die Wahrnehmung kultureller Identität sowie auf die Wechselwirkungen zwischen Forschung, Kulturwirtschaft und Gesellschaft. Im Laufe der letzten drei Jahrzehnte haben viele Institutionen des Kulturerbes eine Fülle digitaler Repräsentation von Kulturgütern (digitale 2D-Reproduktionen von Gemälden, Musiknoten, digitale 3D-Modelle von Skulpturen, Denkmälern, Räumen, Gebäuden), audiovisuelle Daten (Musik, Film, Bühnenaufführungen) und prozessorientierte Forschungsdaten wie Codierungs- und Annotationsformate beigesteuert. Die langfristige Bewahrung und FAIR-Verfügbarkeit (gemäß der FAIR-Prinzipien sollen Daten "Findable, Accessible, Interoperable, and Re-usableßein) von Forschungsdaten aus dem Bereich des Kulturerbes ist von grundlegender Bedeutung, nicht nur für den zukünftigen akademischen Erfolg in den Geisteswissenschaften, sondern auch für die kulturelle Identität des Einzelnen und der Gesellschaft als Ganzes. Bislang fehlt in Deutschland eine koordinierte Initiative für ein professionelles Forschungsdatenmanagement. NFDI4Culture will diese Lücke schließen und eine nutzerzentrierte, forschungsgetriebene Infrastruktur schaffen, die ein breites Spektrum an Forschungsbereichen von Musikwissenschaft,

Kunstgeschichte und Architektur bis hin zu Performance-, Theater-, Film- und Medienwissenschaften abdeckt. Die Forschungslandschaft, die das Konsortium anspricht, ist durch eine starke institutionelle Differenzierung gekennzeichnet. Zu den Forschungseinrichtungen in der Interessengemeinschaft des Konsortiums gehören Universitätsinstitute, Kunsthochschulen, Akademien, Galerien, Bibliotheken, Archive und Museen. Diese heterogene Landschaft zeichnet sich auch durch eine Fülle von Forschungsobjekten, Methoden und ein großes Potenzial für datengetriebene Forschung aus. In einer einmaligen Initiative der Antragsteller und Mitantragsteller des Antrags [1] sowie zehn akademischen Gesellschaften ist diese Community zum ersten Mal durch einen föderalen Ansatz miteinander verbunden, der den Bedürfnissen der beteiligten Forscher in hohem Maße gerecht wird. Zusammenarbeit innerhalb der NFDI zu fördern, Wissen und Technologie zu teilen und die Nutzer umfassend zu unterstützen, waren von Anfang an die Leitmotive des Projekts und werden im Mittelpunkt aller Arbeitsabläufe und Entscheidungsprozesse stehen. Dank dieser Prinzipien hat NFDI4Culture eine starke Unterstützung erfahren, die von einzelnen Forschern bis hin zu hochrangigen Organisationen des Kulturerbes wie der UNESCO, dem Internationalen Museumsrat, der Open Knowledge Foundation und Wikimedia reicht. Auf dieser Grundlage wird NFDI4Culture innovative Maßnahmen ergreifen, die einen kulturellen Wandel hin zu einem reflektierteren und nachhaltigeren Umgang mit Forschungsdaten fördern und gleichzeitig die Qualifizierung und Professionalisierung in der datengestützten Forschung im Bereich des Kulturerbes vorantreiben.



Veröffentlichungen

1. Altenhöner, R., Blümel, I., Boehm, F., Bove, J., Bicher, K., Bracht, C., Brand, O., Dieckmann, L., Effinger, M., Hagener, M., Hammes, A., Heller, L., Kailus, A., Kohle, H., Ludwig, J., Münzmay, A., Pittroff, S., Razum, M., Röwenstrunk, D., Sack, H., Simon, H., Schmidt, D., Schrade, T., Walzel, A.V., Wiermann, B.: Nfdi4culture - consortium for research data on material and immaterial cultural heritage. Research Ideas and Outcomes **6**, e57036 (2020). https://doi.org/10.3897/rio.6.e57036, `https://doi.org/10.3897/rio.6.e57036`

TAPIR

TeilAutomatisiertes Persistent-Identifier-basiertes Reporting

Projekttitel	TeilAutomatisiertes Persistent-Identifier-basiertes Reporting
Projektlaufzeit	1.11.2020 bis 31.10.2022
Kurzbeschreibung	Erprobung von teilautomatisierten Verfahren zur Forschungsberichterstattung im Kontext universitärer und außeruniversitärer Forschung
Zielsetzung / Ergebnisse	Harvestings von Forschungsinformationen und teilautomatisierte Erstellung von Forschungsberichten
Beteiligte Mitarbeiter*innen	Prof. Dr. Ina Blümel
Beteiligte Institutionen	TIB Hannover, Universität Osnabrück
Geldgeber	Bundesministerium für Bildung und Forschung (BMBF)
Finanzvolumen (Blümel)	167.739 €

Im Verbundvorhaben TAPIR werden teilautomatisierte Verfahren zur Forschungsberichterstattung im Kontext universitärer und außeruniversitärer Forschung erprobt. Es werden die Möglichkeiten zur Realisierung eines teilautomatisierten Reportings analysiert und es wird der Frage nachgegangen, inwiefern die dazu notwendige Datenaggregation auf Basis offen verfügbarer Forschungsinformationen mittels persistenter Identifikatoren durchgeführt werden kann.

Im Rahmen des Teilvorhabens "Datenanalyse und Use Cases"werden beim Verbundpartner Universität Osnabrück praxisrelevante Anwendungsfälle identifizert und die Ergebnisse hinsichtlich der Datenqualität evaluiert. Das Teilvorhaben “Technische Entwicklung” an der TIB widmet sich der technischen Umsetzung des Projektes, insbesondere bezüglich des Harvestings von Forschungsinformationen auf Basis persistenter Identifikatoren sowie der teilautomatisierten Erstellung von Forschungsberichten mit der Open-Source-Software VIVO, siehe auch `https://duraspace.org/vivo/`.

Weitere Veröffentlichungen des Clusters

Neben der Forschung in den Projekten gibt es auch Arbeiten, die keinem Projekt zugeordnet sind, oder die aufbauend auf Ergebnissen aus Projekten erst nach der Projektlaufzeit erstellt wurden. Veröffentlichungen aus dieser nicht-projektgebundenen bzw. im Nachlauf von Projekten entstandenen Forschung werden hierunter aufgeführt.

Veröffentlichungen

1. Bruns, R., Dunkel, J., Offel, N.: Learning of complex event processing rules with genetic programming. Expert Systems with Applications **129**, 186–199 (2019). https://doi.org/10.1016/j.eswa.2019.04.007
2. Dötterl, J., Bruns, R., Dunkel, J., Ossowski, S.: Stream-based perception for agents on mobile devices. In: Lujak, M. (ed.) Agreement Technologies. pp. 73–87. Springer International Publishing, Cham (2019)
3. Dötterl, J., Bruns, R., Dunkel, J., Ossowski, S.: Stream-based perception for cognitive agents in mobile ecosystems. AI Communications **32**(4), 271–286 (2019)
4. Dötterl, J., Bruns, R., Dunkel, J., Ossowski, S.: On-time delivery in crowdshipping systems: An agent-based approach using streaming data. In: Proceedings of the 24th European Conference on Artificial Intelligence (ECAI2020). IOS Press, Santiago de Compostela, Spain (2020)
5. Dötterl, J., Bruns, R., Dunkel, J., Ossowski, S.: Evaluating crowdshipping systems with agent-based simulation. In: Agreement Technologies, AT 2020. Thessaloniki, Greece (to appear)
6. Heine, F., Kleiner, C., Oelsner, T.: Automated detection and monitoring of advanced data quality rules. In: Hartmann, S., Küng, J., Chakravarthy, S., Anderst-Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Database and Expert Systems Applications - 30th International Conference, DEXA 2019, Linz, Austria, August 26-29, 2019, Proceedings, Part I. Lecture Notes in Computer Science, vol. 11706, pp. 238–247. Springer (2019). https://doi.org/10.1007/978-3-030-27615-7_18, https://doi.org/10.1007/978-3-030-27615-7_18
7. Heine, F., Kleiner, C., Oelsner, T.: A DSL for automated data quality monitoring. In: Hartmann, S., Küng, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds.) Database and Expert Systems Applications - 31st International Conference, DEXA 2020, Bratislava, Slovakia, September 14-17, 2020, Proceedings, Part I. Lecture Notes in Computer Science, vol. 12391, pp. 89–105. Springer (2020). https://doi.org/10.1007/978-3-030-59003-1_6, https://doi.org/10.1007/978-3-030-59003-1_6
8. Knauf, F., Bruns, R.: A peek into the swarm: Analysis of the gravitational search algorithm and recommendations for parameter selection. In: Proceedings of the Genetic and Evolutionary Computation Conference. p. 30–38. GECCO '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3321707.3321774, https://doi.org/10.1145/3321707.3321774
9. Renners, L., Heine, F., Kleiner, C., Rodosek, G.D.: Adaptive and intelligible prioritization for network security incidents. In: 2019 International Conference on Cyber Security and Protection of Digital Services, Cyber Security 2018, Oxford, United Kingdom, June 3-4, 2019. pp. 1–8. IEEE (2019). https://doi.org/10.1109/CyberSecPODS.2019.8885208, https://doi.org/10.1109/CyberSecPODS.2019.8885208

10. Renners, L., Heine, F., Kleiner, C., Rodosek, G.D.: Concept and practical evaluation for adaptive and intelligible prioritization for network security incidents. Int. J. Cyber Situational Aware. **4**(1), 99–127 (2019). https://doi.org/10.22619/IJCSA.2019.100127, https://doi.org/10.22619/IJCSA.2019.100127

Ausgewählte Fachbeiträge

2

Der rechtliche Schutz von und durch künstliche Intelligenz geschaffener Kunstwerke

Fabian Schmieder

University of Applied Sciences and Arts Hannover, Germany
Faculty III, Dept. of Information and Communication
fabian.schmieder@hs-hannover.de

1 Einleitung

Eigentümer ist, wer das rechtliche Herrschaftsverhältnis über eine Sache ausübt. Nur der Eigentümer kann mit der Sache nach Belieben verfahren (§ 903 Satz 1 BGB), so lange er jedenfalls dadurch nicht gegen Gesetz oder Rechte Dritter verstößt.

Eigentum ist ein eingängiges Konzept, welches sogar für kleine Kinder intuitiv verstehbar ist, weil es die Zuordnung von Gegenständen (Sachen) zu Menschen in seinem Kern trägt und daraus ein ausschließliches Bestimmungsrecht ableitet. Verstärkt wird dies durch die Eigentumsvermutung aus § 1006 Abs. 1 BGB, nach der ein Dritter zunächst davon ausgehen darf, dass der Besitzer einer Sache auch dessen Eigentümer ist und eben über diese Sache bestimmen darf. Das gilt im Kleinen für Förmchen und Bagger im Sandkasten genauso wie im Großen für Autos, Häuser und ganze Fabrikanlagen.

Eigentümer können stets nur Menschen sein. Ein Tier ist zwar keine Sache (§ 90a Satz 1 BGB), gleichwohl finden die Vorschriften über Sachen in Ermangelung spezieller Vorschriften auf Tiere entsprechende Anwendung, so dass die Milch einer Kuh eben nicht im Eigentum der Kuh, sondern als tierisches Erzeugnis (§ 99 Abs. 1 BGB) dem Eigentümer der Kuh zusteht. Das Eigentumsrecht an der Milch wird danach von der Muttersache abgeleitet, mit der es einmal verbunden war.

Verlässt man den Bereich der Alltagsgegenstände und wendet den Blick auf Kulturgüter wie Gemälde, Fotografien oder gar Musik, stößt das Konzept vom Sacheigentum schnell an seine Grenzen. Zwar ist das Sacheigentum auch für Kunstwerke bedeutsam, ist aber naturgemäß auf vergegenständlichte Kunst beschränkt. Ein Gemälde steht – jedenfalls wenn es nicht herrenlos geworden ist – im Eigentum eines Menschen: Zunächst gehört es dem Künstler, welcher es dann verkaufen oder verschenken (und übereignen) kann, damit nun ein anderer bestimmen darf, wo das Gemälde hängt und wer dazu Zugang erlangt. Wem aber „gehört“ ein Musikstück? Das Sacheigentum kann in Ermangelung einer Sachqualität des Musikstücks bei der Beantwortung dieser Frage nicht weiterhelfen. Und auch beim Gemälde ist das Bestimmungsrecht des Eigentümers nur von begrenzter Hilfe, wenn von einem Gemälde eine Fotografie und damit eine Vervielfältigung angefertigt wird. Dadurch entstünde eine

Art abgeleitetes Werk, welches jedenfalls gedanklich mit dem Ursprungswerk verbunden bliebe, da es selben Inhalts wäre. Und auch wenn dieses abgeleitete Werk nicht „das Original“ wäre, so stellt sich doch wenigstens die Frage, wer diese Ableitungshandlung überhaupt vornehmen dürfte und wem die ggf. durch einen Verkauf des Vervielfältigungsstücks (oder Rechts) erlöste Erträgnisse zustünden. Eine Antwort dazu liefert das Urheberrecht, welches – losgelöst vom Sacheigentum (siehe etwa § 44 Abs. 1 UrhG) – die Rechte an persönlich, geistigen Schöpfungen (§ 2 Abs. 2 UrhG) regelt. Die urheberrechtliche Schöpfung ist das Ergebnis eines unmittelbaren und zielgerichteten geistigen Schaffens- bzw. Gestaltungsprozesses, welcher nur von Menschen vollzogen werden kann.[1] Das Urheberrecht ordnet die Verwertungsrechte (§§ 15 ff. UrhG) ausschließlich dem Schöpfer des Werkes (Urheber) unabhängig vom Sacheigentum (z.B. Leinwand, Farben, Notenpapier) zu. Danach ist jede in den Verwertungsrechten festgelegte Handlung (z.B. Vervielfältigen, Verbreiten, öffentliche Wiedergabe – etwa durch das zur Verfügung stellen zum Abruf im Internet) ausschließlich dem Urheber vorbehalten und bedarf dessen Erlaubnis (Lizenz, § 31 UrhG), wenn nicht ausnahmsweise der Gesetzgeber etwas anderes bestimmt hat (sog. Schrankenbestimmungen, §§ 44a – 63a UrhG). Der Urheber erlangt auf diese Weise umfassende rechtliche Möglichkeiten der Kontrolle seines Werks und kann durch die Einräumungen von Nutzungsrechten (z.B. Vervielfältigungsrechte) an Dritte – abseits des einmaligen Verkaufs und anschließender Übereignung des gegenständlichen Originals – sein Werk umfassend monetarisieren.

2 Schutz von Schöpfungen künstlicher Intelligenz de lege lata

Wenn nun nicht mehr Menschen Kunstwerke schaffen oder sich aber die Menschen bei der Schaffung von Kunstwerken in erheblicher Weise durch Computerprogramme unterstützten lassen, stellt sich die Frage, wer der Urheber dieser Werke und damit auch derjenige ist, der die Früchte aus diesen Werken durch die Einräumung von Nutzungsrechte ziehen darf. Die Frage hat durch aktuelle Kunstprojekte[2], die mittels Künstlicher Intelligenz (KI[3] geschaffen wurden, stark an Aktualität gewonnen, neu ist sie indes nicht. Schon Mitte der 60er Jahre haben sich u.a. Fromm und Fabiani der urheberrechtlichen Frage gewidmet, ob „Apparate“ schöpfen können.[4]

Jedenfalls in Kontinentaleuropa ist das Urheberrecht vom Schöpferprinzip bestimmt. Schöpfer eines Werkes ist derjenige, der aus einer urheberrechtlich nicht schutzfähigen Idee durch eine geistige Leistung ein Werk gewisser Originalität erschafft. damit im Sinne des Urheberrechts schöpft und durch diesen Realakt zum Urheber (§ 7 UrhG) mit all den einhergehenden Rechte wird. Ob KI diese geistige Leistung in Zukunft wird erbringen können, soll an dieser Stelle offenbleiben. Das maschinelle Lernen jedenfalls lässt den

[1] *Ahlberg* in: *Ahlberg/Götting*, BeckOK Urheberrecht, § 2 Rn. 52 f.
[2] Etwa das von KI geschaffene Werk „Next Rembrandt“ (`www.nextrembrandt.com`).
[3] Dazu ausführlich *Herberger*, NJW 2018, 2825 f.
[4] *Fromm*, GRUR 1964, 304; *Fabiani*, GRUR Ausl 1965, 422.

Verfasser eine solche, der menschlichen ebenbürtige Geistesleistung für möglich halten, auch wenn die aktuelle, von menschlicher Geistesleistung geschaffene, algorithmenbasierte KI dieses Niveau noch nicht erreicht haben dürfte.[5]

2.1 Rechtsfähigkeit von Maschinen und KI

Eine Zuordnung des Urheberrechts an eine Maschine scheitert im Falle der ausschließlichen Erstellung durch eine Maschine allerdings bereits an deren fehlender Rechtsfähigkeit.[6] Die vom EU-Parlament in Betracht gezogene „elektronische Person" ist (zu recht) alles andere als „beschlossene Sache" und ist wohl ohnehin eher für haftungsrechtliche Fragen gedacht, so dass sich an diesem Fakt auch für die Zukunft nichts ändern dürfte. Zumal sich Vorteile einer Zuordnung von Urheberrechten an eine „elektronische Person"[7] für die Rechtsordnung auch nicht wirklich aufdrängen, sondern vielmehr Folgeprobleme von erheblichem Ausmaß nach sich zögen: etwa die Frage, wie Lizenzverträge geschlossen werden sollen und was mit der finanziellen Gegenleistung geschehen sollte, so es eine gäbe.

2.2 KI und Trainingsdaten als Gegenstand von Schutzrechten

KI liegt stets die Programmierung eines Computers und damit ein Computerprogramm zu Grunde. Computerprogramme sind gemäß § 2 Abs. 1 Nr. 1 UrhG ausdrücklich als Sprachewerke urheberrechtlich geschützt und dürften stets die Anforderungen an eine persönliche, geistige Schöpfung (§ 2 Abs. 2 UrhG, sog. Schöpfungshöhe) erreichen. Das Urheberrecht an der Implementierung eines KI-Algorithmus in ein Computerprogramm liegt beim Programmierer, hat aber natürlich keine Auswirkung auf die von diesem Computerprogramm geschaffenen Kunstwerke, da dem Urheberrecht der derivative Schutz fremd ist. Anders also als bei den oben erwähnten Erzeugnissen (Kuh/Milch), ordnet das Urheberrecht das Erzeugnis eines Computerprogramms nicht dem Urheber des Computerprogramms zu.

Der mit dem Computerprogramm implementierte Algorithmus kann – wenn er jedenfalls über eine mathematische Methode hinausgeht (Art. 52 EPÜ, § 1 Abs. 3 PatG) und die übrigen Voraussetzungen vorliegen – als computer-implementierte Erfindung Patenschutz erlangen.[8]

Die neben dem Algorithmus in aller Regel erforderlichen Trainingsdaten[9] können dem Datenbankschutzrecht sui generis (§ 87a UrhG) unterliegen,

[5] So auch *Herberger*, NJW 2018, 2825, 2827.

[6] So auch *Borges*, NJW 2018, 977.

[7] Draft Report 2015/2103(INL) with recommendations to the Commission on Civil Law Rules on Robotics, S. 11, abrufbar unter `http://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf`.

[8] *Nägerl/Neuburger/Steinbach*, GRUR 2019, 336, 339; *Hetmank/Lauber-Rönsberg*, GRUR 2018, 574, 576.

[9] Für „Next Rembrandt" wurden ca. 150 GigaByte Trainingsdaten bestehend aus 346 hochaufgelösten Scans des Œuvres von Rembrandt verwendet.

wenn jedenfalls für ihre Herstellung eine erhebliche Investition erforderlich war.[10] Wobei an die Höhe der Investition keine erheblichen Anforderungen gestellt werden. Insbesondere sind nicht notwendigerweise finanzielle Mittel aufzuwenden; da der Investitionsbegriff weit zu verstehen ist[11], so dass auch die Aufwendung von menschlichen oder technischen Ressourcen oder Mitteln ausreichend ist[12].

2.3 autonome Schöpfungen von KI

Blickt man auf das Kunstwerk an sich, ist für die rechtliche Beurteilung beim Einsatz von KI zu differenzieren: Ist das urheberrechtlich zu beurteilende Kunstwerk vollständig autonom durch eine KI geschaffen worden, gibt es in Ermangelung einer schöpferischen Eigenleistung eines Menschen (§ 2 Abs. 2 UrhG) keinen Raum für ein urheberrechtliches Schutzrecht.[13]

Anders als das Patentrecht kennt das Urheberecht auch keinen abgeleiteten Schutz, der es ermöglichen würde, das von der KI geschaffene Kunstwerk dem Schutzrechtsinhaber an der KI zuzuordnen. Für solche, vollständig autonom geschaffenen Kunstwerke bleibt es beim eingangs dargestellten Schutz über das Sacheigentum. Das Eigentumsrecht greift sinnvoll freilich überhaupt nur für gegenständliche Werke und bietet dann auch nur einen sehr eingeschränkten Schutz über die Bestimmung des Ausstellungsortes und damit die Bestimmung über den Zugang zum Kunstwerk.[14]

In Ermangelung eines (urheberrechtlichen) Ausschließlichkeitsrechts[15] zugunsten einer Person besteht jedenfalls kein Raum für die Einräumung von Nutzungsrechten oder die Ausübung von Verbotsrechten. Die Anfertigung von Vervielfältigungsstücken wie Fotografien von einem Kunstwerk stellt jedenfalls keine Eigentumsbeeinträchtigung im Sinne des § 1004 Abs. 1 BGB dar, da durch die Vervielfältigungshandlung die Sachsubstanz des gegenständlichen Originals nicht beeinträchtigt wird.[16] Es ist in diesem Zusammenhang auch unerheblich, ob mit den Vervielfältigungen ein kommerzieller Zweck verfolgt wird.[17] Solche Kunstwerke dürften ohne weiteres von Dritten vervielfältigt, verbreitet und öffentlich wiedergegeben werden, insbesondere über das Internet zugänglich gemacht werden. Sie sind gemeinfrei.

---

[10] *Hetmank/Lauber-Rönsberg*, GRUR 2018, 574, 578.
[11] BGH GRUR 2011, 724 Rn. 18.
[12] Siehe dazu ausführlich *Vogel* in: *Schricker/Loewenheim*, § 87a, Rn. 40.
[13] So auch *Hetmank/Lauber-Rönsberg*, GRUR 2018, 574, 577 aE.; *Borges*, NJW 2018, 977, 978; *Schulze* in: *Dreier/Schulze*, UrhG, § 7 Rn. 2.
[14] Dazu weiterführend *Bullinger* in: *Wandtke/Bullinger*, UrhG, § 2 Rn. 164.
[15] Ein solches lässt sich auch nicht aus dem Sacheigentum am Kunstwerk ableiten, wohl aber aus dem Grundeigentum selbst und dem Fruchtziehungsrecht des Grundeigentümers: BGH GRUR 2011, 323, 324 Rn. 15.
[16] *Bullinger* in: *Wandtke/Bullinger*, UrhG, § 2 Rn. 164.
[17] BGH JZ 1998, 1120, 1121.

2.4 Schöpfen mit KI

Setzt der Künstler hingegen in seinem Schaffensprozess KI als Werkzeug ein, kommt er selbst noch als Träger von Rechten in Betracht, da es jedenfalls möglich ist, dass er noch eine eigenschöpferische Leistung erbringt.[18] Voraussetzung für das Entstehen urheberrechtlichen Schutzes zu Gunsten des Künstlers ist mit Blick auf § 2 Abs. 2 UrhG nämlich stets eine persönlich, geistige Schöpfung desselben. Die Schöpfung muss dabei das Ergebnis eines unmittelbaren und zielgerichteten geistigen Schaffens- bzw. Gestaltungsprozesses des Künstlers sein.[19] An einer solchen Unmittelbarkeit fehlt es jedenfalls, wenn das Ergebnis ausschließlich von einem Zufallsgenerator erstellt wurde.[20] Hat der Künstler hingegen die KI durch zielgerichtetes Einstellen von Parametern als Hilfsmittel für die Gestaltung seines Werkes genutzt und sind diese Einstellungen prägend für das Werk, liegt ein Schöpfung im Sinne des § 2 Abs. 2 UrhG und damit ein urheberrechtlich geschütztes Werk vor.[21] Letzteres dürfte vor allem für den Einsatz von KI-basierten Effekten (z.B. das gesichtsbasierte Verflüssigen von Adobe Photoshop) oder aber auch für Auswahl der Medien und die Einstellung der Parameter im Rahmen von Neuronal-Style-Transfers der Fall sein. Dies wird freilich in jedem Einzelfall zu beurteilen sein, wobei es durchaus Stimmen in der Literatur gibt, welche eine bloße Selektionsentscheidung (etwa von Ausgangsmaterial für die Bearbeitung durch eine KI) nicht für die Gewährung des urheberrechtlichen Schutzes als ausreichend erachten.[22] Mit Blick auf die – insbesondere bei bildender Kunst – geringen Anforderungen an die Schöpfungshöhe, spricht jedoch viel dafür, dass dies in der Regel ausreichen dürfte und nur bei trivialen Selektionsentscheidungen die nötige Schöpfungshöhe nicht erreicht wird.

3 Schutz von Schöpfungen künstlicher Intelligenz de lege ferenda

Wenn es jedenfalls für vollständig autonom geschaffene Werke keinen urheberechtlichen Schutz gibt, drängt sich die Frage auf, ob es nicht eines solchen, jedenfalls aber urheberrechtsähnlichen[23] Schutzes bedarf, wer Inhaber eines solchen Schutzrechtes sein soll und unter welche Voraussetzungen es entstehen soll.

