UNIVERSITÄT ZU LÜBECK
INSTITUT FÜR MEDIZINISCHE INFORMATIK

Aus dem Institut für Medizinische Informatik der Universität zu Lübeck
Direktor: Prof. Dr. rer. nat. habil. Heinz Handels

Standardisierte Metadatenintegration für die Sekundärnutzung klinischer Daten

Inauguraldissertation
zur
Erlangung der Doktorwürde
der Universität zu Lübeck

Aus der Sektion Informatik/Technik

vorgelegt von
Hannes Ulrich
aus Brandenburg/Havel

Lübeck, 2022

1. Berichterstatter: Prof. Dr. rer. nat. habil. Josef Ingenerf

2. Berichterstatter: Prof. Dr. rer.nat. habil. Sven Groppe

Tag der mündlichen Prüfung: 15.03.2022

Zum Druck genehmigt: Lübeck, 16.03.2022

Kurzfassung

Durch die voranschreitende Digitalisierung und den Einzug neuer Technologien in unser Gesundheitswesen entstehen immer neue Datenquellen. Ein ganzheitlicher Überblick über den rasant wachsenden Datenbestand kann den medizinischen Akteuren neue individuelle Therapiemöglichkeiten für jeden Patienten bieten und die klinische Forschung beschleunigen. Eine gewinnbringende und effektive Datenverwendung setzt jedoch voraus, dass die uneinheitlichen Quellen in Bezug gesetzt werden können. Dieser notwendige Prozess wird insbesondere durch die stark heterogene IT-Landschaft des Gesundheitswesens erschwert. Diese Voraussetzungen erfordern es, dass Datenquellen zunächst verwendbar gemacht werden, indem sie in die Abläufe integriert werden. Diese Datenintegration und der damit zusammenhängende Datenaustausch wird mit der zunehmenden Anzahl verteilter Systeme und der ständig wachsenden Datenmenge immer wichtiger. Als mögliches Integrationswerkzeug sind Metadaten in der Lage, beliebige Informationen detailreich zu beschreiben. Damit sind sie in der Lage, neben inhaltlichen Angaben auch strukturelle und organisatorische Informationen abbilden. Somit können Metadaten ein wertvolles Hilfsmittel für die Datenintegration sein und sollen daher im Rahmen dieser Arbeit dahingehend untersucht werden, ob und unter welchen Voraussetzungen sie sich für die Integration von Daten der klinischen Forschung eignen. Dazu werden erforderliche Standards eingeführt und eine systematische Untersuchung durchgeführt, die ihren Fokus auf die Struktur und Definition von (klinischen) Metadaten legt. Im Rahmen der Untersuchung wurde ein Problem im Umgang mit Metadaten deutlich: bevor sie zur Datenintegration genutzt werden können, müssen die Metadaten selbst integriert und verfügbar gemacht werden. Die Metadaten liegen wie die dazugehörigen klinischen Daten heterogen verteilt in abgekapselten Systemen. Anhand einer Anforderungsanalyse wurden die Voraussetzungen für einen föderalen Metadatenverbund untersucht und dessen Bedingungen abgesteckt. Basierend auf diesen Ergebnissen wurde eine neue, standardkonforme Schnittstelle QL^{4}MDR speziell für die Kommunikation von Metadaten konzipiert. Es wurde eine Integrationsstudie durchgeführt, welche aktuell genutzte Systeme der klinischen Forschung in Betracht zieht. Darauf aufbauend wurde QL^{4}MDR in den zwei meistgenutzten Systemen integriert und deren Metadaten verfügbar gemacht. Zur Demonstration der durch die Verwendung von QL^{4}MDR gewonnenen neuen Verwendungsmöglichkeiten wurden drei Dienste konzipiert und implementiert. Diese Dienste sollen die Datenintegration mittels Metadaten vereinfachen und ein Metadaten-gestütztes Ökosystem begründen.

Abstract

The advance of digitization and the introduction of new technologies in our healthcare system are creating constantly new sources of data. A holistic overview of the rapidly growing data pool can offer medical experts new individual therapy options for each patient and accelerate clinical research. However, profitable and effective use of data requires the ability to link disparate sources. This necessary process is particularly complicated by the highly heterogeneous healthcare IT landscape. These conditions require that data sources first be made usable by integrating them into workflows. This data integration and the associated data exchange are becoming more and more important with the increasing number of distributed systems and the constantly growing amount of data. As a possible integration tool, metadata are capable of describing any information in great detail. Thus, they are able to map structural and organizational information in addition to content information. Thus, metadata can be a valuable tool for data integration and will therefore be investigated in this thesis to determine whether and under which conditions they are suitable for the integration of clinical research data. For this purpose, required standards will be introduced and a systematic investigation will be conducted, focusing its attention on the structure and definition of (clinical) metadata. During the investigation, one problem in dealing with metadata became clear: before it can be used for data integration, the metadata itself must be integrated and made available. Metadata, like the associated clinical data, resides heterogeneously distributed in encapsulated systems. Based on a requirements analysis, the prerequisites for a federated metadata network were investigated and its conditions were mapped out. Based on these results, a new, standards-compliant interface QL^{4}MDR was designed specifically for metadata communication. An integration study was conducted that considered currently used clinical research systems. Based on this, QL^{4}MDR was integrated into the two most commonly used systems and their metadata made available. To demonstrate the new uses gained by using QL^{4}MDR, three services were designed and implemented. These services are intended to facilitate data integration using metadata and to establish a metadata-driven ecosystem.

Inhaltsverzeichnis

Kapitel 1

Einleitung

Der digitale Wandel verändert unser Gesundheitssystem und seine Fachdisziplinen fortschreitend. Durch die rasante Digitalisierung entstehen auch immer mehr neue klinische Datenquellen, welche erschlossen werden müssen, um miteinander nutzbar zu sein. Dies ist von enormer Bedeutung für die Patientenversorgung, da eine Datenintegration und -fusion aller Quellen einen umfassenden ganzheitlichen Überblick erlaubt. Doch viele Anwendungen im heutigen Gesundheitssystem sind stark veraltet und können mit der stetigen Digitalisierung nicht Schritt halten. Einfach zu bedienende Plug-and-Play-Technologien und -Systeme, wie sie im privaten Bereich zur Routine geworden sind, sucht man im medizinischen Bereich noch vergeblich. Die soziotechnischen Anforderungen der Patienten sind in den letzten zehn Jahren rapide gestiegen. Jedoch haben sich die verfügbaren Technologien und