Wegen der Bindung des Urheberrechts an die persönliche, geistige Schöpfung eines Menschen scheidet ein solcher originär urheberrechtlicher

[18] So für den Einsatz von „Kompositionsautomaten und Apparaten klanglicher oder kinetischer Effekte“ etwa Ahlberg in: *Ahlberg/Götting*, BeckOK Urheberrecht, § 2 Rn. 55.
[19] *Ahlberg* in: *Ahlberg/Götting*, BeckOK Urheberrecht, § 2 Rn. 54.
[20] Ebd.
[21] *Bullinger* in: *Wandtke/Bullinger*, UrhG, § 2 Rn. 16; *Loewenheim* in: *Schricker/Loewenheim*, UrhG, § 2 Rn. 13.
[22] *Hetmank/Lauber-Rönsberg*, GRUR 2018, 574, 577.
[23] Zu patentrechtlichen Überlegungen siehe bei *Hetmank/Lauber-Rönsberg*, GRUR 2018, 574, 580.

Schutz a priori aus. Allerdings kennt das Urheberrechtsgesetz mit den Leistungsschutzrechten und dem Datenbankschutzrecht sui generis auch Schutzrechte ohne die Notwendigkeit einer solchen Schöpfungsleistung. Bevor man allerdings an ein solches Leistungsschutzrecht oder Schutzrecht sui generis denkt, stellt sich die Frage nach dem Schutzbedürfnis. Da schöpferische Leistungen die sich der KI lediglich als Werkzeug bedienen – wie dargestellt – noch durch das Urheberrecht geschützt werden, bleibt eigentlich nur noch der Investitionsschutz als Argument für ein neues Schutzrecht. Die KI selber wird – wie dargestellt – bereits durch bestehende Schutzrechte geschützt, so dass für ein neues Schutzrecht nur das Produkt einer KI bliebe. Es mag im Moment schwer vorstellbar, vor allem aber schwer prognostizierbar sein, welche schützenswerte Investitionen für das Produkt einer KI in Zukunft zu tätigen sein wird. Auszuschließen ist eine solche Entwicklung freilich nicht. Es erscheint in diesem Zusammenhang jedenfalls überlegenswert, Produkte die von einer KI unter Aufwendung einer erheblichen Investition geschaffen wurden mit einem Schutzrecht sui generis in Analogie zum Schutzrechtrecht des Datenbankhersteller zu versehen. Ein denkbarer praktischer Anwendungsfall könnte etwa das Produkt einer KI sein, welches eine besonders langwierige und rechenintensive Datenverarbeitung erforderte oder nur mittels einer erheblichen Investition in Trainingsdaten herzustellen war. Gerade der letzte Fall dürfte in der Praxis zukünftig Relevanz erlangen. Zwar sind die Trainingsdaten im Falle einer erheblichen Investition als Datenbank geschützt (§ 87a UrhG), die mittels dieser Daten hergestellten Produkte hingegen nicht, was möglicherweise dazu führt, dass Investitionen in die Erstellung hochwertiger Trainingsdaten geringer ausfallen oder ausbleiben, weil es in Ermangelung des Schutzes daraus mittels KI hergestellter Produkte keinen attraktiven Markt für diese Daten gibt. Ein solches Schutzrecht sui generis müsste analog dem Datenbankherstellerrecht demjenigen zugeordnet werden, der die Investition getätigt hat.

4 Fazit

Das bestehende Urheberrechtsregime bietet bereits für weite Teile des Einsatzes von KI hinreichenden Schutz. Insbesondere bestehen beim Einsatz von KI als Werkzeug im Schaffensprozess keine gravierenden Schutzlücken. KI ist selbst als Computerprogramm urheberrechtlich geschützt. Die Trainingsdaten der KI sind – eine entsprechende Investition vorausgesetzt – ihrerseits als Datenbanken geschützt. Produkte, die völlig autonom von einer KI geschaffen wurden, sind allerdings derzeit de lege lata ohne immaterialgüterrechtlichen Schutz. Diesen in Analogie zum Datenbankherstellerrecht zu schaffen, scheint daher jedenfalls unter Investitionsschutzgesichtspunkten überlegenswert.

Literatur

1. *Ahlberg* in: *Ahlberg/Götting*, BeckOK Urheberrecht, 29. Edition, München, 2020.

2. *Borges*, Rechtliche Rahmenbedingungen für autonome Systeme, NJW 2018, 977.
3. *Europäisches Parlament*, Draft Report 2015/2103(INL) with recommendations to the Commission on Civil Law Rules on Robotics, abrufbar unter `http://www.europarl.europa.eu/doceo/document/JURI-PR-582443_EN.pdf`.
4. *Fabiani*, Sind Apparate geistige Schöpfer?, GRUR Ausl 1965, 422.
5. *Fromm*, Der Apparat als geistiger Schöpfer, GRUR 1964, 304.
6. *Herberger*, „Künstliche Intelligenz“ und Recht, NJW 2018, 2825
7. *Hetmank/Lauber-Rönsberg*, Künstliche Intelligenz – Herausforderungen für das Immaterialgüterrecht, GRUR 2018, 574.
8. *Nägerl/Neuburger/Steinbach*, Künstliche Intelligenz: Paradigmenwechsel im Patentsystem, GRUR 2019, 336, 339.
9. *Schricker/Loewenheim*, Urheberrecht (zitiert: UrhG), 5. Auflage, 2017.
10. *Wandtke/Bullinger*, Praxiskommentar zum Urheberrecht (zitiert: UrhG), 5. Auflage, München, 2019.

Standard and Advanced Automated Data Quality Monitoring*

Felix Heine[1], Carsten Kleiner[1], and Thomas Oelsner[2]

[1] University of Applied Sciences and Arts Hannover, Germany
[2] mVISE AG, Hamburg, Germany
{felix.heine,carsten.kleiner}@hs-hannover.de,
thomas.oelsner@mvise.de

Abstract. Data is getting more and more ubiquitous while its importance rises. The quality and outcome of business decisions is directly related to the accuracy of data used in predictions. Thus, a high data quality in database systems being used for business decisions is of high importance. Otherwise bad consequences in the form of commercial loss or even legal implications loom.
In this paper we describe the IQM4HD system for automating data quality monitoring. The system uses a domain specific language (DSL) called *RADAR* for data quality rules, that fulfills our main requirements: reusability of check logic, separation of concerns for different user groups, support for heterogeneous data as well as advanced data quality rules like time series-based rules. Also, it allows to automatically suggest potential rules based on profiling historic data. Furthermore, we show initial optimization approaches for the execution of rules on large data sets. Finally, the GUI used to interact with the system is presented.
All in all the system presents a novel approach for a flexible and powerful management of data quality in practical applications while meeting the needs of actual data quality managers in being pragmatic and efficient.

Kurzfassung. Daten werden zunehmend in immer größerer Breite erhoben und ihre Bedeutung steigt mindestens im gleichen Maße. Qualität und Konsequenzen von geschäftlichen Entscheidungen kann man in direktem Zusammenhang zur Qualität der Daten sehen, auf deren Grundlage diese getroffen wurden. Daher ist eine hohe Qualität von Daten in Datenbanksystemen, die Grundlage von geschäftlichen Entscheidungen sind extrem wichtig. Anderenfalls sind negative Folgen bspw. durch wirtschaftliche Verluste oder sogar juristische Konsequenzen nicht unwahrscheinlich.
In diesem Bericht wird ein System zur automatisierten Überwachung von Datenqualität vorgestellt, das im Projekt Intelligentes Qualitäts-Monitoring für heterogene Daten (IQM4HD) entwickelt wurde. Das System nutzt eine domänenspezifische Sprache (DSL), um Datenqualitätsregeln zu beschreiben, die unsere Hauptanforderungen erfüllen können: Wiederverwendbarkeit der Prüflogik, Trennung der Belange unterschiedlicher Arten von Systemnutzern, Unterstützung heterogener Datenquellen sowie die Möglichkeit, auch komplexe Regeln wie bspw.

* The project IQM4HD has been funded by the German Federal Ministry of Education and Research under grant no. 01IS15053A. We would also like to thank our partners SHS Viveon/mVise for implementing the prototype and CTS Eventim for providing important requirements and reviewing practical applicability of the prototype and concepts.

Zeitreihenregeln zu formulieren. Zudem ermöglicht es den Prozess durch das Vorschlagen möglicher Regelkandidaten auf Basis einer automatisierten Analyse historischer Daten. Es werden auch Optimierungsmöglichkeiten für eine effiziente Ausführung großer Mengen von Regeln sowie eine graphische Oberfläche für die Interaktion mit dem System dargestellt.
Insgesamt bietet das System einen innovativen Ansatz für ein flexibles und mächtiges Überwachen der Datenqualität in praktischen Einsatzszenarien, das die tägliche Arbeit eines Datenqualitätsmanagers effizient und leistungsstark unterstützt.

Keywords: Data Quality, Domain Specific Language, Data Quality Monitoring, Rule Based Data Quality, Data Heterogeneity

1 Introduction

Data is the core of many modern businesses. Data is typically stored in different distributed databases and business decisions rely on an integrated view on this data. In order to take right decisions it is important that data used is of high quality. While there are many different definitions of data quality in the literature (e.g. [16] defines 16 dimensions), we focus, in line with practical requirements from our project partners, on the quality dimensions of accuracy, completeness, consistency and integrity. The *IQM4HD*[3] system and its DSL *RADAR*[4] presented in this paper is applicable both to classical data warehouses as well as to heterogeneous, distributed data sources including non-relational data. An important part of data quality monitoring is to check internal consistency of a database and its conformance with external datasets. These checks can be rather simple checks like NOT NULL, structural checks like referential integrity checks, or complex statistical checks like conformance to a distribution or time series model. Our DSL integrates all these kinds of checks.

In order to detect quality problems reliably, it is important to check the data regularly. An extreme solution would be to run checks as database constraints, effectively avoiding the insertion of wrong data. However, this approach has multiple drawbacks: 1) The check logic slows down all processes that modify the data. 2) The check logic must be implemented inside the database and thus depends on the capabilities of the database. Not every check can be implemented on every system, especially more complex checks like statistical checks. 3) These kinds of checks often rely on proprietary languages like stored procedure languages. 4) Implementing these checks in a modular way, in order to apply the same logic to different databases, is often impossible. 5) Some checks, esp. statistical checks, do not always indicate clear errors. They rather result in warnings indicating an unusual data distribution that should be checked. Thus always rejecting non-conformant data is not possible. 6) Checks that span multiple databases (e.g. a relational database and a document database) are impossible with this approach.

Thus, our approach is to run the checks from an external engine that has its own language to encode the checks, called **rules** in our system. For this, we designed the *RADAR* DSL. It has been developed in the context of the project *IQM4HD*, which not

[3] `http://iqm4hd.wp.hs-hannover.de/english.html`
[4] Rule Language for Automated Data Quality Assessment and Reporting

only targets the execution engine for *RADAR*, but also an approach to create and maintain the set of rules with low effort. This approach includes a profiling module that can analyze existing data to make suggestions for rules and a feedback system where the data steward can classify detected data quality problems and warnings (e.g. as real quality problems, exceptions, or new patterns). The feedback system can then use these classifications to suggest new or adapted rules. Note that automated data correction is not in the scope of our work.

The remaining paper is structured as follows. First, the main concepts and design decisions of the DSL *RADAR* along with illustrative examples are presented in Sect. 2. Our prototype implementation is described in the following Sect. 3, with a focus on approaches to efficient code execution. We also present initial performance results showing the effectiveness of the optimization in this section. In Sect. 4, we discuss the profiling component used to generate rule suggestions. The GUI to interact with *IQM4HD* is presented in Sect. 5. After that, Sect. 6 covers related work and compares it with the specific contributions of this paper. Finally, the paper closes with a conclusion and ideas for future work.

2 Data Quality Rule Language RADAR

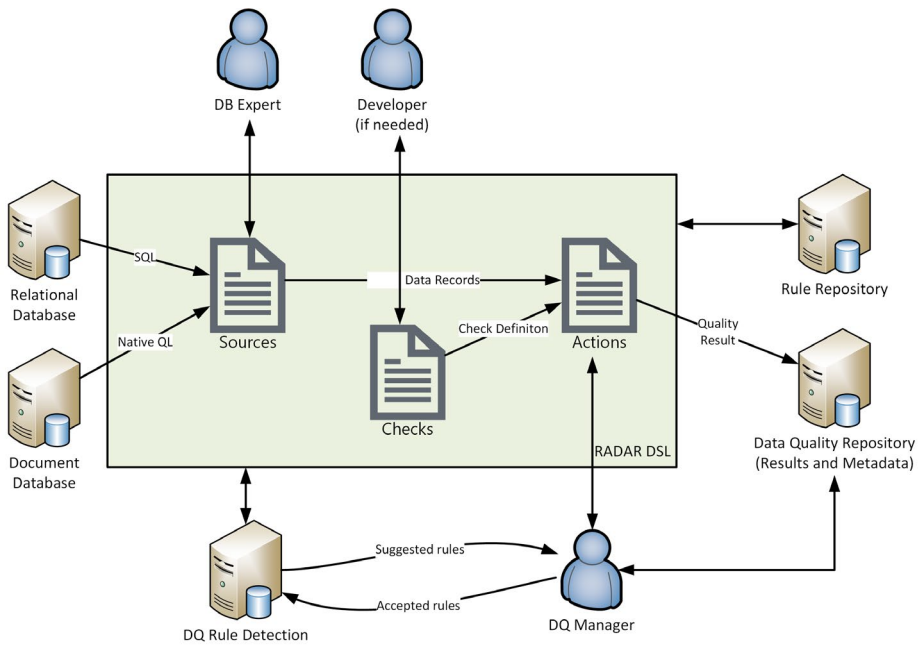


Fig. 1. Overview of DQ Monitoring with *RADAR*.

This chapter introduces the proposed language *RADAR* to specify data quality rules. For presenting details about the language we at first look at what information is needed to compose data quality rules, and secondly which types of users are present in our system. Information needed for quality rules includes (i) *what data should be checked*, (ii) *what quality aspect should the data be checked for* and lastly (iii) *what should be done with the result*. Regarding user types, we suggest a separation between the technical aspects of data retrieval and the subject-specific knowledge needed for writing coherent rules. The technical user has technical knowledge concerning the databases while the data quality manager has business knowledge about the data's meaning. Whereas the technical user just prepares the data for the data quality manager, the latter will - with his business knowledge - compose data quality rules or adjust automatically generated rules. The three aformentioned aspects are reflected in the quality rules in *RADAR* by three different components, namely Sources, Checks and Actions. To achieve this goal the framework is split into different parts. An overview of the system architecture is given in Fig. 1 (not depicted is the rule execution with its entailing processes). Note that a developer is only needed in case the Checks already provided with the system need to be extended. Data quality rules, approved by a data quality manager, in addition to their metadata and execution or result reports are persisted in a rule repository. These DQ rules are either manually created by a domain expert (based on a set of standard rules provided with the system or completely from scratch) or extracted from the databases with data profiling techniques. The automatically generated rules have to be confirmed, and may be adjusted, by a data quality manager in order to be carried over to the rule repository. Beside the generation of completely new rules, another part is the continuous adjustment of the existing rule base. Data grows with its business and thus thresholds - once valid - may shift, therefore, besides manual adjustment, also automated adjustment of existing rules is provided.

In addition to being able to perform standard data quality checks such as cardinality or format checks an important aspect is the specification of advanced data quality rules such as time series checks and multidimensional cube checks that can be monitored automatically. Moreover these rules cannot only be checked in relational data sources but also in other types of data source (e.g. document databases) as our quality language operates on an intermediary internal data model abstracting from the concrete source.

It is important to note though, that for implementation we used a formal specification of *RADAR* in terms of a EBNF grammar definition that has been used to generate a parser for the quality rules language using `antlr`. Details on this have been omitted here due to space constraints, but can be found in [15] for the interested reader. We try to present both simple rules that are easy to understand and more complex ones to document the potential of the language.

2.1 Introductory Example

As a first practical, purposely very simple (in order to focus on the DQ language aspect), example consider checking whether data in a certain value field of customer documents from a MongoDB contains a unique SSN and in another field a unique e-mail address. We have to define the data to be checked as Source (cf. Listing 1) in the MongoDB query language. The Check analyzing whether the attribute value in the SSN field is

```
SOURCE Customer TYPE LIST QUERY ROLES (_id: IDENTIFIER):
NATIVE
    db.customer.aggregate( [ {
        $project : { "name" : 1, "dob" : 1,
                     "email" : 1, "ssn" : 1 } } ] )
END
```

Listing 1. DSL code for a Customer Source from a Document DB with Projection.

```
CHECK UniqueCheck ON LIST src(val):
  RETURN
    SELECT val ROLE value, COUNT(*) ROLE value
    FROM src
    GROUP BY val
    HAVING COUNT(*) > 1;
END
```

Listing 2. DSL code for a unique check.

```
ACTION UniqueCheckCustomerSsnMail:
  EXECUTE UniqueCheck ON Customer(ssn: val)
  RESULT IN Error
END
```

Listing 3. DSL code for checking MongoDB customer attributes for uniqueness.

unique over all rows in the data source is displayed in Listing 2. Note that while the code inside the check looks similar to an SQL statement, it is not implemented in SQL, but rather in our own specification language. An important difference is that our language operates on lists of records in our internal data model and not on relations in a relational database. Thus, `src` in this case is a set of records from a document database as shown in Listing 1. The implementation counts the number of occurrences of each attribute value by a grouping function, detecting non-unique values by having more than a single occurrence. This type of Check is shown for completeness purposes only; it would be provided with the product and not be implemented by the DQ manager or a software developer in his enterprise. Finally, Listing 3 shows how the source data is assigned to the Check in an Action. The same Check could be used for a different attribute (e.g. `{Customer(email)}`) in another Action.

2.2 Internal Data Model

The basic data type in our DSL is a **record**, which is similar to a relational tuple, however, with one important distinction. Each record carries so-called *roles* instead of attribute names. Each attribute can have multiple roles, and each role can be assigned to multiple attributes. As an example, the following illustrates a customer record:

group identifier	id identifier	firstname name	lastname name	birthdate	telephone
A	12345	Jon	Doe	5/7/1980	111-222-33333

In this case, the first attribute has the roles “group” and “identifier”. However, the role “identifier” is assigned to the first two columns. When we access this role in our DSL, we effectively access the two-attribute record (’A’, 12345). This is a very flexible and convenient way to write code that can handle either attributes or compound data using a single identifier. The order of the attributes is relevant. The attributes of the record have basic data types, like strings, integers, doubles or booleans, and there is a NULL value. There is no explicit data type definition, each value carries its own type information. A non-existing role in a record is implicitly treated as value NULL.

The next higher level structure is a **list of records**, similar to a relation. The records in a single list, however, do not need to be homogeneous, each record can have a different set of roles and the data types of the same role could vary across the records.

This way, we can map both structures from relational databases as well as document-oriented databases to our internal model. Even other data sources are supported as long as they can be mapped to a list of records. The relational mapping is straightforward, while the mapping from documents flattens the documents. Sub-documents are mapped into a flat structure using role names with dot notation. Arrays are either ignored or an unwind operation[5] is used to map each array element to a new instance of the surrounding document together with this element, according to needs of the quality check logic.

Since we only need to map results of queries to our internal data model for use in quality checks and not the entire database schema, we do not have to consider more generic schema matching models here.

[5] `https://docs.mongodb.com/manual/reference/operator/aggregation/unwind`

```
SOURCE ageHistogram TYPE LIST CONST ROLES (from, to, perc):
    [ 0, 18, .05], [18, 25, .20], [25, 35, .15],
    [35, 50, .25], [50, 65, .19], [65, 80, .10],
    [80, 95, .05], [95, NULL, .01]
END
```

Listing 4. DSL code for a reference histogram.

```
SOURCE Customer TYPE LIST QUERY ROLES (id: IDENTIFIER):
DATABASE mondrian NATIVE
    SELECT id, name, date_part('year', age(dob)) AS age, email
    FROM customer
END
```

Listing 5. DSL code for a relational customer Source.

2.3 Sources

Sources are the interface to data from databases or other reference values a domain expert may work with when composing data quality rules. There are two main types of Sources. On one hand, there is the *domain mapping layer*, which maps database structures to Sources, like the `Customer` Source (Listing 5), referring to a customer table in a relational database system. On the other hand, there are *constant values* like simple or complex values, or patterns that contain the parameters for statistical models. A simple example for this kind of Source could be an email pattern used to check customers' emails for invalid entries. A statistical parameter set could be a histogram containing a reference distribution to check the age distribution of customers against, as shown in Listing 4.

The domain mapping layer has been designed to create an interface between the technical and the domain-specific side of data quality rules. The goal is to provide an easy way for domain experts to work with the underlying data without having to know details about where and how the original data is stored. Since non-constant Sources are specified in the database's own language, it is an assignment for technical users to prepare Sources for domain experts. The customer example in Listing 5 shows a generic SQL statement selecting the `id`, `name`, `dob` (date of birth), converted to an age, and `email` from the `customer` relation. Additional meta information is added by declaring the field `id` to be an identifier. Having the source layer as an intermediate layer between rules and the source database allows a very flexible way to define data access for any system using the system's native query language and to map structures to the domain layer accessed in the Checks of the DSL.

Beside the simple selection of e.g. relations and fields, it is also possible to use the full power of the data source's query language. This may range from simple pre-filtering of data for use in quality rules up to assembling data in a way that is best suited for the future quality checks, by e.g. using joins in relational sources. Data filtering may be beneficial for either legal reasons (i.e. the DQ manager is only allowed to access data

```
CHECK NullCheck ON value:
    RETURN value IS NULL;
END
```

Listing 6. DSL code for a NULL Check.

```
CHECK QuantityCheck ON LIST list1 WITH LIST list2:
    cnt1 := SELECT COUNT(*) FROM list1;
    cnt2 := SELECT COUNT(*) FROM list2;
    RETURN cnt1 != cnt2;
END
```

Listing 7. DSL code for a quantity Check.

for his country) or practical reasons, e.g. quality rules that are varying over countries due to different formats. Data assembly may be beneficial to create a quality specific view on the source data in an efficient way in the source system. However, a Source should be defined as simple as possible and should not contain calculations related to data quality checks but be constrained to retrieving data for a future quality check only.

2.4 Checks

Checks are the functional skeletons of data quality rules, they can be compared to functions/methods from general purpose programming languages. A Check is basically a method for data quality control. They operate on data in the internal data model (cf. 2.2) and are either already implemented in the system core such as the examples presented below or can be implemented by a developer using the constructs of the Check part of *RADAR*. Note that while the simple Check examples below at first sight seem similar to database constraints, there are two important differences: for one, the Checks can be used on arbitrary Sources, including data sources that do not provide constraints such as many NoSQL systems. Secondly, while constraints will typically be enforced on database operation level (which is not always desirable, esp. in DWH scenarios) our Checks provide a "softer" view on the current state of a quality aspect without direct influence on operations. Checks always include `RETURN` statements, which yield detected quality problems. Thus the overall result is a list of potential quality problems that will later be evaluated by the Action.

Listing 6 shows one of the most basic Checks, a check for `NULL` values. The code in Listing 7 compares two lists and checks whether they have the same amount of entries. Note that the lists may originate from completely different data sources as assigned in the Actions below. Finally, Listing 8 show a check for distribution matches. The `dist` parameter is a reference distribution, that defines bins for the values and a probability p for each bin. The age histogram in Listing 4 is an example. In the check code, first the number of data items is counted. Then, each value `val` is translated to a bucket name (`bin`) using the build-in function `val2bin` according to the buckets defined in the reference distribution `dist`. In the next step, a histogram is constructed and stored

```
CHECK DistChk ON LIST data(val) WITH dist(from, to, bin, perc):
    cnt := SELECT COUNT(*) FROM data;
    bindata := SELECT val2bin(val, dist) ROLE val FROM data;
    hist := SELECT bin, perc, COUNT(*) ROLE acnt
            FROM bindata RIGHT JOIN dist ON val = bin
            GROUP BY bin, perc;
    binscore := SELECT (acnt - cnt*perc) * (acnt - cnt*perc)
                      / (cnt*perc) ROLE score FROM hist;
    RETURN SELECT SUM(score) ROLE score FROM binscore;
END
```

Listing 8. DSL code for a distribution Check.

in the local `LIST`-type variable `hist`. Thereafter the score for each bucket is computed by $(O_i - E_i)^2/(O_i - E_i)$. In this case, the observed count O_i is the bucket count from the previous statement, and the expected count E_i is calculated using the size of the relation times the expected fraction from the reference distribution. The bucket scores are finally aggregated into a χ^2 score. This part shows the SQL-like syntax of *RADAR* for defining the Checks which works nicely with the lists of records internal data model produced by the Sources.

Since Checks are just archetypes of data quality rules, they provide parameters to utilize them for many different concrete usecases. There are two kinds of parameters for Checks: the first parameter provides the information, which data has to be checked for quality issues. The second type of parameters are control parameters, which are used to aid in checking. For example the `QuantityCheck` (Listing 7) expects two parameters, the first being a list of values, e.g. the content of a customer table from a Data Warehouse (DWH), whose quantity should be checked against the amount of values of the second parameter, another list of values (e.g. the content of the original customer table before inserted into the DWH). These parameters are divided via the keywords `ON` for "data-to-be-checked" and `WITH` for control parameters.

Thus the definition of the Checks is generic in two ways. First, the same Check can be applied to many different data sets. Second, the same Check can be used with different control parameters. The connection between Sources and Checks will be defined in the Actions explained below. Hence there is no reference to a specific age distribution in the `DistChk`. The exact same Check can be used for checking against an age reference distribution as well as a salary reference distribution depending on which parameter is passed to the Check as `WITH` parameter later. This significantly reduces the number of Checks required in a data quality monitoring system. Most basic and some complex analysis Checks are predefined in a rule catalog provided with the IQM4HD prototype. The goal is to provide a substantial foundation in order to cover most needs. Additional Checks can be implemented based on the provided language by developers as needed.

To cover all needs without creating a too complex language the expressiveness is three-tiered. First, there are boolean and mathematical expressions described directly by the language in combination with variable usage, `IF` statements and `FOR` loops as known from normal programming languages. Second, there is an SQL-like language to

```
ACTION NullCheckCustomerName:
    EXECUTE NullCheck ON EACH Customer(name) RESULT IN ERROR
END
```

Listing 9. DSL code for NullCheckCustomerName Action.

process data sets using typical operations like projection, selection, join and grouping as shown in Listing 8 already. Third, there are built-in methods to hide complexity, like a method to check multidimensional cube data.

Accessing Sources within the SQL-like statements in the Checks works straightforward for lists of homogeneous records as internal representation of list Sources originating, e.g. from relational or document databases. The name of a Source can be used in the `FROM` part to reference it. Individual columns within such Sources are identified by their attribute names. Attributes are assumed to be of non-complex data types. If a referenced attribute name is not present in any of the list elements (e.g. in the case of heterogeneous documents), this element is excluded from consideration in this Check.

For attributes of complex data types (e.g. sub-documents from a document database) the classical dot notation may be used to access key-value pairs within such an attribute. As documents may be nested hierarchically, path expressions separated by dots may be used to navigate into the hierarchy. In most cases the value specification using dot notation will rather be used in the Action part, because the Checks should be generic and operate on roles that are assigned in the Action part.

In order to present potential data quality issues to the DQ manager later, the `RETURN` statement is used. For proper storage and evaluation of the detected quality problems, a unified format to identify problems is necessary. Thus we classify quality problems according to the unit of data that is related to the problem: 1) Problems that are related to individual records like `NULL` values; 2) Problems that are related to values that relate to multiple records, like uniqueness problems; 3) Problems that are related to a whole data source, like skewed distributions. All returned values are bundled and will later be evaluated by the result part of the Action invoking this Check.

2.5 Actions

Actions are responsible for two things. Firstly, they connect Sources and Checks resulting in data quality rules. Secondly, they deal with the result and therefore conclude data quality rules. Actions were designed with natural language in mind, so that a rule composed in the DSL reads as a self-explaining sentence. Following are a few examples of Actions which cover different kinds of Checks.

Listing 9 shows an Action testing customer names for `NULL` values. The `NullCheck` is designed for single values, because there is no `LIST` keyword in front of the parameter in the Check. Thus the action includes the `EACH` keyword in order to call the check for each item of the data source individually. Listing 10 is an example of a check of the result of an Extract-Transform-Load (ETL) process. It checks whether the amount of tuples in the DWH table fits the original table.

```
ACTION DWHQuantityCheckCustomer:
    EXECUTE QuantityCheck ON DWH_Customer WITH Customer
    RESULT IN ERROR
END
```

Listing 10. DSL code for DWHQuantityCheckCustomer Action.

```
ACTION AgeDistributionCheckCustomer:
    EXECUTE DistributionCheck ON Customer(age) WITH ageHistogram
    RESULT IN WARNING ABOVE 14.07 AND ERROR ABOVE 18.48
END
```

Listing 11. DSL code for AgeDistributionCheckCustomer Action.

A more complex example is Listing 11. It uses the customers dates of birth and a reference age distribution to compare the current customer ages with a saved reference distribution. The result is a score computed by a $\mathcal{X}^2$ test, which is evaluated by the `RESULT IN` part of the Action. The chosen threshold values 14.07 and 18.48 will issue a warning when the 95% value of the $\mathcal{X}^2$ test is not met and an error if even the 99% value of the test is not achieved. These concrete values are typically computed by the profiling component of the *IQM4HD* system.

The notation to access parts of a source definition follows the same logic as described for the Checks previously. Key-value pairs in nested documents within a document can be accessed by using the Source name together with a path expression. Listing 12 shows a `NULL` Check on the `name` key inside the `street` sub-document of the `address` sub-document of the Source `CustomerAddress`.

2.6 Advanced Data Quality Rules

This section presents an example of a value rule, showing how more complex data quality rules are specified in *RADAR*. The basic idea is to use time series modeling to find data quality problems in time-related data, i.e. the sales volume of a company. Such data normally behaves according to some time series model. Unusual values not matching the model could indicate quality problems, like missing data or duplicate data fed into the data warehouse. A basic type of time series models are SARIMA models.

```
ACTION NullCheckCustomerStreetName:
    EXECUTE NotNullCheck
    ON EACH CustomerAddress(address.street.name: value)
    RESULT IN ERROR
END
```

Listing 12. DSL code for NullCheckCustomerStreetName Action.

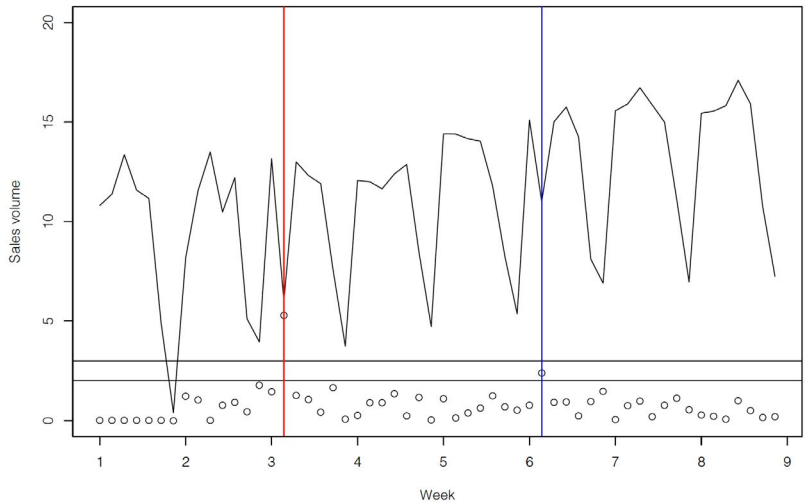


Fig. 2. Sales time series with daily ARIMA scores.

A time series in SARIMA models (see e.g. [14]) is described by multiple components. They are driven by a series of random components e_t with zero mean and equal variance σ^2. The AR component is a linear combination of the previous values of the time series and the current random component. The MA component is a moving average of the previous random components. The "I" stands for "Integrated" meaning that the differenced series is looked at. The seasonal part references back to the values in the previous season, for example $S = 7$ for the same day in the previous week. The seasonal part can also be described using AR, MA or I components. Overall, the order of an SARIMA model is described by $(p, d, q) \times (P, D, Q)_S$, where the first triple defines the order of the AR, I, and MA components in the non-seasonal part, and the second triple defines these orders for the seasonal component.

For example, the model we use for the time series in Fig. 2 is an $(1, 0, 0) \times (2, 1, 0)_7$ model, which means that we have an AR model that only uses the last value and an $S = 7$ weekly seasonal component that uses 2 AR components. Furthermore, the seasonal model is differentiated once as $D = 1$. We have no MA parts here. The equation for this model series is:

$$(1 - \text{ar}_1 B)(1 - \text{sar}_1 B^7 - \text{sar}_2 B^{14})(1 - B^7) y_t = e_t \tag{1}$$

Here, B^l is the backshift operator, meaning $B^l y_t = y_{t-l}$. The parameter values are $\text{ar}_1 = 0.6301$, $\text{sar}_1 = -0.5609$, and $\text{sar}_2 = -0.1503$. The variance of e_t is estimated to $\sigma^2 = 2.12$. Using this model we can calculate a forecast $\hat{y}_t$ by resolving (1) for y_t and setting $e_t := 0$. The forecast interval is calculated using the variance of e_t.

Our check now computes $\hat{y}_t$ for every time step t using the previous values. The absolute difference between the forecast and the actual value, normalized with σ, defines

```
CHECK ArimaCheck ON LIST ts(time, value)
                WITH LIST model(key, index, value):
    RETURN arima(ts, model);
END
```

Listing 13. DSL code for ARIMA check.

```
SOURCE TimeSeriesSales TYPE LIST QUERY
                       ROLES(time_id: IDENTIFIER):
NATIVE
    SELECT '' || week || '-' || day as time_id, sales_volume
    FROM time_series_sales
END
```

Listing 14. Source for time series sales data.

the score for each point:

$$\text{score}_t = \frac{|y_t - \hat{y}_t|}{\sigma}$$

For the forecast calculation of later points, we replace outliers where the score is larger than 2 with the forecasted value to avoid that a single error influences the forthcoming forecasts and produces consecutive errors.

As the check code uses a built-in method `arima` that internally calls a suitable standard function of the R statistics package, the code is rather simple, cf. Listing 13. Data to be checked is specified using a relational data source, see Listing 14.

The model itself is described in *RADAR* using a constant Source, that is not computed by executing an external query, but rather uses a fixed structure throughout. This constant Source lists the order of the model, the coefficients and the variance of e_t. An example matching the above model is shown in Listing 15. Since it is very difficult to find this model manually, especially for the DQ manager, we IQM4HD provices a profiling component described in section 4 to automatically fit a model using historical data.

Finally, an action puts it all together. It calls the ARIMA check using the model and the time series data source, and assigns the appropriate roles, as shown in Listing 16.

```
SOURCE ArimaModelSales TYPE LIST CONST ROLES (key,index,value):
    ["coef",1,0.6301], ["coef",2,-0.5609], ["coef",3,-0.1503],
    ["coefnames",1,"ar1"], ["coefnames",2,"sar1"],
    ["coefnames",3,"sar2"], ["var",1,2.12],
    ["arma",1,1], ["arma",2,0], ["arma",3,2], ["arma",4,0],
    ["arma",5,7], ["arma",6,0], ["arma",7,1]
END
```

Listing 15. Source for sales data ARIMA model.

```
ACTION ArimaCheckSales:
    EXECUTE ArimaCheck
    ON TimeSeriesSales(time_id: time, sales_volume: value)
    WITH ArimaModelSales RESULT IN ERROR ABOVE 3 WARNING ABOVE 2
END
```

Listing 16. Action for sales time series data quality rule.

Continuing the example Fig. 2, we see that the sales volume data has a weekly seasonal pattern. The circles at the bottom denote the computed scores; the upper horizontal line indicates the error threshold of 3, the lower horizontal line indicates the warning threshold of 2. Note that we don't have scores for the first week as we need historic data for forecasting. At time 3-2 (second day of week 3), there is an error, the value of the sales volume is too small for a non-weekend day. Due to the seasonal model, this error can be detected. On the other hand, at time 6-2 there is a warning as the value is rather low but not as exceptional compared to 3-2. Yet, the data steward may know that these values are due to some shops having been closed for a public holiday. On the other hand, it might also be due to an error in data provisioning and in that case the error and the warning should be observed. Thus, days 3-2 and 6-2 will be reported as suspicious. It would be impossible to detect both types without using a seasonal time series, as both values are within the normal range of values.

```
Issue: type=0, severity=ERROR, score=5.30,
  identifier={"value":"3-2","roles":["IDENTIFIER"]}
Issue: type=0, severity=WARNING, score=2.39,
  identifier={"value":"6-2","roles":["IDENTIFIER"]}
```

For the error this is true as the sales volume is only suspicious because it occurred on a second day of the week. It is still higher than volumes on last days of a week, so the weekly pattern is integral for detecting a potential error. Similarly, the warning is only issued as the difference when compared to the same day in the previous week is high. This data points in itself is close enough even to the mean value for that weekday. So again using a seasonal time series model is integral.

Additional to these time series checks, we have implemented a quality check that exploit the semantics of a multidimensional data cube. Such a cube is a set of aggregate cells over a base relation that has dimensional attributes and metric attributes [1]. Almost any cube has a time dimension.

Syntactically, the cube check is pretty similar to the time series check, with the difference, that the Action provides the Check with information as to which attributes contain the cube metric/fact value, cube dimensions and cube time dimension, respectively. The Check itself is rather simple, basically invoking a built-in function that hides the complexity and computing a score for the passed data cube by comparing it against a desired cube model. This model is specified as a constant data source similar to the one presented in Listing 15 for time series, just with different parameters. Based on the computed score the Action can issue warnings or errors, as before. Since the constant data sources needed to provide the models for the advanced quality checks tend to be

lengthy and difficult to discover, it is important to assist the DQ manager in specifying such sources. This is done by our profiling component as will be explained later.

3 Implementation / Optimization

In this section, we briefly touch upon the execution engine that we implemented to execute *RADAR* rules. Basically, the engine consists of a parser that analyzes the DSL code (Sources, Checks, and Actions), and then interprets the code by loading data from the sources and processing the Check logic upon the data. However, always fetching all data from the source and processing the logic internally is inefficient. Thus we focus on optimization approaches in the first sub-section and present an initial performance evaluation to show the gains that can be achieved using these approaches.

3.1 Optimization

An important goal of the optimization is flexibility with respect to the location of the execution of check logic. In principle, checks can be executed in the quality engine itself or moved to the source databases by rewriting the source query. On the one hand, the execution engine should be able to execute all logic on its own, in cases where the target database does not have enough capabilities, either because of resource constraints or because of a lack of features. On the other hand, if possible, logic should be moved to the database for more efficient execution due to co-location with the data.

To reach this goal, we distinguish two cases. First, we have checks for individual records, that are called in a loop for each item of the source database. In this case, we extract what we call *safe conditions* from the Check, i.e. conditions that need to be fulfilled for the Check to find quality problems. These conditions are then translated to filter conditions that are added to the native database query in the source database. This step needs individual code for each native query language and might fail, if the target dialect does not support the relevant conditions. However, if it succeeds, we avoid fetching large numbers of records that will not produce output for the quality check anyway. As a very simple example, consider the `NullCheckCustomerName` in Listing 9. From the referenced check, we extract the condition `value IS NULL`. The role `value` is then back-substituted to `name` and the original query is modified to

```
SELECT * FROM (... original query ...) WHERE name IS NULL
```

For checks that work on `LIST`s, we follow a different strategy. The first component of the strategy is deferred execution. This means, that list expressions are evaluated only symbolically, i.e. variables are substituted with their current values. However, the list expression itself is not evaluated until it is actually needed (i.e. because it is part of a return statement or part of a condition that determines control flow). This leads to larger expressions that are build up part by part.

As an example, consider the following lines from the distribution check:

```
1 bindata := SELECT val2bin(val, dist) ROLE val FROM data;
2 hist := SELECT bin, perc, COUNT(*) ROLE acnt
```

```
3          FROM bindata RIGHT JOIN dist ON val = bin
4          GROUP BY bin, perc;
5  binscore := SELECT (acnt - cnt*perc) * (acnt - cnt*perc)
6                   / (cnt*perc) ROLE score FROM hist;
7  RETURN SELECT SUM(score) ROLE score FROM binscore;
```

In the first line, `bindata` is assigned the list expression on the right side *without evaluation*. The literal `data` is replaced with a reference to the data source stemming from the Action that called the Check, including the native database query statement (i.e. an SQL statement for relational sources). So the value of the variable `bindata` is a list expression, not its actual value (i.e. not a list of records). In line two, again, `hist` is assigned a list expression, where the literal `bindata` is replaced with its value, leading to a more complex list expression with a sub-expression. This process is going on until the `RETURN` statement is reached. The expression becomes a huge integrated list expression containing all previous expressions as sub-expressions.

Once execution is requested, as it is the case for a `RETURN` statement, the whole expression is optimized quite similarly to normal database optimizers. Our optimizer tries to push as much logic towards the data sources, where the logic is added to the original database query (similar to the above modification for `NullCheckCustomerName`). However, as soon as different source databases are involved in an evaluation, no further pushing of operations to the Sources is possible and the quality engine has to take over evaluation from this point on. For example, a join of two sources from different databases has to be processed internally in the data quality engine.

3.2 Performance Evaluation

In this section, we look at the efficiency of our execution engine and especially at the effectiveness of our optimization approaches as performance will be a critical feature in large scale use cases. The goal is not an exhaustive evaluation of all types of rules, but rather to show the potential of our optimization strategies. For this, we use three different scenarios. First, we look at a simple NOT NULL check (NN), which is an example of a single record check. In this case, the optimizer modifies the source statement so that only critical records (i.e. NULL values) are selected. Second, we look at a referential integrity check (RI) which checks which records in one list do not reference valid records in the other list. This contains more complex list expressions that are translated to SQL by the optimizer in case of a relational data source. As a final, more complex, example of this list expression to SQL translation, we use the distribution check (Dist; see Sect. 2.4) that checks to which degree records in a source comply with a reference distribution.

For evaluation, we use a `person` (approx. 5,000,000 records) and a `plays` relation (approx. 8,400,000 records) from a movie database[6]. For each scenario, we look at different sizes of the data to check (1%, 10%, and 100% of the original data). As the runtime might be influenced by the number of errors in the first two scenarios, we also

[6] `ftp://ftp.fu-berlin.de/pub/misc/movies/database/`

	Errors	0%			1%			10%		
	Size	1%	10%	100%	1%	10%	100%	1%	10%	100%
NN	Std.	0.25	1.57	16.45	0.22	1.35	18.15	0.21	1.36	17.22
	Optim	0.06	0.05	0.33	0.04	0.06	0.48	0.05	0.23	2.85
	SQL	0.01	0.04	0.53	0.01	0.04	0.53	0.01	0.12	1.85
RI	Std.	0.66	5.44	106.38	0.56	5.62	142.90	0.67	11.98	458.45
	Optim	0.15	0.79	16.38	0.12	0.84	14.11	0.11	1.28	16.74
	SQL	0.11	0.76	11.24	0.10	0.87	10.75	0.12	1.37	11.01
Dist	Std.	0.82	6.40	74.01						
	Optim	1.07	7.80	73.04						
	SQL	1.23	7.97	70.21						

Table 1. Evaluation results (averages over 5 runs; time in sec.)

use different numbers of errors (no errors, 1% errors, 10% errors).[7] The distribution check calculates an overall score for the target relation taking each record into account, thus here we have no difference in effort related to the error ratio. As benchmark, we also list run times of SQL statements running directly on the database that perform the same check. This is only possible as long as all data originates from a single database and also sacrifices all other benefits of *RADAR* as explained in Section 2.

The results are summarized in Table 1. For the NN check, we see that the optimized version performs much better than the standard version. The standard version suffers from a large number of calls to the check (for each person record) and is almost independent of error ratio. The optimized version is a significant improvement that is the better the smaller the error ratio (since fewer data has to be transferred to the engine). Also, the scaling behavior looks good as the runtime is roughly proportional to the data size. Runtimes for the optimized version are within the same order of magnitude as the raw SQL statement (that does not include any DSL processing). As the optimized version can only filter correct records, the runtime becomes larger with larger error ratio.

For the RI check in the standard version the scaling is not as good, both in terms of data size and error ratio. We assume that this is mainly due to our join implementation in the engine, which is currently not very efficient and subject to future improvements. However, the optimizer works quite well for this scenario again, reaching runtimes in the same order of magnitude as the raw SQL benchmark and similar scaling behaviour. Also, as expected, the runtime is almost agnostic of the error ratio in this case.

Finally, the Dist scenario shows a different picture. Here, the scaling works well, however the optimized version does not yield any improvement. A comparison with the raw SQL statement shows, that this statement already consumes nearly all of the runtime so that no improvement is possible. A closer look revealed that this is due to the stored procedure we used to translate the `val2bin` method to SQL. A hand-crafted SQL version that avoids this procedure can be much faster, so we plan to modify our SQL generation to output this result in the future.

[7] The errors were introduced by updating values to NULL or to non-existing FK values for a random subset of the records.

Overall, we can conclude that the optimizer works very well in general and establishes runtimes close to raw SQL statements that implement the same logic. However, as we cannot always count on it in cases when the underlying database has limited query capabilities (NoSQL) or when data from multiple sources is combined, we will also continue to improve the DSL interpreter. Efficiency improvements w.r.t. the engine itself are more tied to improvements in small details (e.g. join implementation, implementation of the expression evaluation) and not to general architectural issues. Thus the prototype shows already promising results and underlines that the chosen approach works well. However, further improvements still remain an important issue for future work (cf. Section 7.2).

4 Automated Rule Detection

In order to ensure data quality, a huge set of rules is necessary to regularly test as much quality aspects and parts of the data set as possible. When writing rules, two challenges arise. First, the effort to write huge numbers of rules is high and thus a good coverage of the data with rules might not be achieved. Second, finding good parameters and thresholds for complex rule types by hand is extremely difficult. This holds true especially for rules that define a statistical model for the data and check the conformance of the current data with the model (see section 2.6). To simplify these tasks, we developed a profiling component that examines existing data and suggests rules and parameters for these rules. The data quality steward can then inspect these suggestions and decide, using additional business knowledge, which candidates should be switched to active.

We follow a modular approach and provide various profiling modules. Each module is focused on one type of profiling and outputs rules of a specific type. As a simple example, the NOT NULL profiling module counts the current number of NULL and NOT NULL entries in a column. Based on these statistics, the module suggests either a *hard* rule as shown in Listing 9, lines 1 to 3, or a *soft* NOT NULL rule, that states that no more than a certain fraction of the values may be NULL, or both. For the soft rule, the profiling module suggests reasonable parameter values. Another example is a profiling module for a time series rule, that has to estimate the parameters of the time series model.

The user interface (UI) to steer the profiling is two-fold. First, we have a basic profiling UI that targets the generation of huge amounts of basic rules. This UI shows the data sources and their attributes in a grid-like view and automatically suggests basic profiling modules that match the data type of the attributes, i.e. numeric range profiling for numeric columns, NULL profiling for every column, etc. The goal is to quickly generate simple rules for individual column values with a good coverage. The profiling functionality here is similar to commercial data profiling tools, but with the extension, that quality rule candidates are automatically generated based on the profiling results. The other UI, called advanced profiling, targets more complicated rules where the profiling needs more information to run. An example is the time series profiling, which needs to know which attribute of a data source contains the time and which contains the value of the time series. Here, running automatically over every combination would produce extreme load and generate lots of meaningless rules. Thus we assume the data

steward can steer the process better using business knowledge about data. For the basic profiling UI, we use one generic kind of dialog. For the advanced profiling UI, each module defines its own UI components and parameters that are needed to steer this module.

In the following, we present details of selected profiling modules in order to better show our approach and the way of thinking behind these modules. In general, every module might either output rules that are completely satisfied by the current data or rules that already declare some data as errors or warnings. For these rules, every module has a "max errors" parameter that steers how many errors are acceptable for a rule to ultimately be suggested.

4.1 NOT NULL profiling

In the rule catalog, we have two kinds of NOT NULL rules. The first is a hard rule that outputs an error when there is at least one NULL entry. The second is a soft variant that is based on statistical assumptions. The assumption is that each value has a certain probability p of being NULL, independently of the other values. Thus the random variable X that counts the number of NULL values follows a Binomial distribution with the parameters p and n, the current size of the data source. The soft NULL rule essentially performs a one-sided Binomial test, and the thresholds determine the significance level. However, we use 0.95 and 0.99 and $P(X < x)$ instead of 0.05 and 0.01 and $P(X \geq x)$ as our system uses the general convention that larger scores indicate more severe quality problems.

The NULL profiling module starts by counting the current number x of NULL values in the chosen attribute, and the overall size of the relation n. The module suggests a hard rule in case that the overall number of NULL values is smaller than the "max errors" parameter. A soft rule is suggested in case that at least one NULL value exists, using the estimate x/n for the parameter p. However, when this estimate for p is larger than 0.5, we don't suggest any rule, as in this case NULL values are more probable than NOT NULL values.

As an example, assume we have a column with 1000 entries where only 2 are NULL. In this case, it is possible that there is a business reason for these NULL values, or these values might be data quality problems. Accordingly, the profiling module will suggest a hard NOT NULL rule and output that this rule will generate two errors based on the current status, and it will additionally suggest a soft NOT NULL rule with $p = 0.002$. The steward then must make a decision between these two rules based on his domain knowledge. In case that a column does not have any NULL values, only the hard NOT NULL rule is suggested.

4.2 Range profiling

A range rule states that the value x of a column must be between two values l and u. The score is 0 when $l \leq x \leq u$, and it is the absolute deviation from the lower or upper value in case that it is smaller than l or larger than u. The profiling module for this rule examines the current column values. The first rule that is output is a rule that uses the minimum and maximum value occurring as l and u, respectively. This rule does not

return any error for the current data set. The second rule copes with outliers. For this, the lower and upper quantile (Q_1 and Q_3) of the values are computed and the typical outlier definition using the interquartile distance IQD $= Q_3 - Q_1$ is applied. Thus we define $l = Q_1 - 1.5 \cdot \text{IQD}$ and $u = Q_3 + 1.5 \cdot \text{IQD}$. Using these boundaries, we count the number of outliers and check whether the number is not larger than the "max outliers" parameter. When this is true, we suggest the alternative range rule.

4.3 Foreign key dependencies

Foreign key profiling works essentially exactly as NULL profiling, because we use the current number of violations of the foreign key as basic statistic. Based upon this statistic, the module chooses to suggest a hard and/or a soft rule to check the foreign key. However, foreign key profiling is an advanced method, because it would be too inefficient to try every possible attribute combination between any data sources. Thus, we expect the data steward to steer the profiling by defining foreign key candidates that should be checked.

4.4 Time series profiling

This module uses the R statistics package to estimate a time series model based on the data. For this, we need a data source with columns that carry the time identifier and the time series value. The whole time series is read and sent to R in order to retrieve an SARIMA model. The SARIMA model is converted into a constant data source (cf. Listing 15) and will be stored in the rule repository along with the rule in case the rule is accepted by the data steward.

Based upon the SARIMA model, the time series is checked as described in section 2.6. The value of every data point is compared to the 95% and 99% prediction interval. In case it is outside either interval, the point is declared to be a warning or an error, respectively. The warnings and errors are counted and displayed to help the steward to decide whether to accept or decline the suggestions. In the future, we also plan to integrate a graphical display of the time series, its model, and the detected warnings and errors in the UI component.

4.5 Cube profiling

To check the data of a cube, we profile how the data in the cells varies over time using the time dimension, and define a cube model based upon these statistics. Thus, a cube model is essentially a set of cell models.

During profiling, we need to choose the subset of cells for which models are computed, as the overall number of all possible cells is typically extremely large (growing exponentially in the number of dimensions). Our main criteria is the availability of enough data for a cell to be included in the model. First, the user chooses a time interval for the profiling (i.e. daily, weekly, monthly). Then the profiling process calculates a cube for every interval separately. In order to limit the computational power and memory requirements, we limit these cubes to Iceberg cubes [20] using a threshold parameter. This means that only cells where the metric value is above the threshold

are included in the cube. Then, we run through all time intervals and collect statistics about all cells that exist in every time interval. This ensures a focus on the higher-level aggregate cells that contain enough information for reliable models.

Currently, we use a Gaussian distribution to model each cell's value. Thus we use the sequence of metric values computed for the current cell for every time interval to estimate μ and σ. The overall cube model consists of all these values for the selected cells. The cube model is provided in the format of a constant data source of our quality rule language (similar to time series above) and can thus be easily used in quality checks, if the corresponding suggested rules are accepted by the data stewart. The final steps are also identical to the time series profiling: we use the model to calculate the current number of warnings and errors and suggest the rule together with this information.

5 Managing DQ Rules in Practice

In addition to the definition and implementation of the DQ language *RADAR*, it is important to consider the practical deployment of these results into practice in order to actually provide a DQ management system as outlined in Fig. 1. In large scale enterprise installations where the DQ language will be used to monitor data quality in large DWHs such as the application scenarios the company partners in the *IQM4HD* project have been using, the system is decomposed into micro services and these will be deployed as in other enterprise micro service architectures. However, for smaller company installations as well as in the research area, the overhead incurred by employing a micro service architecture is too big. For these type of applications a small GUI tool facilitating management, execution and result analysis of DQ rules in *RADAR* is beneficial. We have developed such a GUI tool in the scope of a student project and will briefly outline the main features of this tool in this section.

The primary use cases of the tool whose main window is displayed in Fig. 3 are as follows:

- Management of Sources, Checks and Actions organized in DQ projects
- Management of database execution environments for DQ rules
- Manual execution of quality checks on databases (both local and remote execution)
- Display of rule execution results with analysis capabilities
- Collecting and persisting expert feedback on detected quality rule violations

We will discuss some details on each of these use cases in the following paragraphs. Note that the profiling component for automated rule detection and suggestion has not yet been integrated into the GUI.

5.1 Artefact Management

The basic unit of management is a DQ project which consists of the different artifacts as displayed in Fig. 1. All information for such a DQ project can be stored in the file system or in a database system. The main artifacts Sources, Checks and Actions present in a DQ project are displayed in their respective tab on the left (cf. Fig. 3). Icons indicate

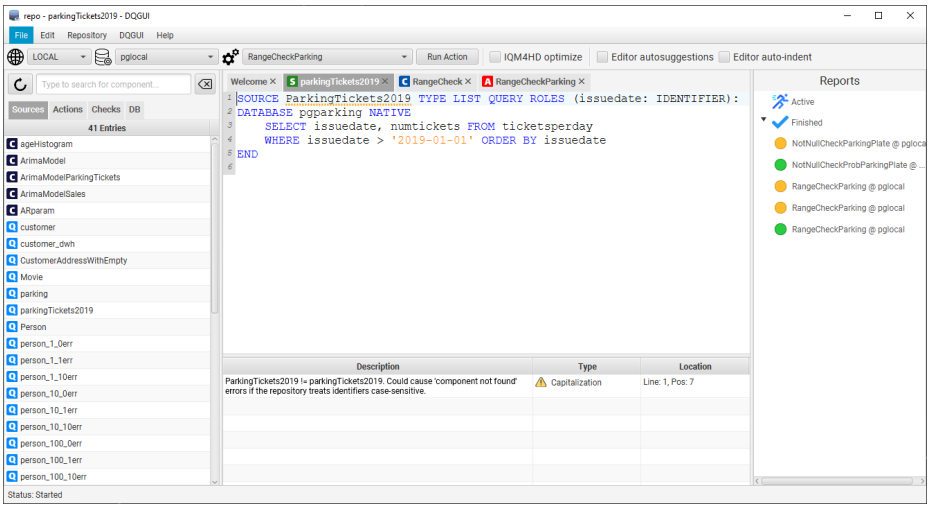


Fig. 3. Main window of DQ GUI

the subtype of artifact (e.g. constant vs. query source or system-provided vs. individual Check). Moreover, in the editor window in the middle the source code of the artifact can be edited. Editing is supported by syntax highlighting and auto completion based on the antlr grammar definition of the DQ language. The editing support is thus dynamic and automatically adjusts to future modifications of the language. Furthermore, below the code editor there is a window listing errors or warnings detected by the tool with regard to the current artifact in the editor. A search feature to find artifacts of a certain name is also provided as is auto navigation from one artifact to another references in the source editor window. Individual Checks can be added to the project and then either be kept local or promoted to a global Check that is then added to the core system.

5.2 Database Management

In a realistic setting a DQ manager will work with different databases, both different types of database systems and different application areas. This is reflected in the database management component of the DQ GUI as shown in Fig. 4. The DQ manager can create different DB environments (e.g. for test and production) and different database connections within these environments. These environments and database connections are created by typical database wizards asking for the specific connection parameters. When executing Actions later on a database from the supplied connections can be chosen. This makes it possible to re-use the same DQ rule for both test and production with only a minimal change in the code. Currently, MongoDB and PostgreSQL are supported by the wizard, but an extension interface for databases of a similar paradigm is also provided.

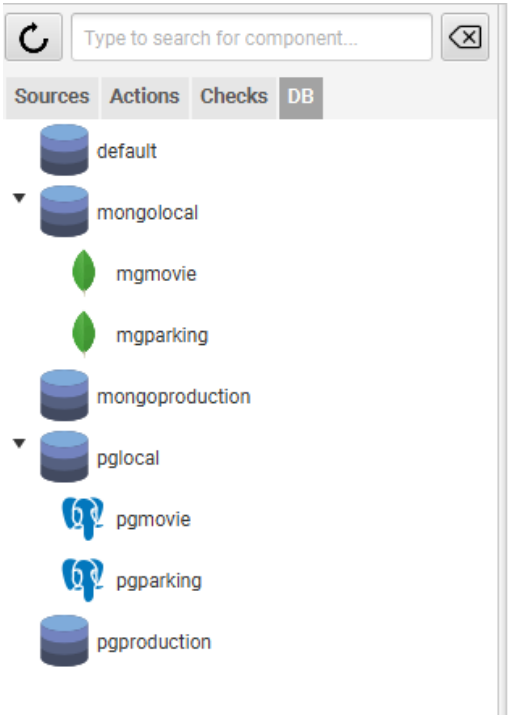


Fig. 4. Database Management with DQ GUI

5.3 Rule Execution

DQ rules (i.e. Actions in *RADAR*) can be executed from the DQ GUI. A drop down menu is used to select the database connection against which an Action (selected from another drop down menu listing all defined Actions in the repository) should be executed (cf. Fig. 5). The execution can be local on the maching running the GUI itself or remotely in case a remote server running the DQ engine is specified. This is particularly useful for long running, compute intensive rules where the machine running the GUI should not be affected by rule execution. Rules are always executed asynchronously and a notification is displayed in the GUI when an execution is completed. All rule execution results are also stored with the DQ project and displayed in the Reports part of the GUI (cf. Fig. 6). All executed Actions are listed with name and database connection as well as outcome using the traffic light system. Red indicates an execution failure, whereas yellow indicates executions resulting in detected quality errors. Green indicated executions without quality errors. Executions that have not been finished are listed under "Active". Double-clicking on any of the executed rules opens the window for execution result analysis which is explained in the next subsection.

Fig. 5. Rule execution start with DQ GUI

5.4 Analysis of Execution Results

Results of rule executions can be analyzed in the detail window shown in Fig. 7. On the left hand side the desired report can be selected. In the parts for "Info", "Database Trace" and "Log" the corresponding textual output from the execution on the respective log level can be seen (e.g. duration of a run and number of quality issues detected). In the "Issues" part opened in Fig. 7 the detected quality issues are displayed. Specifically, the severity level of the issue according to the definition in the Action, the score (in case of a numerical return value), details about the record violating the quality rule and information on the conditions checked in the rule. For each violating record an identifier and the violating attribute value are shown. The information regarding the identifier is defined in the data source definition; it is important for the DQ manager to actually locate the problematic record in the database of the data source. The first row in Fig. 7 for example shows that the record for January 2, 2019, has a checked value of 8012 (parking tickets per day) which is outside of the defined valid range from 0 to 6000 (with a badness score of 2012 = 8012-6000).

5.5 Feedback Collection

In the "Identifier Overview & Feedback" tab of the report detail the DQ manager is able to check the individual quality issues and assign his judgement on the issue. The

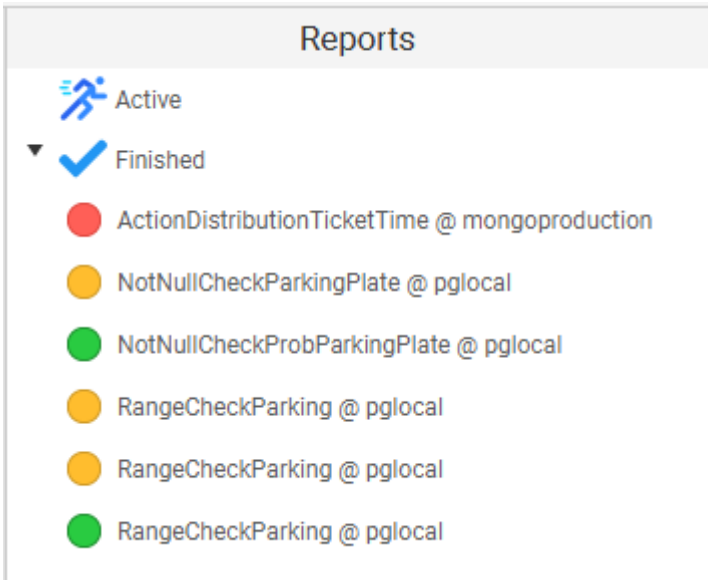


Fig. 6. Rule execution result overview

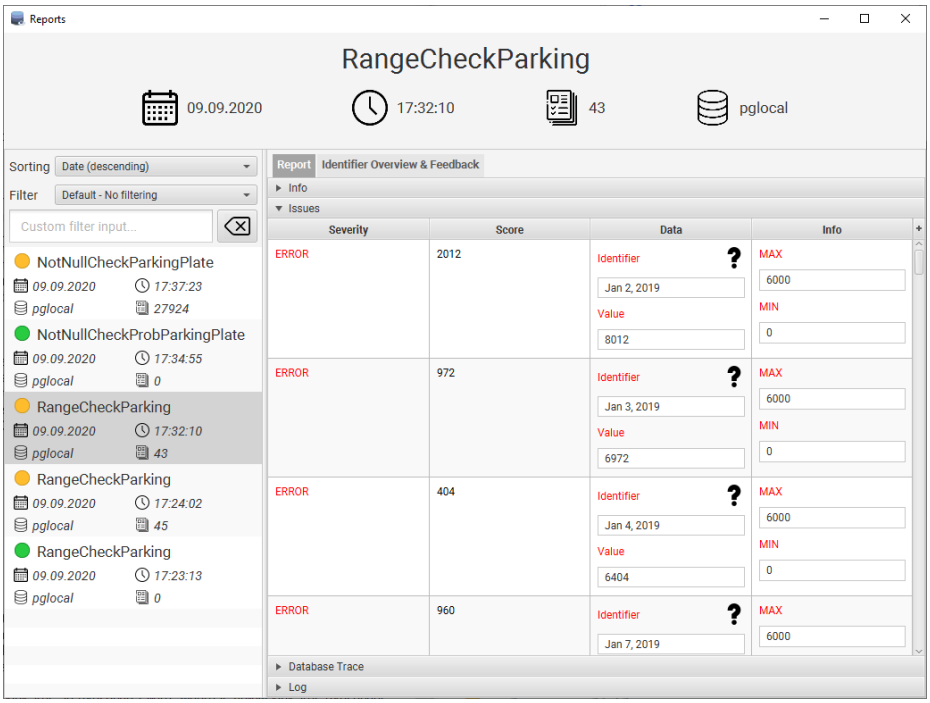


Fig. 7. Analysing quality issues

feedback from the domain expert can e.g. be used to modify quality rules for the future. As explained in section 2 this can even be done in a semi-automated manner by the profiling component which has not yet been integrated into the DQ GUI, however. In more detail, as shown in Fig. 8 the options are:

- to mark an issue as an actually correct data item which should not be reported as quality issue in the future; this information can be used for rule modifications
- to mark an issue as a false positive, essentially an outlier that is acceptable in this case but not in general; these should not lead to rule modifications
- no feedback (the default) which means this item is actually a DQ issue and the data in the data source should be corrected (outside of the scope of the project)

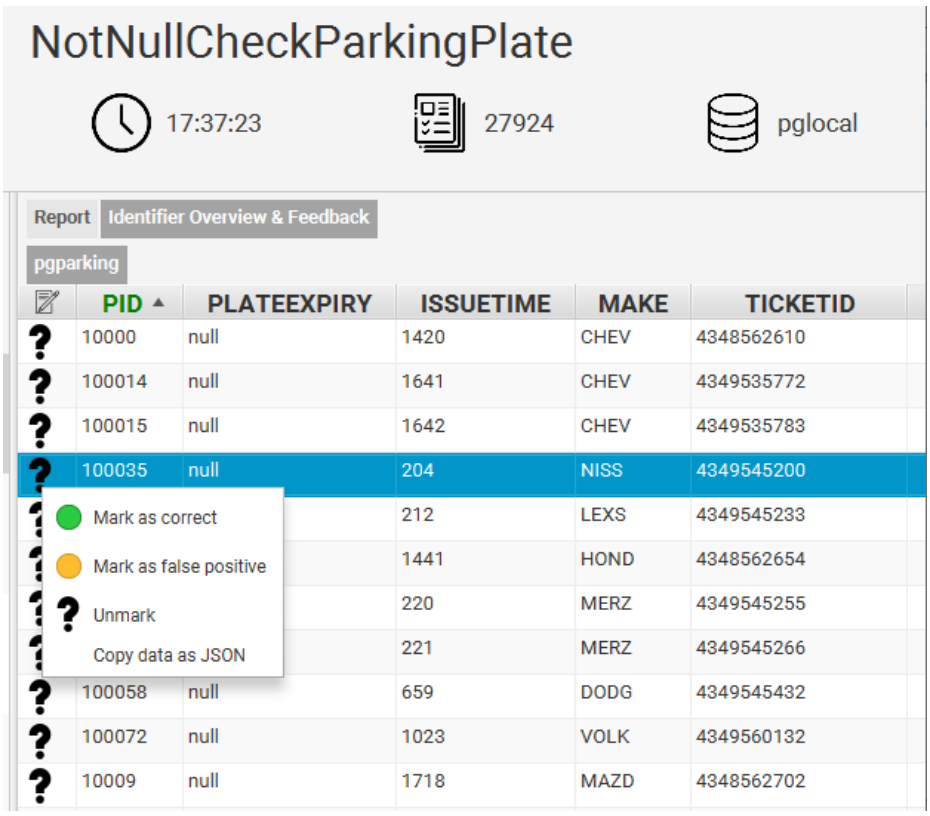


Fig. 8. Provding feedback to quality issues

The feedback is stored together with the report inside of the DQ project and can be accessed and processed further from there. This is the component "Data Quality Repository" as depcited in the architecture in Fig. 1.

6 Related Work

Older papers, including our own predecessor project Data Checking Engine (DCE) [12], that address partial aspects discussed in this paper (e.g. heterogeneous data integration) are typically not applicable for non-RDBMS data sources and thus cannot satisfy an important requirement of our solution. However, many solutions and tools for DQ measurement and monitoring exist that also address data not stored in RDBMS. For a recent tool survey, e.g. see [7]. We now describe examples of individual tools and then summarize the key differences.

In Apache Griffin[8] (based on Spark and Hive), there is a DSL. However, it only contains the parameters to predefined rule types. E.g., for the type "accuracy", the rule essentially contains a logical expression that provides the link between two tables that are to be compared. In general, the parameters are used to instantiate predefined SQL patterns according to the rule type. For more specific rules, direct SQL can be inserted.

Our main components Sources, Checks, and Actions are structurally similar to MobyDQ[9], however our Checks and Actions can be much more complex than indicators and alerts there and, in addition, data sources can be more diverse in *RADAR*.

Endler et al. developed an architecture for continuous data quality monitoring in medical centers [9]. They handle data quality monitoring with a rule-based approach. The most basic shown rules are simple boolean operations directly defined for the underlying database.

In [18], Schelter et al. present an approach with the goal to automate data quality checks in large data sets. A wide range of basic types of checks is present; there are methods to automatically suggest potential quality rules, and integration of heterogeneous data sources is possible. They use a sophisticated optimization procedure to efficiently process the checks based on Spark and address differential quality checking in detail.

The work of Ehrlinger [8] focuses on automated and frequent execution of the DQ checks. Similarly, [6] also focuses on a highly efficient evaluation of rather simple rules (and with a more specific target, namely improvement of machine learning algorithms). In principle, these approaches complement each other as these efficient execution engines could be applied in principle to our simple rules part as well. However, the actual implementation technologies differ.

In summary, the main difference of our approach to these works is our DSL *RADAR*. It allows to extend the core functionality of the system with user-defined rule types. The logic of the rules is independent of the concrete data sources and independent of the target execution environment (i.e. Hive or a RDBMS or MongoDB). The action part of the DSL allows to instantiate rules in combination with specific sources. With its well-defined internal data model, the DSL also independent of the data model of the underlying DBMSes. Each rule can even execute on heterogeneous environments, combining data from different systems for cross-checking. The engine can execute the logic either by itself, thereby retrieving the source data without any transformations, or it can decide to execute parts of the logic using the features of the underlying DBMS. To

[8] `http://griffin.apache.org`
[9] `https://github.com/ubisoftinc/mobydq`

the best of our knowledge, these are unique features, making our approach extremely flexible with respect to rule types and heterogeneous execution environments, while still allowing less technical users to define quality measurements. Furthermore, our DSL allows for advanced rule types like distribution checks (cf. [13]).

From the operational and query side, our work touches multiple lines of work in the database research that we are going to mention here. First of all, our *RADAR* language allows to formulate queries integrated into an imperative language. This is similar on the one hand to approaches like Oracle's PL/SQL [10] or other stored procedure languages, and on the other hand to LINQ-approaches (language integrated queries), see [5]. Yet another approach is to integrate SQL with functional programming [3]. However, our combination is specific to the domain of data quality and offers specific features.

Furthermore, *RADAR* can act as a mediator [11] allowing to query and integrate multiple sources using a single piece of code. Thus the query optimization issues are quite similar to those in mediator-based systems. However, we don't define a global schema but allow the checks to directly access the individual sources, which is more appropriate for data quality checks. Similarly, a lot of work has been published in federated query processing which can complement our system in the future.

RADAR can access both relational and NoSQL databases from the same code, which again is a generic problem in heterogeneous environments. Here, both query languages (i.e. UnQL, [4]) and systems have been developed. There are multiple mediator approaches, e.g. [2,21,17] that allow to query relational and NoSQL-data from a unified SQL-like language. They are to some degree comparable, however, they lack the parametrizing features vital for our approach.

Parametrizing SQL in the way we do it – i.e. parameterizing also schema elements like attributes and relations and with the option to e.g. use a compound foreign key as actual parameter for a single formal parameter – is, as far as we know, a new feature. From our point of view it is vital to achieve the flexible modularization and reuse-capabilities we are targeting. The only comparable approach uses the Maude system [19] to rewrite parametrized SQL statements. However, we integrated the parametrization capabilities into our core language.

7 Conclusion and Future Work

In this paper we have presented the *IQM4HD* data quality monitoring system and its DSL *RADAR* used to specify data quality rules. The DSL has been developed based on a fixed set of requirements that have been set up together with company partners as potential users.

We have discussed the main principles of the DSL, using both simple and complex rules to illustrate the ideas. Our prototype is able to execute these quality checks on heterogeneous types of databases providing the source data. We have also shown how our engine optimizes rule execution by pushing parts of the execution logic towards the sources whenever possible. The implementation of the DSL allows for easy extensibility in case additional functionality is required. Other complex quality rules can also be specified manually and simple as well as complex quality rules can also be generated automatically based on analysis of existing data (profiling).

Within the project *RADAR* rules to check quality of real-world data provided by our project partner CTS Eventim, a large European online ticket seller, have been successfully used. In particular, rules have been applied to web tracking data to detect potential data quality issues, e.g. caused by a malfunctioning detection of user agent types. In summary, *RADAR* together with the prototypical implementation provides an important step forward in efficiently managing data quality in data warehouses.

In the remainder of this section, we first discuss main DQ monitoring requirements and explain how they are supported by *IQM4HD* and *RADAR*. Finally, we discuss future work in the context of the *IQM4HD* system.

7.1 Requirements Review

In this section, we review the main features of our proposed DSL *RADAR* and the *IQM4HD* system and discuss how they meet important DQ monitoring requirements.

Extensibility and Reusability We think that an extensible rule logic is essential, so that users can specify new rule types in a flexible and reusable way. To achieve reusability, the logic behind the user-defined rules must be decoupled from the actual data source, so that the same logic be applied to multiple data sources. The separation between Sources, Checks and Actions fulfills this requirement. The check logic is independent of the data source, and of the type of the data source.

Heterogeneous Data Sources Data is stored using various technologies and different data models. A DQ system must be able to access and compare data from such heterogenous sources. This requires flexibility w.r.t. technological aspects and aspects of the data model. We resolve this by translating the data into the common internal format, so that all further components can be agnostic of the particular type of data source. A single Check can e.g. use data from a relational and a document database to test consistency constraints between the two databases.

Optimizability As data sources might be huge, execution speed is an important issue, even though the checks are typically scheduled and executed in batch mode. For this, it is beneficial to make use of the features of the underlying DBMS like query optimization and available indices. However, as the available features differ, our optimizer component can decide which parts of the logic are executed using native DBMS features or within our execution engine.

User groups In DQ teams, multiple persons with different expertise are involved. First, there are experts for the data sources to be monitored, that understand the data models of these systems. Second, we have domain experts that understand the business logic and can define which data is valid from a business point of view. Third, there are programmers that understand how to code the business logic programmatically. We need the respective parts of the language to be accessible by the corresponding groups. Again, the separation in Sources, Checks and Actions supports this requirement. Sources are defined by database experts that know the underlying database models; Checks are either pre-defined by our system or can be extended by technical experts, and Actions are defined by domain experts.

Advanced Quality Rules Apart from rather simple data quality checks such as non-null or range checks the rule language should also provide the option to specify advanced data quality checks. The DistributionCheck shown in Sect. 2.4 is one example

of an advanced rule. Other examples include outlier identification in time series data and checks on cube data analyzing relevant aggregations of the data (cf. Sect. 2.6).

7.2 Future Work

Several issues remain to be resolved in the future. While the applicability to different types of data sources has been shown in principal by using relational as well as document databases in the prototype, an extension to more types of data sources needs to be implemented. As stated before, execution optimization remains a major task in future work. From the practical experiments with the CTS Eventim data we can conclude that further optimization for relational data sources is beneficial. Separating functionality between DQ system and native source has to be addressed for non-relational data sources as well. Nevertheless, the DSL itself will remain unchanged even for different types of sources. Finally, future extensions of the DSL capabilities, particularly in the body of Checks, might be necessary to provide functionality that has not been necessary so far.

References

1. Agarwal, S., Agrawal, R., Deshpande, P., Gupta, A., Naughton, J.F., Ramakrishnan, R., Sarawagi, S.: On the Computation of Multidimensional Aggregates. In: Proceedings of the 22th International Conference on Very Large Data Bases. pp. 506–521. VLDB '96, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (1996), `http://dl.acm.org/citation.cfm?id=645922.673497`
2. Atzeni, P., Bugiotti, F., Rossi, L.: SOS (Save Our Systems): A Uniform Programming Interface for Non-relational Systems. In: Proceedings of the 15th International Conference on Extending Database Technology. pp. 582–585. ACM, New York, NY, USA (2012)
3. Binnig, C., Rehrmann, R., Faerber, F., Riewe, R.: FunSQL: It is Time to Make SQL Functional. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops. pp. 41–46. ACM, New York, NY, USA (2012)
4. Buneman, P., Fernandez, M., Suciu, D.: UnQL: a query language and algebra for semistructured data based on structural recursion. The VLDB Journal **9**(1), 76–110 (Mar 2000)
5. Cheney, J., Lindley, S., Wadler, P.: A Practical Theory of Language-integrated Query. In: Proceedings of the 18th ACM SIGPLAN International Conference on Functional Programming. pp. 403–416. ICFP '13, ACM, New York, NY, USA (2013)
6. Ehrlinger, L., Haunschmid, V., Palazzini, D., Lettner, C.: A DaQL to Monitor Data Quality in Machine Learning Applications. In: Database and Expert Systems Applications - 30th International Conference, DEXA 2019, Linz, Austria, August 26-29, 2019. LNCS, vol. 11706, pp. 227–237. Springer (2019)
7. Ehrlinger, L., Rusz, E., Wöß, W.: A Survey of Data Quality Measurement and Monitoring Tools. arXiv preprint 1907.08138 (2019)
8. Ehrlinger, L., Wöß, W.: Automated data quality monitoring. In: MIT International Conference on Information Quality. vol. 22, pp. 19–1–19–8 (2017)
9. Endler, G., Schwab, P.K., Wahl, A.M., Tenschert, J., Lenz, R.: An Architecture for Continuous Data Quality Monitoring in Medical Centers. MEDINFO (2015)
10. Feuerstein, S., Pribyl, B.: Oracle PL/SQL Programming. O'Reilly Media, Inc. (2005)

Usability Testing of Visual Policy Evaluation for Network Security Event Detection*

Volker Ahlers und Bastian Hellmann

University of Applied Sciences and Arts Hannover, Germany
Faculty IV, Dept. of Computer Science
volker.ahlers@hs-hannover.de

Zusammenfassung. The network security framework VisITMeta allows the visual evaluation and management of security event detection policies. By means of a "what-if" simulation the sensitivity of policies to specific events can be tested and adjusted. This paper presents the results of a user study for testing the usability of the approach by measuring the correct completion of given tasks as well as the user satisfaction by means of the system usability scale.

Kurzfassung. Das Netzwerksicherheits-Framework VisITMeta ermöglicht die visuelle Evaluation und Verwaltung von Erkennungsregeln für Sicherheitsvorfälle. Mit Hilfe einer "Was-wäre-wenn"-Simulation kann die Sensitivität der Regeln für bestimmte Vorfälle getestet und angepasst werden. Diese Arbeit stellt die Ergebnisse einer Nutzerstudie zur Untersuchung der Gebrauchstauglichkeit des Ansatzes vor, in der die korrekte Bearbeitung vorgegebener Aufgaben sowie die Nutzerzufriedenheit mittels des *System Usability Scale* gemessen werden.

Keywords: Network Security · Policy Evaluation · Information Visualization · Visual Analytics · User Interfaces · Usability Testing.

1 Visual Analytics of Network Security

The monitoring and protection of network security is usually performed by a combination of different detection systems such as firewalls, network access control (NAC), and intrusion detection systems (IDS). Several approaches have been proposed for the visualization of the – in general time-varying – security information, e. g., by means of graphs showing connections between hosts and users or between security-relevant events [8,4], or by means of information dashboards displaying the information of separate security detection components within a single screen [6].

* This work was supported by German Federal Ministry of Education and Research (BMBF) within the projects VisITMeta and SIMU (grant nos. 17PNT032, 16KIS0045). The fruitful collaboration with Gabi Dreo Rodosek, Felix Heine, Carsten Kleiner, the Trust@HsH group, and our industry partners is gratefully acknowledged.

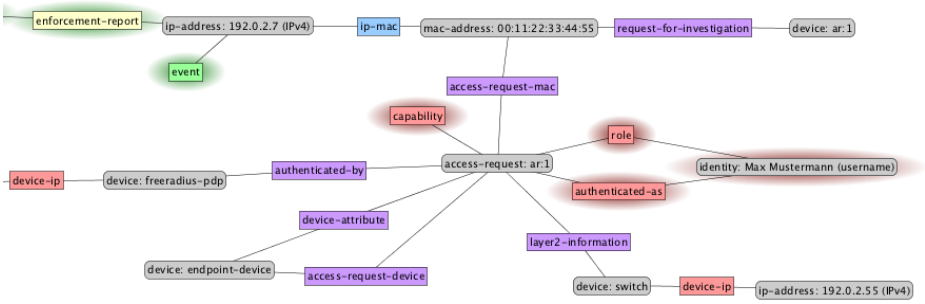


Abb. 1. Part of an example IF-MAP graph visualized with force-directed layout. Identifier nodes are drawn with rounded edges, metadata nodes with sharp edges. The VisITMeta user interface adds a green and red glow effect to nodes that have been added and removed during a selected time interval, respectively.

In order to overcome the limitations of separate information sources, the *Interface for Metadata Access Points* (IF-MAP) specification defines a system architecture for the collection of data from various physical and logical network components (MAP clients) by a central MAP server [9]. The graph-based data model consists of two types of nodes named identifiers and metadata, which are connected by edges. Identifiers represent physical and logical entities such as devices, IP and MAC addresses, or users. Metadata represent different types of connections between identifiers as well as detected network security events.

Our visual analytics framework VisITMeta employs the IF-MAP data model and architecture to visualize network security information, as shown in Fig- 1. A core feature is the storage of historic network data in a graph database, which allows the comparison of network states at different time instances as well as the visualization of network changes within a selected time period. Details on the underlying concepts and the implementation can be found in our previous works [2,3]. The open-source software VisITMeta and related projects are available via GitHub [10].

2 Visual Policy Evaluation

By an extension of the IF-MAP specification, the VisITMeta framework offers the visualization of a detected security event together with the current network state and the detection policy that triggered the event, as shown in Fig. 2. This enables the security administrator to trace the network state and check the plausibility of the detected event, thereby keeping the mental model of the graph-based visualization. In case of false positives, i. e., wrongly detected events, or false negatives, i. e., known events that remain undetected, the administrator can adjust the rules of the policy and re-evaluate it against the historical network state. The underlying “what-if” simulation uses the data model and the data exchange mechanism of the IF-MAP specification [7].

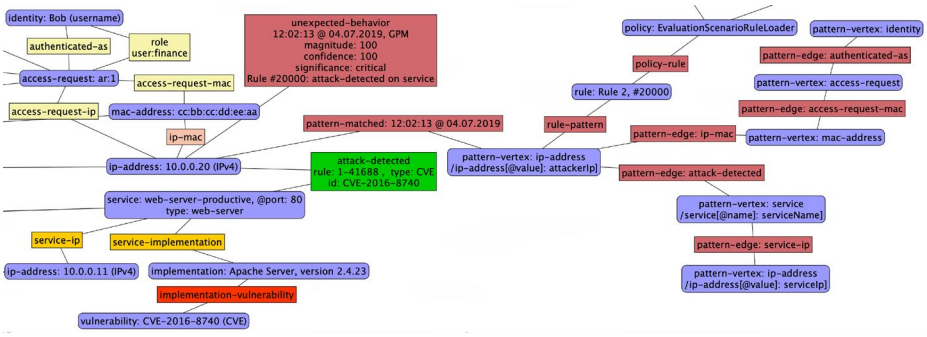


Abb. 2. Example of a detected security event (green) visualized together with the current network state (left) and the policy that triggered the event (right). The relevant identifiers of the network state and the policy are connected via a "pattern-matched" metadata node.

3 Usability Testing

Study Design. In order to test the usability of our IF-MAP-based visualization framework VisITMeta as well as the visual policy evaluation, we conducted a user study consisting of three scenarios with different tasks to be solved on a laptop computer [1]:

A: Find out information on an enterprise network from log files and security component reports (variant A1) as well as from the VisITMeta visualization (variant A2) in random order A1 $\leftrightarrow$ A2.
B: Examine the correctness of policy rules for a detected event.
C: Adapt the detection policy using the "what-if" simulation.

The details and questions of the tasks will be described below.

The two principle variables that have been measured are the task correctness, i. e., the proportion of correctly solved tasks, and the processing time, i. e., the time required to process all three tasks. Furthermore, a retrospect evaluation of scenario 1 based on the system usability scale as well as a general questionnaire for all three scenarios have been applied after the completion of the tasks.

Study Execution. The study has been conducted with $n = 12$ test subjects: 7 computer science students, 2 research assistants, 2 IT administrators, and 1 IT security consultant. The test subjects were asked to self-assess their prior aggregated knowledge (AK) and experience in the following fields on a scale from 0 (no prior knowledge) to 4 (extensive prior knowledge):

AK1: analysis of log output,
AK2: usage of general software and/or hardware for network security,
AK3: knowledge of the IF-MAP specification,
AK4: experience with the VisITMeta framework.

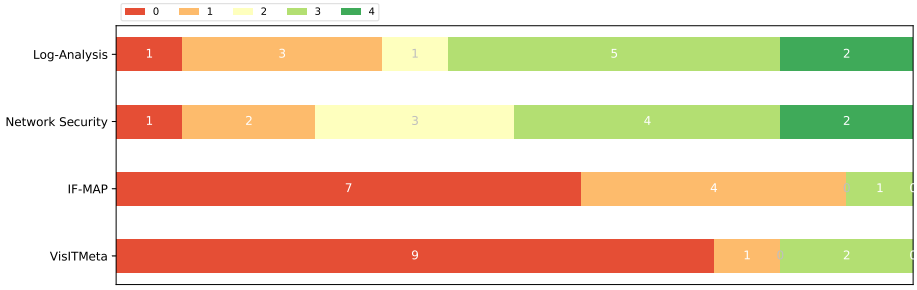


Abb. 3. Prior aggregated knowledge of test subjects ($n = 12$) in different fields by self-assessment on a scale from 0 (no prior knowledge) to 4 (maximum prior knowledge).

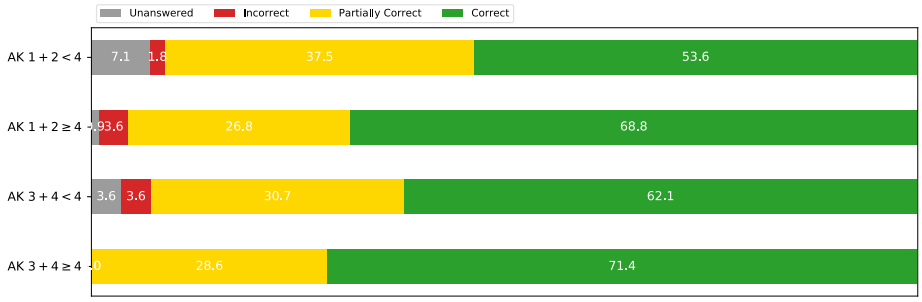


Abb. 4. Overall task correctness (percentage) for all three scenarios in dependence of prior aggregated knowledge AK1 + AK2 and AK3 + AK4, respectively (cf. text for explanation).

The distribution of prior knowledge is shown in Fig. 3. While more than half of the test subjects judge themselves to have good knowledge of log analysis and general network security, very few have prior experience with IF-MAP or even the VisITMeta framework.

Influence of Prior Aggregated Knowledge. The proportion of correctly solved tasks for all three scenarios in dependence of prior knowledge is shown in Fig. 4. While a higher degree of prior knowledge leads to better results (as expected), the difference is more pronounced for the fields AK1 + AK2 (log analysis and general network security) than for the fields AK3 + AK4 (IF-MAP and VisITMeta), indicating that the visualization does not require experience with the underlying technologies.

Processing Time. The median processing time for all tasks of the three scenarios has been 106 minutes, with a minimum of 86 minutes and a maximum of 130 minutes. The processing time includes a short introduction of the study supervisor. Furthermore the test subjects had the opportunity to ask the supervisor if they had problems understanding scenarios or the tasks.

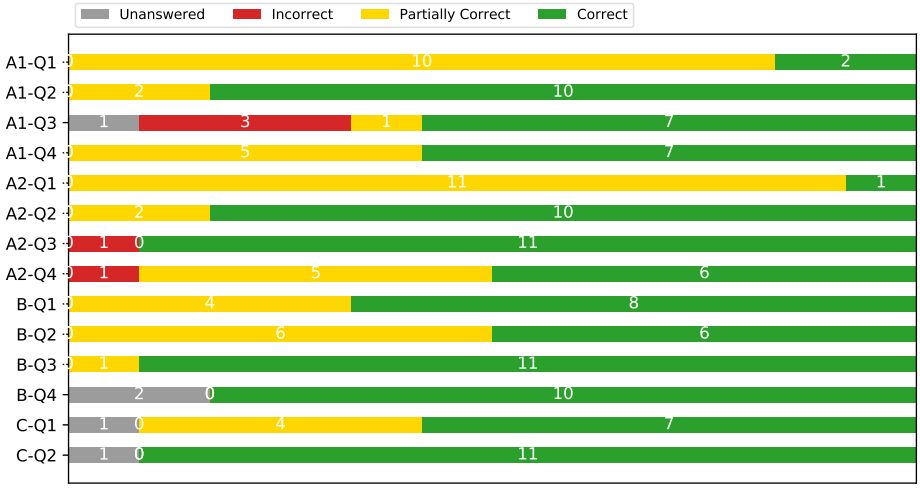


Abb. 5. Task correctness ($n = 12$) for questions Q1 to Q4 of scenarios A and B and Q1 to Q2 of scenario C. Scenario A has the variants A1: monitoring based on log files and security component reports, and A2: monitoring using the VisITMeta visualization (cf. text for further explanation).

Task Correctness. The tasks of the three scenarios A to C included two to four questions for each scenario that had to be answered by the test subjects. These questions can be summarized as follows.

A (network monitoring): List all IP addresses (Q1) and users (Q2) with associated parameters, identify a network component that authenticated a certain user (Q3), describe any detected vulnerabilities of the network (Q4).

B (security event analysis): Identify any events detected by a certain security component (Q1), state the timestamps of each event (Q2), find out which software was running on target device (Q3), decide which of two possible rules produces more false positives (Q4).

C (policy evaluation): Identify the policy rule responsible for a false positive event (Q1), modify the rule to prevent further false positives.

As explained above, scenario A has two variants: In variant A1 the network monitoring is based on the analysis of log files and security component reports, while in variant A2 the VisITMeta visualization is used. The two variants are later compared based on the system usability scale (see below).

The correctness of the answers to the individual questions is displayed in Fig. 5. A few incorrect or partially correct results were found to be the result of a misunderstanding of the task descriptions.

For scenario A (network monitoring), the task correctness for the two variants A1 (log data and security component reports) and A2 (VisITMeta visualization) is comparable. A significant difference only exists for question Q3 (identification of authentication component). One possible explanation for the close results

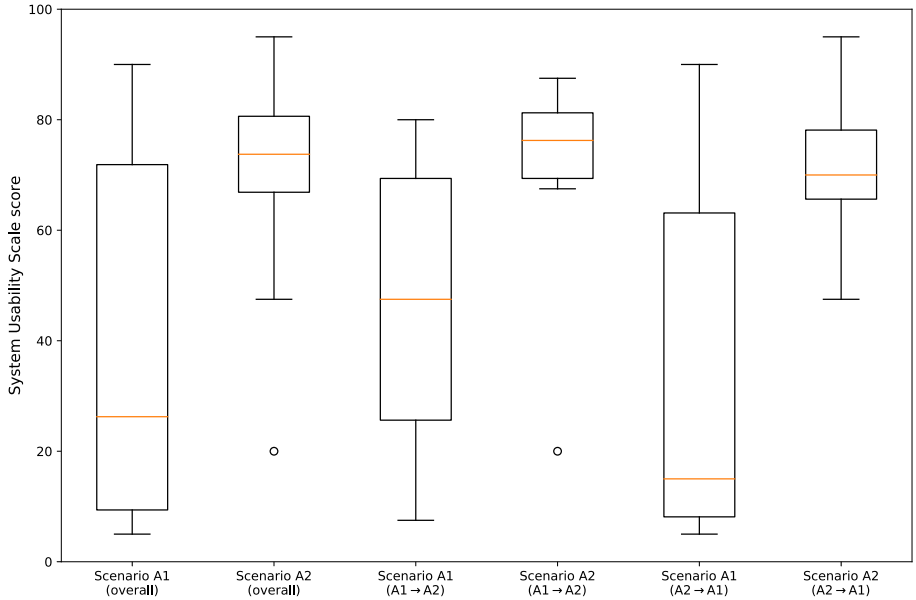


Abb. 6. System usability scale for scenario A (network monitoring) with variants A1: monitoring based on log files and security component reports, and A2: monitoring using the VisITMeta visualization, in dependence of the random order of presenting both variants to the test subjects.

is that most of the test subjects were used to work with log data and security reports, but had to familiarize themselves with the VisITMeta application in a very short time.

The results of scenarios B (security event analysis) and C (policy evaluation) can be considered good, given the fact that most test subjects (9 out of 12) had no prior experience with the VisITMeta framework.

System Usability Scale. The two variants of scenario A (network monitoring) have further be compared by means of the system usability scale (SUS) [5], which consists of 10 standard questions that have been slightly adopted to the prototypical nature of the software. The results are shown in Fig. 6. Summarizing over all results independent of the order the of presenting both variants to the test subjects, the median SUS is 26.3 for variant A1 (log data and security component reports) compared to a median SUS of 73.8 for variant A2 (VisITMeta visualization). Given the common interpretation that a system is usable if the SUS is higher than 68, this is a clear support of the visualization approach. The difference is more pronounced for the order A2 $\rightarrow$ A1, i. e., first using VisITMeta and afterwards the log data and security reports, but it is also significant for the order A1 $\rightarrow$ A2, i. e., the other way round.

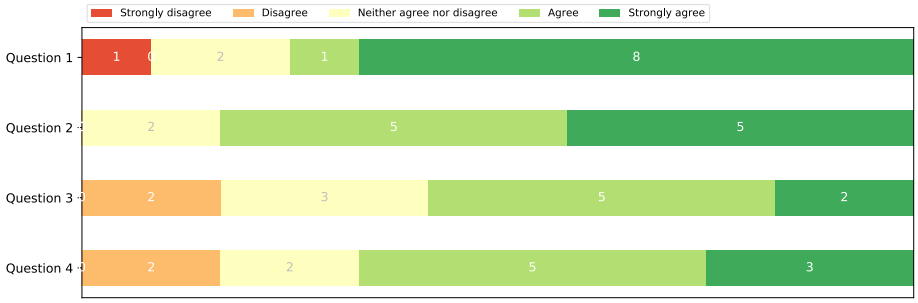


Abb. 7. Answers to the questionnaire with the following four questions ($n = 12$):
1. Can the first scenario be easier solved with visualization rather than with log output?
2. Is the consistent visualization between different scenarios helpful?
3. Were you aware of the fact that data was gathered from different components?
4. Is the integration of detection rules as well as their interconnections helpful?

Questionnaire. Finally, the test subjects were asked to fill in a questionnaire with four questions as stated in the caption of Fig. 7. The results also give a clear indication that the VisITMeta visualization as well as the visual policy evaluation approach are considered helpful.

4 Conclusion

We have presented results of a user study for the usability testing of the visual analytics framework VisITMeta for monitoring the network security. The framework is based on the IF-MAP specification and employs a graph-based visualization of physical and logical network components. It furthermore offers a visual evaluation and management of security event detection policies by means of a "what-if" simulation using historical network data.

While the correctness of typical network monitoring tasks was found to be comparable when using either classical log files and security component reports or the VisITMeta visualization, the usability in measures of the system usability scale was assessed significantly higher for the visualization solution. This as well as the helpfulness of the visual policy evaluation approach were further supported by the answers to a questionnaire.

The scalability to large and highly dynamic networks remains a major challenge of our approach. Finding solutions for these issues will be the subject of future work.

Literatur

1. Ahlers, V., Hellmann, B., Dreo Rodosek, G.: A user study of the visualization-assisted evaluation and management of network security detection events and policies. In: 2019 10th IEEE International Conference on Intelligent Data Acquisition and

Advanced Computing Systems: Technology and Applications (IDAACS). vol. 2, pp. 668–673. IEEE, Piscataway, NJ, USA (2019). https://doi.org/10.1109/IDAACS.2019.8924439
2. Ahlers, V., Heine, F., Hellmann, B., Kleiner, C., Renners, L., Rossow, T., Steuerwald, R.: Integrated visualization of network security metadata from heterogeneous data sources. In: Proc. GraMSec 2015. pp. 18–34. Springer International Publishing, Cham, Switzerland (2016). https://doi.org/10.1007/978-3-319-29968-6_2
3. Ahlers, V., Hellmann, B.: Visual analytics of network security metadata (invited paper). In: Proceedings of IWEIC 2019 Hiroshima: International Workshop on Electronics, Information and Communication. pp. 1–4. Hiroshima City University, Hiroshima, Japan (2019)
4. Angelini, M., Prigent, N., Santucci, G.: Percival: proactive and reactive attack and response assessment for cyber incidents using visual analytics. In: Proc. VizSec 2015. pp. 1–8. IEEE, Piscataway, NJ, USA (2015)
5. Brooke, J.: SUS – a quick and dirty usability scale. Usability Evaluation in Industry **189**(194), 4–7 (1996)
6. Goodall, J.R., Ragan, E.D., Steed, C.A., Reed, J.W., Richardson, G.D., Huffer, K.M.T., Bridges, R.A., Laska, J.A.: Situ: Identifying and explaining suspicious behavior in networks. IEEE Transactions on Visualization and Computer Graphics **25**(1), 204–214 (2019). https://doi.org/10.1109/TVCG.2018.2865029
7. Hellmann, B., Ahlers, V., Dreo Rodosek, G.: Integrating visual analysis of network security and management of detection system configurations. In: 2017 9th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS). vol. 2, pp. 1020–1025. IEEE, Piscataway, NJ, USA (2017). https://doi.org/10.1109/IDAACS.2017.8095240
8. Liao, Q., Striegel, A., Chawla, N.: Visualizing graph dynamics and similarity for enterprise network security and management. In: Proc. VizSec 2010. pp. 34–45. ACM, New York, NY, USA (2010). https://doi.org/10.1145/1850795.1850799
9. Trusted Network Communications Working Group: TNC IF-MAP binding for SOAP, version 2.2, revision 10 (March 2014), `https://trustedcomputinggroup.org/wp-content/uploads/TNC_IFMAP_v2_2r10.pdf`
10. Trust@HsH Group: Iron/VisITMeta project suite on GitHub, `https://github.com/trustathsh/`

Screen Reject Data Warehouse – Klinisches Data Warehouse zur Abstoßungsdiagnostik nach NTx: Ein Werkstattbericht

Matthias Katzensteiner, Nina Schewe, Maximilian Zubke, Kim Nörtemann, Bianca Frenzel und Oliver .J. Bott

Fakultät III – Medien, Information und Design – der Hochschule Hannover (HsH)
matthias.katzensteiner@hs-hannover.de

Zusammenfassung. Im Rahmen der Nachsorge nierentransplantierter Patient*innen kommt der Abstoßungsdiagnostik besondere Bedeutung zu, um eine etwaige Abstoßungsreaktion rechtzeitig zu erkennen und behandeln zu können. Im Rahmen des Screen-Reject-Verbundprojekts wird unter anderem ein klinisches Data Warehouse in Kombination mit einem Expertensystem entwickelt, das die Abstoßungsdiagnostik durch Datenaufbereitung und Integration eines Prädiktionssystems für das Vorliegen einer Abstoßung unterstützen soll.
Der Werkstattbericht stellt die Methodik und bislang erreichten Ergebnisse dieses Teilprojekts vor. Entwickelt wurde ein relationales Data Warehouse, das die relevanten klinischen Daten der klinischen Primärsysteme in einem Datenmodell unter anderem unter Einsatz von Text Mining Methoden zusammenführt, sowie eine auf dem internationalen openEHR-Standard basierende und diese Daten bereitstellende elektronische Patientenakte. Weiterhin vorgestellt wird der Stand der auf diesen Datenbeständen aufbauenden Arbeiten zur Entwicklung eines Prädiktionssystems auf Basis von Machine Learning Ansätzen sowie zur Datenvisualisierung für klinische Expert*innen.

1 Hintergrund

Bedingt durch den demografischen Wandel und der damit älter werdenden Bevölkerung ist mit einer Zunahme der Anzahl schwerwiegender Nierenerkrankungen und daher mit einer zunehmenden Anzahl notwendiger Nierentransplantationen (NTx) zu rechnen. Dieser steht ein zuletzt sinkendes Angebot verfügbarer Spenderorgane gegenüber, wobei bereits heute eine erhebliche Unterversorgung besteht: die Warteliste für Nierentransplantation betrug (Stand 31.12.2019) mehr als 7.100 Patient*innen [10]. Zudem ist bei Nierentransplantationen die Überlebenszeit des Transplantats begrenzt. In Europa liegt die 10-Jahres Überlebensrate eines Transplantats gemäß [9] bei 56%, weswegen transplantierte Patient*innen nicht selten mehrfach transplantiert werden müssen. Mit Blick einerseits auf die Zunahme von Nierentransplantationen sowie andererseits auf die zuletzt zunehmende Unterversorgung mit Spenderorganen ist es umso wichtiger, dass nach

durchgeführten Transplantationen kein durch eine nicht rechtzeitig behandelte Abstoßungsreaktion induzierter Verlust des Spenderorgans eintritt. Leider sind Abstoßungsreaktionen nicht selten: akute zelluläre Abstoßungen treten im ersten Jahr nach NTx zu 30% auf, humorale Abstoßungen über den gesamten Verlauf einer NTx sogar in bis zu 70 oder 80% der Fälle [6].

Der Gold-Standard für die Abstoßungsdiagnostik ist Biopsie, d.h. die histologische Untersuchung einer Gewebeprobe des Transplantats und einer rein klinischen Diagnostik deutlich überlegen [2]. Allerdings handelt es sich bei einer Nierenbiopsie um einen invasiven Eingriff, der, wenn auch sehr selten, zu schwerwiegenden Komplikationen führen kann [20]. Im von der Leibniz-Universität Hannover (LUH), der Medizinischen Hochschule Hannover (MHH) und der Hochschule Hannover (HsH) durchgeführten Verbundprojekt Screen-Reject wird ein Diagnostikum, der Lateral Flow Test, in Kombination mit einem Expertensystem entwickelt, das frühzeitig und nicht invasiv Abstoßungen bei NTx-Patient*innen feststellen soll. Unterstützt werden soll dadurch das Monitoring des Transplantats mit dem Ziel, dessen Funktionsfähigkeit zu erhalten und damit die Lebenssituation nierentransplantierter Patient*innen zu verbessern.

Im Teilprojekt „Screen-Reject/KDWH: Klinisches Data Warehouse zur Abstoßungsdiagnostik nach NTx“, das Gegenstand dieses Zwischenberichts ist, wird ein Klinisches Data Warehouse (das Screen-Reject DWH: SR-DWH) konzipiert und implementiert, das die bei der Entwicklung des Lateral Flow Test Testsystems gewonnenen Informationen aufbereitet und der wissenschaftlichen Begleitforschung strukturiert zur Verfügung stellt. Weiterhin bildet das SR-DWH den Ausgangspunkt der Entwicklung eines Expertensystems zur Unterstützung der Abstoßungsdiagnostik und Indikationsstellung diagnostischer Maßnahmen für Transplantationszentren und NTx-Versorgungseinrichtungen. Um dies zu erreichen werden im Rahmen des Teilprojekts auf dem SR-DWH basierende entscheidungsunterstützende Verfahren auf Grundlage von klinischen Versorgungsdaten entwickelt [11]. Um die entwickelten Verfahren später auch in unterschiedlichen Systemumgebungen einsetzen zu können, verfolgt das Projekt eine auf semantische Interoperabilität ausgerichtete Datenmodellierung und integration. Um diesen Anspruch der Interoperabilität zu erfüllen, erfolgt die Datenintegration in Anlehnung an entsprechende Konzepte des HiGHmed-Konsortiums der Medizininformatik-Initiative [1], auf Basis des internationalen openEHR-Standards für elektronische Gesundheitsakten [18].

Im Folgenden wird über Zielsetzung, Methodik und erste Ergebnisse des Teilprojekts Screen-Reject/KDWH berichtet, dessen Laufzeit nach Beginn am 01.05.2017 Ende Dezember 2020 enden wird.

2 Zielsetzung

Zentrale Zielsetzung des Teilprojekts Screen-Reject/KDWH, ist die Entwicklung des klinischen Data Warehouse (SR-DWH), wobei dieses zunächst als relationales Data Warehouse zur Datenübernahme aus den Primärsystemen des behandeln-

den Krankenhauses entwickelt wird [13] und nachfolgend eine Übertragung der Daten in eine openEHR-basierte elektronische Patientenakte erfolgt. Für beide Teilziele ist die Entwicklung eines zur Datenintegration geeigneten Datenmodells von besonderer Bedeutung, das zum einen geeignet sein muss, die Daten aus den Primärsystemen des Krankenhauses in das relationale SR-DWH aufzunehmen und zum anderen in Bezug auf die Datenübernahme in eine auf openEHR basierende elektronische Patientenakte auf idealerweise international abgestimmten Datenmodellen für klinische Konzepte aufbauen muss. Für die Datenzusammenführung gemäß openEHR sind dafür geeignete Archetypen (Datenmodelle für einzelne klinische Konzepte wie z.B. ein Kreatinin-Laborwert) und Templates (Datenmodelle für komplexere klinische Zusammenhänge wie z.B. ein aus mehreren klinischen Informationen bestehender Pathologiebefund) zu entwickeln.

Auf der Entwicklung des SR-DWH aufbauend wird das Ziel der Entwicklung eines Prädiktionssystems basierend auf Machine Learning Modellen verfolgt, um die Wahrscheinlichkeit des Vorliegens einer Abstoßung basierend auf den jeweils vorliegenden klinischen Patientendaten zu bestimmen und damit als Expertensystem die Abstoßungsdiagnostik des Behandlungsteams zu unterstützen. Hierfür sind neben der Bestimmung der patientenbezogenen Abstoßungswahrscheinlichkeit die weiteren für die Bestimmung der Abstoßungswahrscheinlichkeit relevanten Patientendaten für das klinische Behandlungsteam und die diagnostische Entscheidungssituation geeignet zu visualisieren.

3 Material und Methoden

Zur Zusammenführung der relevanten Daten der in die klinische Studie des Projektes eingeschlossenen Patient*innen in das relationale SR-DWH (s. 3.1) wurden zunächst ein relationales Datenmodell und ein entsprechendes Datenbankschema entwickelt (s. 3.1.1). Basierend auf der technologischen Plattform des Enterprise Clinical Research Data Warehouse (ECRDW) der Medizinischen Hochschule Hannover [8] wurde das entwickelte Datenbankschema in der Forschungsdatenbank-Schicht des ECRDW umgesetzt. Um darauf aufbauend eine für die klinische Fragestellung hinreichend umfassende Datenbasis für das SR-DWH zu generieren, wurden mittels ETL-Programmierung Daten aus dem ECRDW in das SR-DWH übernommen (s. 3.1.2). Zur Unterstützung dieser Integrationsprozesse in das SR-DWH wurden relevante Daten des ECRDW in einem sogenannten „Data Mart“ („Datenschnitt) vorab zusammengefasst und zur Integration bereitgestellt. Insofern relevante Datenbestände nicht bereits im ECRDW vorhanden waren, wurde eine Übernahme der Daten direkt aus den betreffenden klinischen Primärsystemen erforderlich. Zu zwei Befundungssystemen ist keine Schnittstelle zu maschinenlesbaren Datenformaten verfügbar, so dass die Extraktion der Schlüsselinformationen aus den natürlichsprachlichen Befundtexten mittels Text Mining Methoden erfolgt (s. 3.1.3, [27]). Im Rahmen der Datenintegration und im Hinblick auf die

Entwicklung des Prädiktionssystems wurde die Datenqualität der medizinischen Daten analysiert und soweit erforderlich und möglich Maßnahmen zur Datenaufbereitung und Verbesserung der Datenqualität ergriffen (s. 3.1.4). Unter Nutzung eines ETL-Tools [23] wurden die Daten in eine openEHR-Plattform [22] überführt und vorbereitend hierfür ein interoperables Datenmodell entwickelt (s. 3.2). Zunächst basierend auf dem SR-DWH und nachfolgend nach Anpassung basierend auf der interoperablen, auf openEHR basierenden elektronischen Patientenakte, werden geeignete Machine Learning Modelle zur Prädiktion der Wahrscheinlichkeit des Vorliegens einer Abstoßung ausgewählt, trainiert und bezüglich der Genauigkeit der Prädiktion evaluiert (s. 3.3) sowie ein webbasiertes, auf der openEHR-basierenden Patientenakte basierendes Anwendungsprogramm zur patientenbezogenen Visualisierung des Prädiktionswertes sowie der für die Abstoßungsdiagnostik relevanten klinischen Patientendaten zur Unterstützung des Behandlungsteams bei der Abstoßungsdiagnostik entwickelt (s. 3.4).

3.1 Datenintegration in das SR-DWH

Die relevanten Patientendaten der Studienpatient*innen wurden zunächst in einem relationalen Data Warehouse zusammengeführt. Hierfür wurde zunächst ein geeignetes relationales Datenmodell und Datenbankschema entwickelt und darauf aufbauend ETL-Prozesse zur Übernahme der Daten aus den Quellsystemen programmiert. Neben strukturiert vorliegenden Daten waren wesentliche Informationen aus Freitexten wie z.B. Befundtexten zu extrahieren, wofür Text Mining Methoden genutzt wurden. Weiterhin erforderlich waren Maßnahmen zur Datenaufbereitung und qualitätssicherung.

Datenmodell des SR-DWH Im Verlauf der stationären Behandlung von NTx-Patient*innen sind verschiedene klinische Abteilungen involviert, die häufig über eigene klinische Anwendungssysteme mit eigener Datenhaltung verfügen, den sogenannten Primärsystemen. Im relationalen SR-DWH werden Daten aus diesen und weiteren für die Abstoßungsdiagnostik relevanten klinischen Datenquellen des Krankenhauses integriert. Zu diesem Zweck wurde zunächst mittels Entity-Relationship-Modellierung ein Datenmodell für das SR-DWH entwickelt, das die Zusammenführung system- oder fachabteilungsspezifischer Datenstrukturen und Terminologien der zu berücksichtigenden Datenquellen ermöglicht und alle für die Abstoßungsdiagnostik verfügbaren, klinisch relevanten Informationen umfasst. Die Entwicklung dieses Datenmodells erfolgte ausgehend von einer Literaturrecherche zu für die Abstoßungsdiagnostik relevanten klinischen Parameter bzw. Datenattribute [12] in Zusammenarbeit mit den klinischen Expert*innen der Projektpartner. Aus der Literaturrecherche zu Datenmodellen mit Bezug auf die Nierenabstoßungsdiagnostik wurde eine erste Unterteilung des Modells in fünf semantisch übergeordnete Datenkategorien abgeleitet: Empfänger (recipient), Spenderorgan (graft), Spender (donor), Diagnostik (diagnosis) und Post-Transplantation (post transplant) sind die

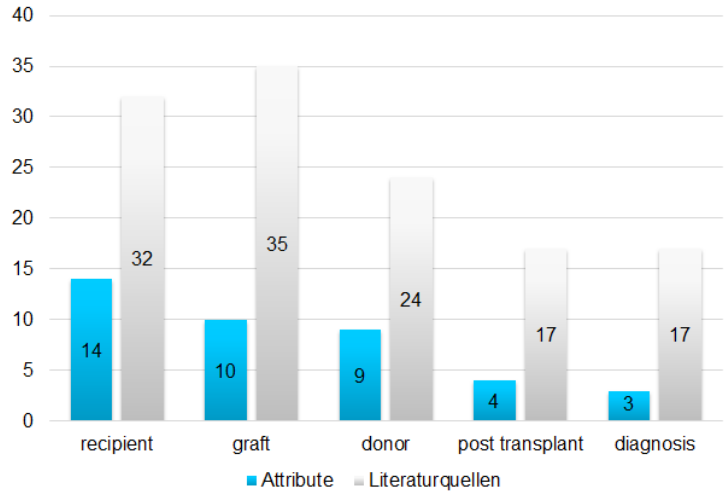


Abb. 1. Ergebnisse der Literaturrecherche: Übersicht über die Verteilung der Datenattribute auf die Datenkategorien und die damit assoziierten Literaturquellen (5 Kategorien , 40 Attribute und 125 Literaturstellen)

übergeordneten Kategorien, denen im weiteren Verlauf der Datenmodellierung die relevanten Datenattribute zugeordnet wurden (s. Abb. 1).

Auf diese fünf Kategorien verteilen sich 40 aus der Literaturrecherche abgeleitete klinisch relevante Attribute, die in ein erstes relationales Datenmodell überführt wurden. Zu den relevanten Attributen zählen Parameter der Patientenbehandlung wie die Primärdiagnose, das Alter bei Transplantation, die Immunsuppression und die spezifische Labordiagnostik. Bezüglich der befundspezifischen Abstoßungsdiagnostik werden zusätzlich die pathologische Begutachtung von entnommenen Gewebeproben und der immunologische HLA-Abgleich von Spender und Empfänger berücksichtigt. Die spezifische Labordiagnostik umfasst die Laborparameter Kreatinin, CRP-Spiegel, Hämoglobin, Albumin und die glomeruläre Filtrationsrate (GFR) sowie Ergebnisse der virologischen Diagnostik. Das Datenmodell wurde parallel zur Literaturrecherche im Rahmen mehrerer Expertenbefragungen validiert und erforderlichenfalls ergänzt und abschließend in ein relationales Datenbankschema des Data Warehouses überführt. Im Projektverlauf wurde das Datenbankschema im Rahmen der sukzessiven Integration von Primärsystemen weiter verfeinert.

Datenintegration allgemein Auf der Entwicklung des Datenmodells aufbauend wurde eine Systemanalyse der medizinischen Primärsysteme des klinischen Projektpartners durchgeführt, die zur Behandlung der nierentransplantierten Patient*innen genutzt werden. Hierbei wurden sieben relevante Datenquellen

und vorhandene Schnittstellen sowie in Bezug auf das Datenmodell relevante Datenfelder der Quellsysteme identifiziert.

Im Verlauf der Datenintegration wurden sukzessive alle sieben klinischen Datenquellen mit Extract-Transfer-Load-Prozessen (ETL-Prozessen) an das Screen-Reject Data-Warehouse-System (SR-DWH) angebunden. Soll das SR-DWH in einem anderen Systemkontext einer anderen Klinik zur Datenintegration genutzt werden, sind diese ETL-Prozesse entsprechend anzupassen.

Neben bereits voraggregierten Daten aus dem ECRDW der MHH konnten Daten der Studienteilnehmer*innen aus einem Spezialdokumentationssystem der Transplantationsambulanz, einem Befundsystem für Transplantationsimmunologie, dem klinischen Arbeitsplatzsystem Krankenhausinformationssystem (KAS), dem Dokumentenmanagementsystem (DMS), dem Laborinformationssystem (LIS) und dem Pathologieinformationssystem des Krankenhauses integriert werden. Neben den sieben relevanten Informationsquellen wurde der Kommunikationsserver des klinischen Partners zur Zusammenführung der relevanten klinischen Daten genutzt (s. Abb. 2)

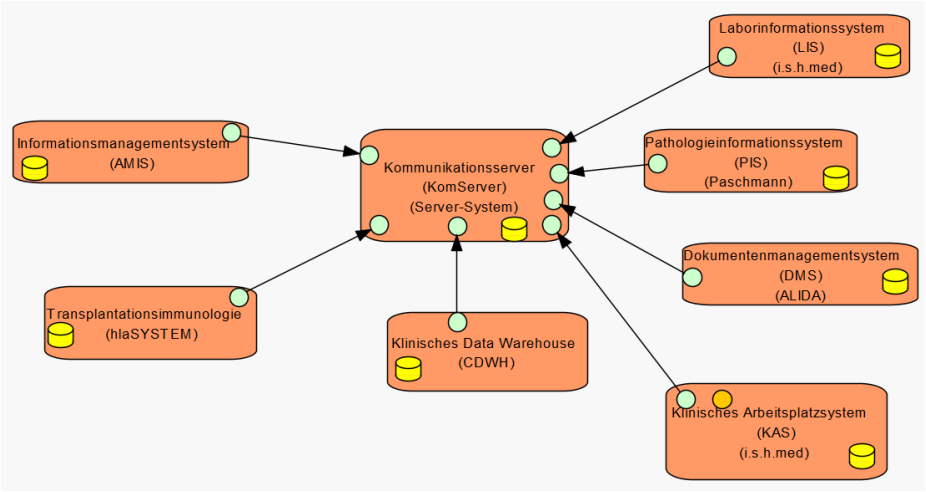


Abb. 2. Logisches Modell der für die Datenzusammenführung relevanten Quell- bzw. Primärsysteme gemäß 3LGM[2]–Methodik[2][24]

Das entwickelte Data Warehouse basiert auf der technologischen Plattform des ECRDW der MHH und nutzt entsprechend einen Microsoft SQL Server auf Basis von Microsoft Windows Server 2019. Für die Entwicklung der ETL-Prozesse wurden die SQL Server Data Tools für Microsoft Visual Studio genutzt.

[2] Three-layer Graph-based Meta Model für die Modellierung von Informationssystemen im Gesundheitswesen

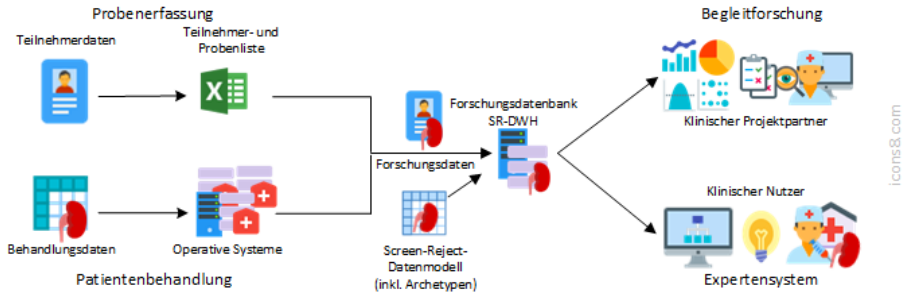


Abb. 3. Datenfluss im SR-DWH

Der Datenfluss folgt üblichen ETL-Abläufen (s. Abb. 3): Die Behandlungsdaten werden aus den operativen Quellsystemen zunächst nach Studienteilnehmer*innen gefiltert und dann transformiert und gegebenenfalls aggregiert in das SR-DWH überführt. Auf Basis des SR-DWH werden dann die weiteren Schritte zur Datenanalyse im Rahmen der Begleitforschung und Entwicklung des Expertensystems unternommen.

Text Mining Als klinisch relevant ergab die Anforderungsanalyse unter anderem die Banff-Klassifikation als Ergebnis der histologischen Beurteilung des Vorliegens und der Art der Abstoßungsreaktion auf Basis einer Gewebeprobe nach Biopsie sowie die Laborergebnisse der transplantationsimmunologischen HLA-Typisierung (Typisierung des Humanen Leukozytenantigen-Systems) von Spender und Empfänger. Konkret wurden folgende Anforderungen erfasst:

- Die Verteilung der Banff-Klassifikation über alle in die Studie eingeschlossenen Teilnehmenden soll automatisiert ermittelt werden können.
- Teilnehmende sollen anhand der Banff-Klassifikation selektiert werden können.
- Das SR-DWH soll automatisiert die dokumentierten HLA-Mismatches ermitteln können.
- Spender/Empfänger-Paare mit bestimmten HLA-Mismatches sollen selektiert werden können.

Da die relevanten Informationen in den betreffenden Primärsystemen in unstrukturierter Form als Freitexte oder Freitextsegmente dokumentiert vorliegen, war ein automatisiertes Selektieren der Zielvariablen anhand der Filterung von Wertebereichen nicht möglich. Aus diesem Grund wurden zur Informationsextraktion die ETL-Prozesse um eine Text Mining Funktionalität erweitert. Wie in Abbildung 4 gezeigt, wurden sämtliche unstrukturierte Quelldaten (Arztbriefe, Laborberichte) in einem Data Lake gesammelt. Innerhalb des Data Lakes wurden die Textdokumente kategorisiert und für jede Kategorie eine Verarbeitungspipeline zur Extraktion der Schlüsselinformationen konstruiert.

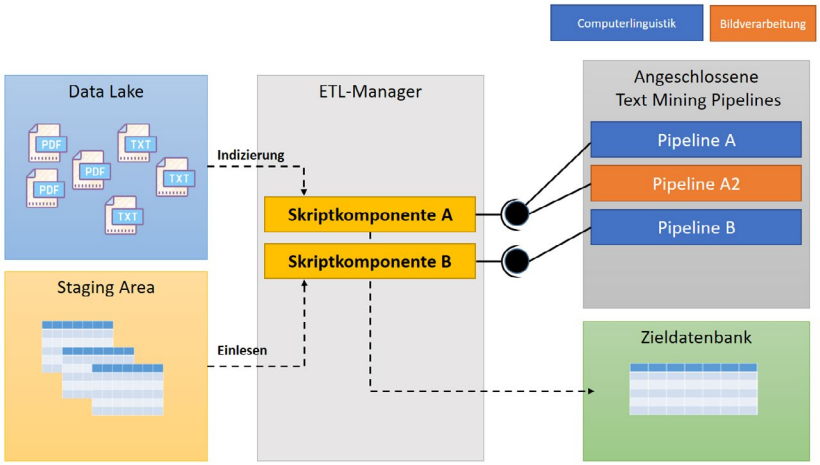


Abb. 4. Workflow zur Integration von Freitexten im Rahmen des zentralen ETL-Managers

Dabei kann es sich bei einer Verarbeitungspipeline um eine beliebige Softwarekomponente handeln, welche über eine definierte Schnittstelle mit einem Script-Task im Microsoft-basierten ETL-Manager verknüpft werden kann. In diesem Projekt wurden insgesamt drei Pipelines entwickelt, wobei die zweite dieser Pipelines auf computerlinguistischen Methoden, und die dritte Pipeline auf Methoden der digitalen Bildverarbeitung basiert. Der Bedarf der Bildverarbeitung ergab sich dadurch, dass einige Dokumente ausschließlich als eingescannte Bilddateien in einer Form vorliegen, für die eine Texterkennung mit OCR-Tools keine verwendbaren Ergebnisse produziert. Konkret konnten die verfügbaren OCR-Tools textuelle Inhalte zwar korrekt rekonstruieren, jedoch Strukturen wie beispielsweise Tabellen nicht ausreichend erfassen. Aus visuellen Gesichtspunkten können Tabellen als offensichtliche und klar abgrenzbare Dokumentationsstrukturen definiert werden. Daher wurden Tabellen als Bilddatei interpretiert und zunächst alle vertikalen und horizontalen Linien detektiert. Um die einzelnen Spalten abgrenzen zu können, wurden in dieser Folge die Schnittpunkte vertikaler und horizontaler Linien berechnet, welche in Abbildung 5 blau dargestellt sind. Dadurch konnten kohärente Informationen anhand ihrer Koordinaten im Bild erfasst werden: Abschnitte mit derselben X-Koordinate gehören zur selben Zeile, Abschnitte mit derselben Y-Koordinate zur selben Spalte. Im finalen Schritt der semantischen Textextraktion wurden die generierten Informationen in die entspechenden Staging-Tabellen des SR-DWH eingetragen (s. Abb. 6). Hierdurch konnten die initial ermittelten Anforderungen erfüllt werden.

Datenaufbereitung und Datenqualitätssicherung Damit die Daten im Hinblick auf die Datenanalyse verwendet werden können, müssen sie eine hierfür geeignete Form und eine ausreichende Datenqualität aufweisen. Um die origi-

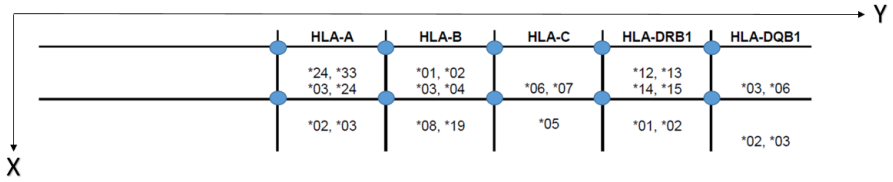


Abb. 5. Eingescannte Tabelle: Zellwerte sind Antigen-Stereotypen, die die HLA-Genorte in der Überschrift derselben Spalte erweitern. Jede Zeile beschreibt eine Probe einer Person (Empfänger oder Spender).

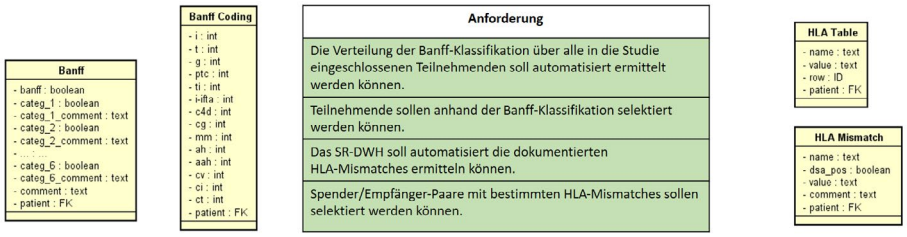


Abb. 6. Strukturierte Dokumentation von Schlüsselinformationen aus Freitexten

nalen Daten nicht zu verändern und die Nachvollziehbarkeit zu gewährleisten, wurden Sichten bzw. Views auf die Originaldaten erstellt. Im Rahmen eines Studierenden-Projektes wurden bereits erste Views entworfen [19], auf deren Grundlage weiter an der Datenaufbereitung und Verbesserung der Datenqualität gearbeitet werden konnte. Die Views basieren auf vordefinierten Abfragen, sodass die Ursprungsdaten nicht verändert werden. Die zu verwendenden Informationen wurden in Bezug auf die Datenqualität auf folgende Aspekte überprüft [3]: Vollständigkeit, Validität, Einheitlichkeit, Aktualität, Eindeutigkeit, Relevanz, Verständlichkeit (s. Abb. 7).

Für die Entwicklung des auf Machine Learning basierenden Prädiktors für das Vorliegen einer Abstoßung ist eine Voraussetzung, Informationen zum Vorliegen einer Abstoßungsreaktion auf das Nierentransplantat verfügbar zu machen. Das Vorliegen einer Abstoßung wurde aus der Banff-Klassifikation als Ergebnis der histologischen Untersuchung einer Gewebeprobe nach Biopsie abgeleitet. Da diese zentrale Information auf Basis von Text Mining-Verfahren aus den Pathologiebefunden abgeleitet wird (s.o.) und damit potenziell unsicher ist, wurde ein manueller Vergleich des Ergebnisses des Text Minings mit dem Befundtext vorgenommen.

Die Validität der Daten wurde zudem dahingehend geprüft, ob diese mit den für die Datenanalyse benötigten Datentypen kompatibel sind, so dass z.B. Datumsangaben auch mit dem Datentyp “Date” und Laborwerte mit dem Datentyp “Integer” bzw. “Float” hinterlegt sind. Außerdem wurde

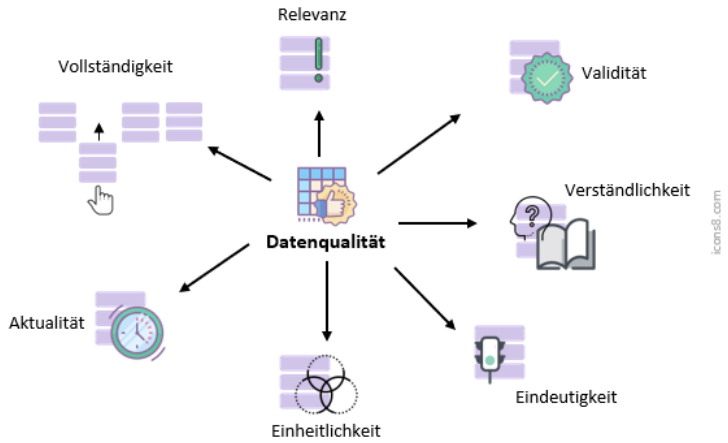


Abb. 7. Datenqualitätsbereinigung anhand von sieben Qualitätskriterien

eine Normierung der Laborwerte durchgeführt. Zudem wurden die Daten in Bezug auf ihre Relevanz für die Entwicklung des Prädiktionssystems zur Abstoßungsdiagnostik aufbereitet. So wurden Daten, welche nicht für die Analyse relevant waren, sowie doppelte Einträge aus den Tabellen beim Erstellen der Views entfernt (z.B. Angaben über das Quellsystem und Messwertzeiten, welche dieselbe Information abbildeten). Es existieren außerdem pro teilnehmender Person eine Vielzahl von Laboruntersuchungen und -werten, welche jedoch nicht alle für die weitere Datenanalyse relevant sind. Daher wurden beispielsweise nur Laborwerte selektiert, die aktuell zum Zeitpunkt der Biopsie erhoben wurden. Im Hinblick auf die Einheitlichkeit des Datenbestands wurden Laborwerte zu unterschiedlichen Probenarten (z.B. Urin und Serum) oder mit verschiedenen Maßeinheiten in verschiedenen Views zugänglich gemacht. Für die Verständlichkeit der Daten wurden unter anderem Spaltenbezeichner angepasst. Im Hinblick auf die Vollständigkeit der Daten wurden leere Felder (NULL-Werte) entfernt. Da sich einige Referenzwerte hinsichtlich des Geschlechts unterscheiden, können die entsprechenden Laborwerte nicht eindeutig miteinander verglichen werden. Daher wurden die entsprechenden Laborwerte um Informationen zu deren Verhältnis zum jeweiligen Referenzintervall ergänzt.

3.2 Datenintegration in die openEHR-Plattform

Durch die Nutzung des openEHR-Standards, einem internationalen Interoperabilitätsstandard für Aufbau und Austausch elektronischer Patientenakten kann der Austausch von Gesundheitsdaten zwischen verschiedenen Gesundheitsdienstleistern ermöglicht werden [7]. Zum Aufbau der openEHR-Plattform für das Screen-Reject-Projekt wurden Templates und Archetypen zur Abbildung

klinischer Begriffe und zur kontextbezogenen Datenzusammenstellung modelliert [21], mit deren Hilfe Daten aus dem SR-DWH in die openEHR-Plattform übertragen werden können (s. Abb. 8). Für die Entwicklung der Archetypen und der Zusammenstellung Templates wurden zwei Modellierungs-Tools von Ocean Health Systems verwendet: der Archetype Editor [16] und der Template Designer [17].

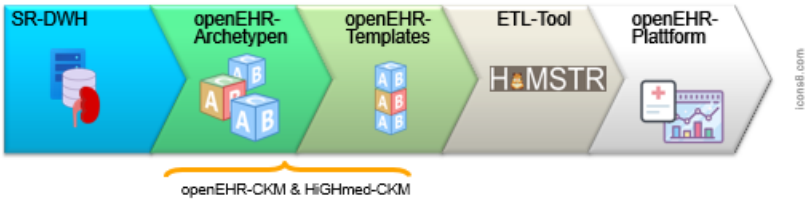


Abb. 8. Prozess der Template- und Archetypenentwicklung bis hin zur Datenübertragung in die openEHR-Plattform

In einem studentischen Projekt wurden vorbereitend Archetypen modelliert und zu einem umfangreichen Template zusammengesetzt [4]. Dies erschwerte jedoch die sukzessive automatisierte Übertragung der Daten in die EHR-Plattform, sodass mehrere inhaltlich voneinander abgrenzbare Templates modelliert wurden. Zur Verbesserung des Datenmodells wurden zunächst die bereits verwendeten Archetypen auf deren Eignung zur Abbildung der entscheidenden medizinischen Informationen geprüft, sowie die Verwendbarkeit mit Blick auf deren Aktualität eingeschätzt. Dazu wurde, durch die Projektverwandtschaft und bestehende Kooperation bedingt, insbesondere der Clinical Knowledge Manager (CKM) des HiGHmed-Projektes [25] verwendet, der ein umfangreiches Repository bereits modellierter Archetypen bietet. Dabei war es wichtig, dass die Archetypen für die Abbildung der relevanten klinischen Konzepte geeignet sind und alle benötigten Elemente beinhalten, sodass die Informationen aus dem SR-DWH dargestellt werden können. Des Weiteren wurde überprüft, inwieweit die Archetypen dem aktuellen Stand entsprechen, oder ob diese in der Zwischenzeit überarbeitet und aktualisiert wurden.

Da im HiGHmed-CKM für einige im SR-DWH befindlichen Daten keine passenden Archetypen vorhanden waren, wurden diese im Rahmen des Projektes modelliert und letztendlich für 15 medizinische Bereiche (Kreatinin, Albumin, CRP, GFR, Virologie, Immunologie, Pathologie, Spender, Empfänger, Versorgungsfall, Diagnosen, Hämoglobin, Transplantat) zu einzeln voneinander abgrenzbaren Templates zusammengesetzt. Im Projektverlauf sind zu den bisherigen im SR-DWH konsolidierten Daten noch weitere Informationen aus den Primär- bzw. Quellsystemen hinzugekommen. Diese Daten wurden bei der Modellierung der Archetypen und Templates berücksichtigt. Für zwei klinische Kontexte „HLA“ und „Medikation“ waren zum Zeitpunkt der Modellierung noch

keine konkreten Daten im Data Warehouse verfügbar [13], sodass für diese Bereiche entsprechende Muster-Entwürfe entwickelt wurden.

Die abschließend modellierten Templates wurden im Hinblick auf die automatisierte Datenübertragung in die openEHR-Plattform integriert und können dort weiterverarbeitet werden. Zudem gibt es die Möglichkeit, die Templates als Formulare abzuspeichern und so die Daten nicht automatisiert, sondern durch manuelle Eingabe in der Plattform zu speichern. Für die automatisierte Dateneingabe bzw. -übertragung vom SR-DWH in die openEHR-Plattform wurde das ETL-Tool „HAMSTRETL-Builder" [23] verwendet (s. Abb. 9). Diese Visual Basic-Anwendung kann mithilfe einer benutzerfreundlichen Oberfläche eine Verbindung zum SR-DWH herstellen und auf die im Data Warehouse befindlichen Daten zugreifen , um diese dann in die openEHR-Plattform zu übertragen. Die Anbindung an die openEHR-Plattform wird durch Nutzung einer REST-API-Schnittstelle und entsprechender Nutzerauthentifizierung gewährleistet. Mittels SQL-Syntax konnten die Daten aus dem Data Warehouse einzeln für jede*n Patient*in abgerufen bzw. selektiert werden.

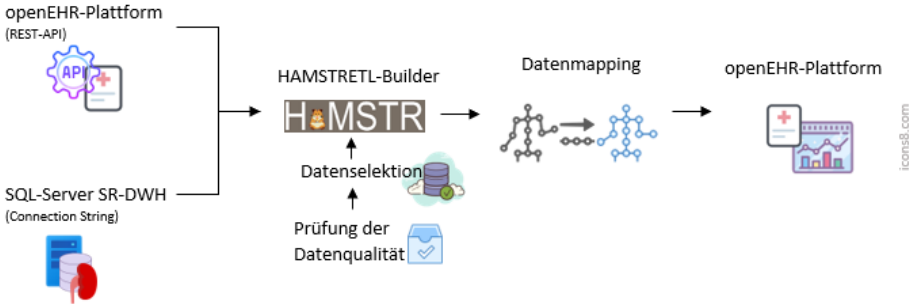


Abb. 9. Detaillierter Ablauf der Datenübertragung aus dem SR-DWH in die openEHR-Plattform mittels modellierter Templates

In einem letzten Schritt werden dann, entsprechend der selektierten Daten, die passenden in der openEHR-Plattform verfügbaren Templates ausgewählt und die Daten des SR-DWH auf die vorhandenen Elemente der Templates abgebildet bzw. gemappt. Um eine korrekte Identifikation des die Daten betreffende*n Patient*in in der openEHR-Plattform zu erreichen, muss ein Mapping der Patienten-ID im SR-DWH auf die EHR-ID der openEHR-Plattform erfolgen. Die EHR-ID ist hierbei die interne ID, welche von der openEHR-Plattform vergeben wird und den Identifikator der elektronischen Gesundheitsakte darstellt. Die Datenübertragung erfolgt dann automatisiert mittels der modellierten Templates vom SR-DWH in die Plattform.

3.3 Datenanalyse zur Prognose einer Abstoßungsreaktion

Aufbauend auf der Datenbasis des SR-DWH wurden konkrete Eigenschaften von Abstoßungen in kompakten Datenmengen festgehalten. Wie in Abbildung 10 exemplarisch dargestellt, wurde für jeden Wert und jede Kombination von Werten definiert, ob diese einer nachfolgenden Abstoßungsreaktion zuzuordnen sind oder nicht. Die Auswahl und Priorisierung der für die Abstoßungsdiagnostik relevanten klinischen Merkmale einer Abstoßungsreaktion wurde durch eine vorangehende Literaturrecherche vorbereitet [14].

Basierend auf diesen händisch kombinierten Merkmal- bzw. Featuresets wurde ein Benchmarking von etablierten maschinellen Lernverfahren (SVM, KNN, Random Forest Trees, Naive Bayes) durchgeführt und die für den Use Case der Vorhersage einer Abstoßungsreaktion am besten performenden Varianten ermittelt.

Zudem werden auch zeitliche Verläufe von Laborwerten mit künstlichen neuronalen Netzen auf Anomalien untersucht und auf diese Art als weitere Information in die Mustererkennung der Abstoßungserkennung integriert.

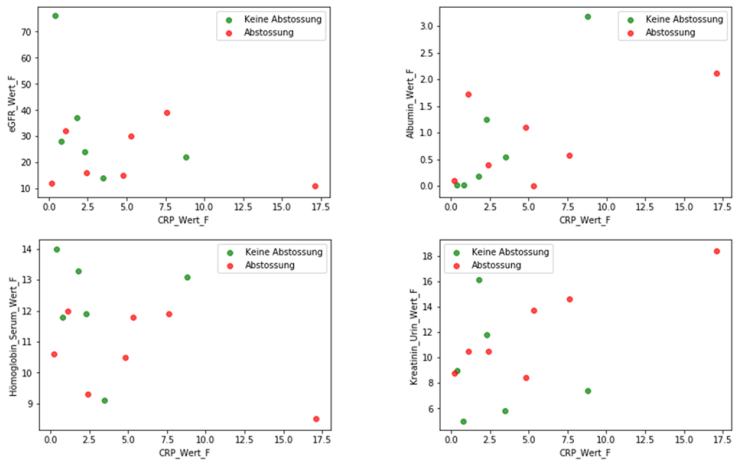


Abb. 10. Routinedaten als mögliche Marker Biomarker für eine Abstoßung. Entscheidungsgrenzen werden durch maschinelles Lernen ermittelt.

3.4 Datenvisualisierung für klinische Experten

Auf Grundlage der in der openEHR-Plattform verfügbaren relevanten Patientendaten und dem Ergebnis des Prädiktionssystems zur Wahrscheinlichkeit des Vorliegens einer Abstoßung, soll eine webbasierte Datenvisualisierung realisiert werden. Die openEHR-basierte elektronische Patientenakte wird zu diesem Zweck um den entsprechenden Prädiktionswert ergänzt und dieser

Wert zusammen mit den weiteren für die Abstoßungsdiagnostik relevanten Daten der Patientenakte grafisch aufbereitet den klinischen Experten in der Entscheidungssituation zur Abstoßungsdiagnostik verfügbar gemacht. So kann eine plattformunabhängige-, und auch interoperable Lösung zur Visualisierung von Daten geschaffen werden, welche die Entscheidungssituation eines klinischen Fachexperten bezüglich des potenziellen Vorliegens einer Abstoßungsreaktion optimal unterstützt.

Das auf Basis der openEHR-Implementierung der Firma Marand bereitgestellte Application Programming Interface (API) EhrScape bietet die Möglichkeit, Gesundheits-Apps einfach und in einer benutzerfreundlichen Umgebung zu entwickeln. Zudem beruht EhrScape auf dem openEHR-standard und erleichtert die Verknüpfung mit der Better-Plattform der Firma Marand [5]. Die Visualisierung soll mittels JavaScript, HTML und CSS implementiert werden, als Entwicklungsumgebung wird Visual Studio Code genutzt. Durch die Erweiterungen jQuery und Morris.js [5] besteht die Möglichkeit, zum einen mit der REST-API-Schnittstelle der openEHR-Akte mithilfe von JSON zu interagieren und zum anderen grafische Darstellungen, wie z.B. verschiedene Diagramme und Zeitverläufe darzustellen. Diese Diagramme umfassen unter anderem Laborverläufe oder auch Vitalparameter (s. Abb. 11).

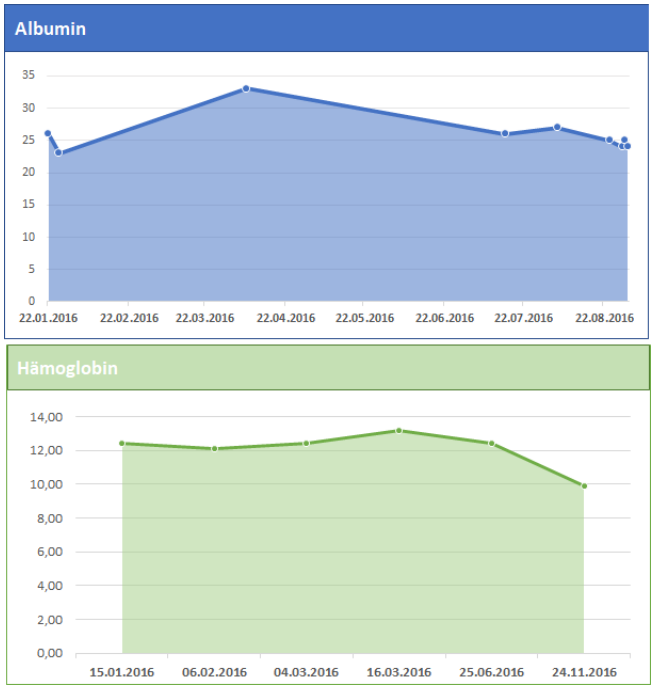


Abb. 11. Beispielhafte prototypische grafische Darstellung eines Laborparameterverlaufs

4 Ergebnisse zum aktuellen Projektstand

Der Entwicklungsstand des relationalen SR-DWH, einschließlich der programmierten ETL-Prozesse und des aktuellen Datenbestandes, wird in 4.1 beschrieben. Ausgehend von den entwickelten Datenstrukturen des relationalen SR-DWH ist eine interoperable openEHR-Plattform durch Bereitstellung von Archetypen und Templates vorbereitet und die Übertragung der Daten aus dem SR-DWH in die Plattform implementiert und getestet worden (s. 4.2).

Die Entwicklung der Prädiktionskomponente zur Abstoßungsdiagnostik konnte basierend auf implementierten Maßnahmen der Datenqualitätsbereinigung und Datenaufbereitung begonnen werden (s. 4.3). Die Entwicklung der Datenvisualisierungskomponente ist zum Zeitpunkt dieser Veröffentlichung gerade begonnen worden, weswegen hierzu noch keine Ergebnisse berichtet werden können.

4.1 Das SR-DWH

Die fünf im Datenmodell abgebildeten Kategorien beinhalten 40 Attribute, die zum aktuellen Zeitpunkt im Data Warehouse in insgesamt 66 Tabellen verarbeitet werden (s. Abb. 13). Für das vorbereitende Staging der Daten aus den Primärsystemen werden hiervon 44 Tabellen genutzt, die das initiale Laden und die Transformation ermöglichen.

Das SR-DWH führt aktuell Patientendaten von 108 Studienteilnehmer*innen aus klinischen Systemen zusammen, die zum Teil bereits im ECRDW der MHH voraggregiert wurden, aber auch Daten aus noch nicht integrierten Systemen werden für das SR-DWH berücksichtigt. Dabei handelt es sich unter anderem um Spezialdokumentationssysteme der transplantationsimmunologischen Befundung, sowie um eine Spezialanwendung zur Erfassung von Parameter zur Post-Transplantations-Überwachung in der Transplantationsambulanz. (s. Abb. 12).

Die Datenextraktion aus verschiedenen Quellsystemen ist dabei dadurch erschwert, dass wichtige Befundinformationen per Bildverarbeitung und Text Mining-Methoden aus 1700 heterogenen PDF-Befunden extrahiert und übertragen werden müssen [13]. Zum Berichtszeitpunkt ist die Datenzusammenführung noch nicht vollständig abgeschlossen, da ein für die Nachsorge der Patient*innen relevantes System während der Projektlaufzeit in eine neue Plattform mit neu zu entwickelnden Schnittstellen migriert wurde, wodurch sich der Zugang zu den Daten verzögerte.

Im Data Warehouse wurden 5 Katalogtabellen angelegt, um anhand von Kodierungen in den Data Warehouse-Tabellen Verknüpfungen zu analysespezifischen Metadaten zu ermöglichen: Diagnosekataloge (ICD10), Operations- und Prozedurenschlüssel (OPS), spezifische NTx-assoziierte Erkrankungskataloge und Laborkataloge zu allen durchgeführten Laboranalysen. So können beispielsweise durch Anbindung der Metadaten Diagnosen nach Gruppen geclustert- oder Zusatzinformationen (wie die Ergebnis-Einheit einer Untersuchung) zu bestimmten Laboranalysen assoziiert werden. Vorbereitet ist zudem

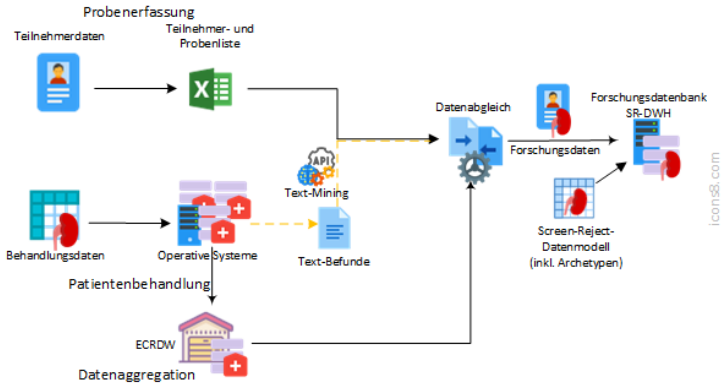


Abb. 12. Datenfluss im SR-DWH unter Nutzung des ECRDW der MHH

die Integration der LOINC-Kodierung in den Laborkatalogtabellen, die zur Verbesserung der Labordatenanalyse durch standardisierte Codierung genutzt werden kann, sobald die Labordaten aus den Quellsystemen entsprechend codiert vorliegen.

Das eigentliche Data Warehouse umfasst 11 (von 66) Tabellen (s. Abb. 13), die verschiedene Kategorien rund um die Nierenabstoßungsdiagnostik abbilden. 7 Tabellen davon bilden Ergebnisse von Laboruntersuchungen ab. Neben regulären Laborbefunden wie Albumin, dem C-reaktivem Protein (CRP), Hämoglobin und virologische Befunde werden auch die glomeruläre Filtrationsrate (GFR), Kreatinin und immunologische Befunde abgebildet. Zusätzlich zu den Labordaten werden Fallinformationen zum jeweiligen Aufenthalt und den durchgeführten Eingriffen (insb. Transplantationen und Biopsien) erfasst sowie übliche demografische Informationen wie Alter und Geschlecht der Patient*innen.

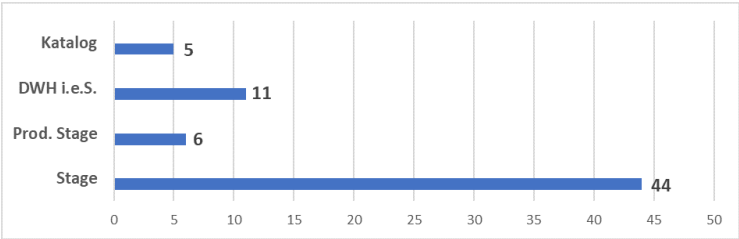


Abb. 13. Anzahl der Tabellen des SR-DWH in den Bereichen Staging, Prod. Stage, DWH i.e.S. und Kataloge

Neben den 44 Staging-Tabellen existieren 6 sogenannte "Produktive Staging Tabellen", die noch nicht abschließend in das fertige Data Warehouse überführt werden konnten (s. Abb. 13). Dabei handelt es sich unter anderem um befundspezifische Informationen wie Pathologiebefunde durchgeführter Biopsien und transplantationsimmunologische HLA-Befunde zum Abgleich der genetischen Voraussetzungen zwischen Spenderorgan und Empfänger. Die Tabellen haben aktuell einen Arbeitsstatus, sodass sie bereits für weiterführende Analysen genutzt werden können, jedoch noch nicht final in die abschließende DWH-Struktur übernommen sind. Zudem sind noch patientenspezifische Informationen zu den Diagnosen und zur Medikation (insb. Immunsuppression) zu transformieren.

Die Tabellen befinden sich in einem relationalen Aufbau und sind über verschiedene Schlüsselfelder in Beziehung zueinander gesetzt. Neben den klassischen Patientenidentifikatoren sind Labor- und Diagnosedaten über einheitliche Identifikatoren assoziiert, um eine möglichst umfangreiche Metadatenanreicherung zu ermöglichen.

Insgesamt wurden für die Datenintegration mit den SQL-Server-Data-Tools 18 Arbeitspakete mit 148 verschachtelten ETL-Strecken angelegt. Von 108 Studienteilnehmer*innen konnten damit insgesamt 8.784.495 Datensätze in das Data Warehouse integriert werden. Davon 81.591 Laborbefunde, 1.478 allgemeine Pathologiebefunde (davon 430 Biopsie-spezifische), 1.709 transplantationsimmunologische HLA-Befunde, 118 Transplantationsdaten, 11.713 Fall-bezogene Aufenthalte, 6.935 Diagnosen und 17.009 operative Eingriffe generell. Die Katalogtabellen umfassen zusammen 554.672 Datensätze.

Die Banff-Klassifikation wurde aus 218 Pathologieberichten und die HLA-Mismatches aus 1.700 PDF-Dateien extrahiert, aufbereitet und strukturiert im SR-DWH integriert, sodass auch diese Informationen in Datenanalysen einbezogen werden können.

4.2 Die openEHR-Plattform

Es wurde ein interoperables, auf openEHR basierendes semantisches Datenmodell als Datenintegrationsziel entwickelt, das als Ausgangsbasis zu weiteren Analyseschritten genutzt werden kann und aktuell Gegenstand weiterer klinischer Reviews ist. Dabei konnten mithilfe von Archetypen, teilweise aus dem Clinical Knowledge Manager (CKM) gespeist, 18 Templates modelliert werden. Fünf Archetypen wurden selbst entwickelt und bilden unter anderem die nephrologiespezifische „Banff"-Klassifikation ab (s. Abb. 14).

Für die Datenübertragung aus dem SR-DWH in die openEHR-Plattform wurden 18 Templates für verschiedene medizinische Bereiche entwickelt: Laboruntersuchungen wie Kreatinin, Albumin, Virologie, Hämoglobin, glomeruläre Filtrationsrate (GFR), C-reaktives Protein (CRP), Immunologie, HLA, Diagnosen wie Haupt- und Nebenerkrankung, Transplantat, Informationen zum Spender und Empfänger, Medikation/Immunsuppression, Pathologie und Versorgungsfall.

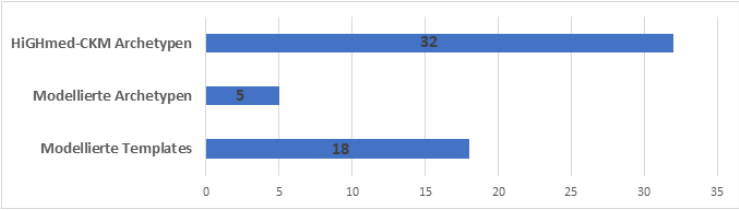


Abb. 14. Anzahl verwendeter Archetypen aus dem HiGHmed-CKM sowie ergänzend im Projekt modellierter Archetypen und Templates

Dabei wurden für die Templates der klinischen Kontexte "Transplantat" und "Medikation" zwei verschiedene Versionen angefertigt. Dies erfolgte zum einen, da die Abbildung der benötigten Inhalte mit mehreren unterschiedlichen Archetypen möglich war und zum anderen, weil im SR-DWH noch nicht final alle benötigten Daten, insbesondere aus den Spezialdokumentationssystemen (s. 3.1.2), zur Verfügung standen. Durch den nahezu gleichen Aufbau der Laboruntersuchungen konnte neben den einzelnen Templates der verschiedenen Blutwerte auch ein einzelnes universelles Template für die Abbildung von verschiedenen Laboranalyseergebnissen modelliert werden.

Die Templates wurden zum größten Teil mit bereits im HiGHmed-CKM vorhandene Archetypen modelliert. Insgesamt wurden 32 Archetypen aus dem CKM wiederverwendet. Fünf Archetypen (Auftragsinformationen, HLA-Details, Details zum Transplantat, Banff-Klassifikation und Erkrankungsdetail) wurden neu modelliert, da keine vergleichbaren Strukturen im CKM vorhanden waren, mit denen die Abbildung der benötigten Informationen möglich war.

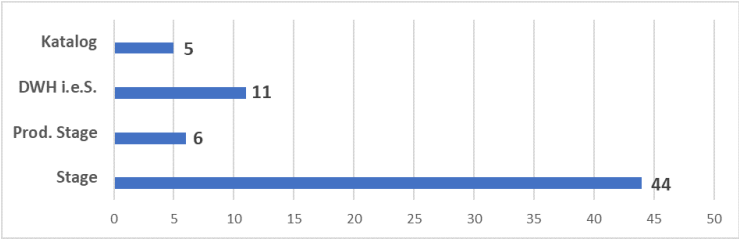


Abb. 15. Übersicht über die übertragenen Daten aus dem SR-DWH in die openEHR-Plattform mittels ETL-Tool

Nach der Zusammenstellung der Templates konnten die medizinischen Daten mittels ETL-Tool als Proof-of-concept aus dem SR-DWH in die openEHR-Plattform übertragen werden. Für sechs Patient*innen wurden unter anderem Pathologiebefunde mit Angabe der nephrologischen Banff-Klassifikation und Laborbefunde für die Blutwerte Albumin und Kreatinin sowie virologische Laborbefunde übertragen. Insgesamt konnten auf diese Weise

125 Albumin-Datensätze, 136 Virologie-Datensätze, 10 Pathologie-Datensätze und 534 Kreatinin-Datensatze übertragen und abgespeichert und so das Datenübertragungskonzept erfolgreich getestet werden (s. Abb. 15).

4.3 Erste Ergebnisse der Datenanalyse zur Prognose einer Abstoßungsreaktion

Im Zuge der Vorbereitung der Datenanalyse erwies sich die Aufbereitung der Datenqualität und Datenaggregation der Routinedaten als erheblich aufwändig. Vorbereitende Entwicklungsarbeiten der Datenanalyse mittels maschinellen Lernverfahren auf Basis des SR-DWH sind bereits erfolgt, in ersten Experimenten konnten 77% der getesteten Fälle auf Basis der vorhandenen Patientendaten korrekt klassifiziert werden. Eine Weiterentwicklung des Prädiktionsalgorithmus zur Erhöhung der Performance der Abstoßungserkennung stellt den wesentlichen Fokus der weiteren Forschungstätigkeiten dar.

5 Diskussion und Ausblick

Nach Entwicklung des relationalen Datenmodells konnten die als relevant identifizierten Daten zum größten Teil in das SR-DWH integriert werden. Die Integration des Datenbestandes eines spezifischen Primärsystems konnte aufgrund einer Systemumstellung innerhalb der MHH erst verzögert begonnen werden und dauert noch an. Zum jetzigen Zeitpunkt wird der Datenbestand dieses Systems analysiert und schrittweise mit dem Ziel integriert, die darin enthaltenen relevanten Daten bis zum Projektende in das SR-DWH übernehmen zu können. Insgesamt hat sich die Datenintegration und Sicherstellung einer hinreichenden Datenqualität als deutlich aufwändiger dargestellt, als im Rahmen der Projektplanung angenommen. Vor allem nicht strukturiert in Freitexten vorliegende Daten haben den Aufwand der Informationsextraktion erheblich erhöht, da hierfür auf Text Mining Methoden zurückgegriffen werden musste und die entwickelten Text Mining Algorithmen zusätzlich in Bezug auf die Güte der Extraktionsqualität durch manuelle Überprüfung der Freitexte aufwändig evaluiert werden mussten.

Insgesamt ist mit den entwickelten ETL-Prozessen ein Framework entstanden, das auf andere Kliniken übertragen werden kann, aber je nach in einer Klinik vorhandenen Primär- und/oder EPA-Systemen daran anzupassen ist. Mit dem Ziel, diesen Anpassungsaufwand gering zu halten und die Datenanalyse zudem auf Grundlage eines semantisch interoperablen Datenmodells durchführen zu können, wurde zusätzlich zur Datenintegration in dem relationalen SR-DWH auf eine Datenbereitstellung in einem auf openEHR basierenden EPA-System fokussiert. Nach Abschluss der Modellierung bzw. Bereitstellung der zur Abbildung des relevanten Datenbestandes erforderlichen Archetypen und Templates und deren Bereitstellung in einer openEHR-Plattform können mithilfe eines ETL-Tools die relevanten medizinischen Daten aus dem SR-DWH

in eine openEHR-Plattform übertragen und darin bereitgestellt werden. Verwendet eine Klinik bereits eine openEHR-basierte elektronische Patientenakte erübrigt dieser Zwischenschritt der Übertragung über das relationale SR-DWH. Durch die Verwendung des auf semantische Interoperabilität ausgerichteten openEHR-Standards, der auch bei der Entwicklung und Bereitstellung klinischer Daten im Medizinisches Datenintegrationszentrum (MeDIC) des HiGHmed-Konsortiums der Medizininformatikinitiative Verwendung findet [1], sowie insbesondere der modellierten und über den HiGHmed-CKM öffentlich zugänglichen openEHR-Templates und Archetypen können klinisch verwandte Projekte wie Nephro-Digital [19], Rocket [15] oder die in Entwicklung befindliche Forschungsplattform des Zukunftslabors Gesundheit des Zentrums für digitale Innovation Niedersachsen [26] von diesen Entwicklungen profitieren.

Aktuell wird auf Grundlage des Datenbestandes des SR-DWH weiterhin an der Identifikation der für die Prädiktion der Abstoßungsdiagnostik geeigneten Machine Learning Modelle und der Evaluation der Genauigkeit der Prädiktion gearbeitet. Wird eine hinreichende Prädiktionsqualität erreicht, wird das Prädiktionsmodell implementiert und als Bestandteil der aktuell ebenfalls noch in Entwicklung befindlichen Datenvisualisierungsanwendung zusammen mit den weiteren, für die Abstoßungsdiagnostik relevanten, Patientendaten im Sinne eines Expertensystems visualisiert. Für einen Einsatz des Expertensystems in der Routine sind im Nachgang des Projekts weitere Schritte der Validierung der Software gemäß Medizinprodukterichtlinie erforderlich.

Danksagung

Wir danken unseren Verbundprojektpartnern aus dem Institut für technische Chemie der Leibniz Universität Hannover der ebd. tätigen Verbundprojektkoordinatorin Prof. Dr. Cornelia Blume sowie dem Institut für Transfusionsmedizin der Medizinischen Hochschule Hannover und dem dortigen Teilprojektleiter Prof. Dr. Stephan Immenschuh. Wir danken weiterhin den klinischen Projektpartnern aus der Klinik für Nieren- und Hochdruckerkrankungen der MHH unter Leitung von Prof. Dr. Hermann Haller sowie den technischen Projektpartnern aus dem Zentrums für Informationsmanagement der MHH und speziell dessen Abteilung Systeme für Forschung und Lehre unter der Leitung von Dr. Svetlana Gerbel für die Unterstützung in Fragen der Datenintegration auf Basis des Enterprise Clinical Research Data Warehouse (ECRDW) der MHH. Weiterhin danken wir den Projektpartnern des Peter L. Reichertz Instituts für Medizinische Informatik der MHH unter der Leitung von Prof. Dr. Michael Marschollek für die Unterstützung in wissenschaftlichen und technischen Fragen die openEHR-Plattform und diesbezügliche Datenmodellierung und Datenübertragung betreffend. Last but not least danken wir dem Transplantationszentrum der MHH für die Unterstützung in Fragen der Integration von Spezialsystemen zur Transplantationsnachsorge sowie allen in der obigen Zusammenstellung ggf. nicht genannten Personen, die das Projekt unterstützt haben.

Das Projekt wird im Rahmen des EFRE-Förderprogramms mit Mitteln der EU und des Landes Niedersachsen gefördert. (Laufzeit bis 12/2020)

Die Hochschule Hannover ist Mitglied des Zukunftslabors Gesundheit, gefördert vom Niedersächsischen Ministerium für Wissenschaft und Kultur unter Fördernummer ZN3491 im Niedersächsischen Vorab der Volkswagen Stiftung und betreut vom Zentrum für digitale Innovationen (ZDIN).

Literatur

1. Highmed - an open platform approach to enhance care and research across institutional boundaries. Methods of information in medicine **57**(S 01), e66–e81 (2018). https://doi.org/10.3414/ME18-02-0002
2. Al-Awwa, I.A., Hariharan, S., First, M.R.: Importance of allograft biopsy in renal transplant recipients: correlation between clinical and histological diagnosis. American journal of kidney diseases : the official journal of the National Kidney Foundation **31**(6 Suppl 1), S15–8 (1998). https://doi.org/10.1053/ajkd.1998.v31.pm9631859
3. Apel, D., Behme, W., Eberlein, R., Merighi, C.: Datenqualität erfolgreich steuern: Praxislösungen für Business-Intelligence-Projekte. Safari Tech Books Online, dpunkt.verl., Heidelberg, 3., überarb. und erw. aufl. edn. (2015), `http://proquest.safaribooksonline.com/9781457193408`
4. Breckner, A., Kiss, S., Wulff, A., Katzensteiner, M., Bott, O.J.: Herleitung von openehr-archetypen im kontext von abstoßungsdiagnostik nach nierentransplantation (hakan). https://doi.org/10.3205/18GMDS177
5. ehrscape.com: ehrscape.com: Flexible gui building blocks (2014), `https://www.ehrscape.com/gui-examples.html`
6. Einecke, G., Sis, B., Reeve, J., Mengel, M., Campbell, P.M., Hidalgo, L.G., Kaplan, B., Halloran, P.F.: Antibody-mediated microcirculation injury is the major cause of late kidney transplant failure. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons **9**(11), 2520–2531 (2009). https://doi.org/10.1111/j.1600-6143.2009.02799.x
7. Garde S.: Semantisch interoperable elektronische gesundheitsakten auf basis des openehr-ansatzes: einfach archetypisch! **8**(4), 100–104 (2006), `https://docplayer.org/18427098-Forum-der-medizin-dokumentation-und-medizin-informatik-jahrgang-8.html`
8. Gerbel, S., Laser, H., Schönfeld, N., Rassmann, T.: The hannover medical school enterprise clinical research data warehouse: 5 years of experience. In: Auer, S., Vidal, M.E. (eds.) Data Integration in the Life Sciences: 13th International Conference, DILS 2018, Hannover, Germany, November 20-21, 2018, Proceedings, Lecture Notes in Bioinformatics, vol. 11371, pp. 182–194. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-06016-9\textunderscore
9. Gondos, A., Döhler, B., Brenner, H., Opelz, G.: Kidney graft survival in europe and the united states: strikingly different long-term outcomes. Transplantation **95**(2), 267–274 (2013). https://doi.org/10.1097/TP.0b013e3182708ea8
10. organspende info.de: Statistiken zur organspende für deutschland und europa (2019), `https://www.organspende-info.de/zahlen-und-fakten/statistiken.html`
11. Katzensteiner, M., Zubke, M., Blume, C., Immenschuh, S., Gerbel, S., Schönfeld, N., Marschollek, M., Haller, H., Ludwig, W., Bott, O.: Screen Reject – Klinisches Data Warehouse zur Abstoßung nach Nierentransplantation – erste Schritte. In: 63.

Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V.(GMDS). German Medical Science GMS Publishing House (2018). https://doi.org/10.3205/18GMDS052
12. Katzensteiner, M., Ludwig, W., Marschollek, M., Bott, O.J.: Results of a literature review to prepare data modelling in the context of kidney transplant rejection diagnosis. Studies in health technology and informatics **258**, 179–183 (2019)
13. Katzensteiner, M., Zubke, M., Blume, C., Immenschuh, S., Gerbel, S., Schönfeld, N., Marschollek, M., Haller, H., Ludwig, W., Bott, O.J.: Zusammenführung klinischer Daten im Screen Reject Datawarehouse zur Optimierung der Abstoßungsdiagnostik nach Nierentransplantation. https://doi.org/10.3205/19GMDS177
14. Knöfel R, Köhler S, Leonhardt D, Rudolph L, Uthe M, Van den Boom A.: MMI-Projekt: Künstliche Intelligenz in der Analyse von Behandlungsdaten für die klinische Forschung. Internes projektdokument, Hochschule Hannover, Hannover (2020)
15. era learn.eu: Project: Reclassification using omics integration in kidney transplantation (rocket) (2020), `https://www.era-learn.eu/network-information/networks/eracosysmed/2nd-joint-transnational-call-for-european-research-projects-on-systems-medicine/reclassification-using-omics-integration-in-kidney-transplantation-rocket`
16. oceanhealthsystems.com: Archetype editor | ocean health systems (1999-2020), `https://oceanhealthsystems.com/applications/archetype-editor/`
17. oceanhealthsytems.com: Template designer | ocean health systems (1999-2020), `https://oceanhealthsystems.com/applications/template-designer/`
18. openEHR Foundation: openehr: Open industry specifications, models and software for e-health (2020), `https://www.openehr.org/`
19. Pape, L., Schneider, N., Schleef, T., Junius-Walker, U., Haller, H., Brunkhorst, R., Hellrung, N., Prokosch, H.U., Haarbrandt, B., Marschollek, M., Schiffer, M.: The nephrology ehealth-system of the metropolitan region of hannover for digitalization of care, establishment of decision support systems and analysis of health care quality. BMC medical informatics and decision making **19**(1), 176 (2019). https://doi.org/10.1186/s12911-019-0902-0
20. Patel, M.D., Young, S.W., Scott Kriegshauser, J., Dahiya, N.: Ultrasound-guided renal transplant biopsy: practical and pragmatic considerations. Abdominal radiology (New York) **43**(10), 2597–2603 (2018). https://doi.org/10.1007/s00261-018-1484-5
21. Schewe, N., Katzensteiner, M., Bott, O.J.: Archetypen-basierte Modellierung von Daten zur Abstoßungsdiagnostik nach Nierentransplantation (NTx). In: Tagungsband der 65. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS). German Medical Science GMS Publishing House (Zur Veröffentlichung angenommen)
22. Tute, E. (ed.): OpenEHR platform concept for semantic interoperability in project HaMSTR, Biomedical Engineering Biomedizinische Technik Joint journal of the German Society for Biomedical Engineering in VDE and the Austrian and Swiss Societies for Biomedical Engineering, vol. s1. De Gruyter, Swiss (2016)
23. Tute, E., Haarbrandt, B.: Integrating relational data into clinical information model based data repositories. In: 62. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e. V.(GMDS). German Medical Science GMS Publishing House (2016). https://doi.org/10.3205/17GMDS145
24. Winter, A., Brigl, B., Wendt, T.: Modeling hospital information systems. part 1: The revised three-layer graph-based meta model 3lgm2. Methods of information in medicine **42**(5), 544–551 (2003)

25. Wulff Antje, Sommer Kim Katrin, Ballout Sarah, HiGHmed Consortium, Haarbrandt Birger, Gietzelt Matthias: A report on archetype modelling in a nationwide data infrastructure project. Studies in health technology and informatics **258**, 146–150 (2019). https://doi.org/10.3233/978-1-61499-959-1-146
26. zdin.de: Zukunftslabor Gesundheit (27102020), `https://www.zdin.de/zukunftslabore/gesundheit`
27. Zubke, M., Katzensteiner, M., Bott, O.J.: Integration of unstructured data into a clinical data warehouse for kidney transplant screening - challenges & solutions. Studies in health technology and informatics **270**, 272–276 (2020). https://doi.org/10.3233/SHTI200165

Unsere Autoren

Volker Ahlers ist Professor für Simulation und Mathematik an der Fakultät IV, Abteilung Informatik der Hochschule Hannover. Er ist Gründungsmitglied des Forschungsclusters Smart Data Analytics und leitet gemeinsam mit Felix Heine und Carsten Kleiner die Forschungsgruppe Trust@HsH. Nach Studium und Promotion in Physik an den Universitäten Göttingen, Edinburgh und Potsdam arbeitete er mehrere Jahre in der Software-Entwicklung für Anwendungen der Bildverarbeitung und Mustererkennung, bevor er 2005 an die Hochschule Hannover berufen wurde. In Lehre und Forschung beschäftigt er sich mit den Themen Computergrafik, Visualisierung, Machine Learning, Mathematik und Statistik.

Oliver J. Bott promovierte 2001 zum Dr.-Ing. im Fach Medizinische Informatik an der Technischen Universität Braunschweig. Seit 2007 ist er Professor für Medizinische Informatik an der Hochschule Hannover (HsH). Von 2011 bis 2013 war er Gründungsdirektor des E-Learning Services Center und von 2014 bis 2020 CIO und ab 2016 bis 2020 Vizepräsident für Forschung, IT und Informationsmanagement an der HsH. Seine Forschung und Lehre konzentriert sich auf Informationssysteme im Gesundheitswesen und in der klinischen Forschung speziell im Hinblick auf Fragen der Integration, des Managements und der Analyse klinischer Daten. Zudem forscht er zu digital unterstützter Lehre in Medizin und Medizinischer Informatik. Er ist Gründungsmitglied des Forschungsclusters Smart Data Analytics der HsH, in dem fakultätsübergreifend an Fragen zu Data Integration, Data Management, Data Analytics & Visualization sowie Data Security & Privacy geforscht wird. 2010 erhielt er den Wissenschaftspreis für Forscher an Fachhochschulen des niedersächsischen Ministeriums für Wissenschaft und Kultur; 2009 erhielt er einen Summa cum Laude Award und einen Excellence in Design Award der Radiological Society of North America.

Ralf Bruns ist Professor für Softwareentwicklung an der Abteilung Informatik der HsH. Er leitet gemeinsam mit Jürgen Dunkel die Forschungsgruppe Softwarearchitektur und ist stellvertretender Sprecher des Forschungsclusters Smart Data Analytics. Als promovierter Informatiker lehrt er in den Gebieten Software Engineering und Intelligente Systeme. Seine Forschungsschwerpunkte sind komplexe Softwarearchitekturen, Datenstromverarbeitung und Schwarmalgorithmen.

Bianca Frenzel befindet sich zurzeit als Studierende des Bachelorstudiengangs Medizinisches Informationsmanagements an der Hochschule Hannover im siebten Semester mit den Vertiefungsrichtungen „Management klinischer Studien und Statistik“ und „Methoden der Medizinischen Informatik“. Während des Studiums absolvierte sie ein Praktikum im Bereich der Softwareentwicklung bei CGI. Seit Frühjahr 2020 befindet sie sich als studentische Hilfskraft im Peter L. Reichertz Institut an der Medizinischen Hochschule Hannover im Projekt Zukunftslabor Gesundheit und befasst sich dort mit dem Thema Modellierung von Gesundheitsdaten und schreibt im Rahmen des Screen-Reject Data Warehouse Projektes ihre Bachelorarbeit im Bereich webbasierte Visualisierung für klinische Experten.

Felix Heine hat u.a. als IT-Berater und Consultant im Bereich DataWarehousing gearbeitet. Er wurde an der Universität Paderborn promoviert und ist seit 2010 Professor für Datenbanken und Informationssysteme an der Abteilung Informatik der Hochschule Hannover. Seine Forschungsinteressen liegen im Bereich Datenbanken, Datenqualität und Datenanalysen. Er ist Co-Leiter der Arbeitsgruppe Trust@HsH und Mitbegründer des Forschungsclusters Smart Data Analytics. Er war und ist an zahlreichen Forschungsprojekten beteiligt, u.a. zum Thema Datenqualität, Data Profiling sowie Datenanalysen, insb. im Umfeld der IT-Sicherheit.

Bastian Hellmann hat sein Studium an der Hochschule Hannover als Master of Science in der Vertiefungsrichtung Graphik und Visualisierung im Jahr 2010 abgeschlossen. Seit 2008 arbeitete er zunächst als wissenschaftliche Hilfskraft und ab Ende 2010 als wissenschaftlicher Mitarbeiter an der Fakultät IV, Abteilung Informatik der Hochschule Hannover im Rahmen der Forschungsgruppe Trust@HsH. Hier hat er sich in den Forschungsprojekten VisITMeta und SIMU mit der Visualisierung von Metadaten zur Netzwerksicherheit beschäftigt.

Matthias Katzensteiner schloss sein Studium des Medizinischen Informationsmanagements mit der Vertiefungsrichtung „Medizinische Informatik“ 2016 mit dem Master ab. Im Laufe seines Studiums arbeitete er als wissenschaftliche Hilfskraft im Institut für Humangenetik der Medizinischen Hochschule Hannover sowie bei der Magrathea Informatik GmbH in Hannover. Sowohl in Abschlussarbeiten als auch in studentischen Projekten befasste sich Herr Katzensteiner vertieft mit den Themen Datenbanken, Data Warehousing und Datenintegration. Seit Juli 2017 ist Herr Katzensteiner als wissenschaftlicher Mitarbeiter an der Fakultät III der Hochschule Hannover tätig. Beginnend mit dem Forschungsprojekt Screen-Reject Data Warehouse ist er seit 2020 zudem im Projekt Zukunftslabor Gesundheit des ZDIN beschäftigt.

Carsten Kleiner hat u.a. als Softwareentwickler, IT- Berater und Softwareprojektleiter im Bereich der Datenbankanwendungssysteme gearbeitet. Er wurde an der Universität Hannover im Jahre 2002 promoviert und ist seit Ende 2004 Professor für Sichere Informationssysteme an der Abteilung Informatik der Hoch- schule Hannover. Seine Forschungsinteressen liegen im Bereich Datenbanksysteme und -anwendungen, Datenqualität und Analyse von Daten. Er ist Co-Leiter der Arbeitsgruppe Trust@HsH und Mitbegründer des Forschungsclusters Smart Data Analytics. Er war und ist an zahlreichen Forschungsprojekten beteiligt, u.a. zum Thema Datenqualität, Datenanalysen und Datenmanagement insb. im Umfeld der IT-Sicherheit.

Kim Jana Nörtemann befindet sich zurzeit als Studierende des Bachelorstudiengangs Medizinisches Informationsmanagement an der Hochschule Hannover im siebten Semester. Sie absolvierte ihr Praktikum im Zentrum für Informationsmanagement der Medizinischen Hochschule Hannover in der Abteilung Systeme für Forschung und Lehre und befasste sich dort bereits mit dem Forschungsprojekt Screen-Reject Data Warehouse. Seit Dezember 2019 arbeitet sie als studentische Hilfskraft für die Hochschule Hannover weiterhin an dem Projekt und beschäftigt sich vor allem mit der Datenqualitätsbereinigung und der Integration relevanter Forschungsdaten, worüber sie auch ihre Bachelorarbeit schreiben wird.

Thomas Oelsner hat den Master of Science Abschluss in den Vertiefungsrichtungen IT Sicherheit und Informationssysteme in 2015 an der Hochschule Hannover absolviert. Während des Studiums arbeitete er in der Forschungsgruppe Trust@HsH, in welcher er nach dem Abschluss als wissenschaftlicher Mitarbeiter übernommen wurde. Dort begleitete er ein Forschungsprojekt das sich mit Datenanalyse und Sicherstellung von Datenqualität befasste. Seit 2018 ist er als Security Consultant im Bereich Security Assessment von IT Systemen mit Schwerpunkt BI und Cloud tätig.

Nina Schewe hat ihr Masterstudium Medizinisches Informationsmanagement mit den Vertiefungsrichtungen „Klinisches Datenmanagement“ und „Methoden der Medizinischen Informatik“ an der Hochschule Hannover im Jahr 2020 abgeschlossen. Während des Studiums arbeitete sie als studentische Hilfskraft in der Verkehrsunfallforschung und im Institut für Epidemiologie, Sozialmedizin und Gesundheitssystemforschung der Medizinischen Hochschule Hannover. Während der Abschlussarbeit befasste sie sich mit dem Thema Interoperabilität im Forschungsprojekt Screen-Reject Data Warehouse. Seit Frühjahr 2020 ist sie als wissenschaftliche Mitarbeiterin in den Projekten Screen-Reject Data Warehouse der Hoch- schule Hannover und NEPHRO-DIGITAL des Peter L. Reichertz Instituts der Medizinischen Hochschule Hannover beschäftigt und befasst sich neben der Interoperabilität mit den Themen Datenqualitätsbereinigung und Datenmanagement.

Fabian Schmieder studierte Rechtswissenschaften an der Juristischten Fakultät der Leibniz Universität Hannover. Er war wissenschaftlicher Mitarbeiter am Institut für Rechtsinformatik der Leibniz Universität und promovierte an der Schnittstelle zwischen Strafprozessrecht, Datenschutzrecht und Medienrecht. Seit 2015 ist der Professor für Medienrecht an der HsH und lehrt und forscht insbesondere im Bereich des Datenschutz- und Urheberrechts. Am 1. September 2020 begann seine vierjährige Amtszeit als nebenberuflicher Vizepräsident für Digitalisierung und IT.

Christian Wartena studierte Computerlinguistik in Nijmegen (Niederlande) und Potsdam, wo er 1999 promovierte. Er arbeitete für Unternehmen in Heidelberg und Enschede (Niederlande). Seit 2011 ist er Professor für Sprach- und Wissensverarbeitung an der HsH und lehrt in den informationswissenschaftlichen Studiengängen der Fakultät III sowie im Master Data Analytics an der Universität Hildesheim. Er leitet verschiedene Projekte im Bereich Sprachverarbeitung und ist seit 2020 Sprecher des Forschungsclusters Smart Data Analytics.

Maximilian Zubke studierte an der Hochschule Hannover Angewandte Informatik mit den Vertiefungsrichtungen Computergrafik und Informationssysteme, bevor er 2016 als wissenschaftlicher Mitarbeiter eingestellt wurde. Während seiner Laufbahn war er an mehreren Projekten des Forschungsclusters Smart Data Analytics beteiligt wo er sich vor allem mit der automatisierten Analyse von Freitexten und Bilddaten beschäftigt hat. Seit April 2018 ist er Doktorand der Medizinischen Hochschule Hannover, wo er Systemarchitekturen zur robusten Bereitstellung und Selektion von Deep Learning basierten Analyseverfahren in der Radiologie erforscht.

HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS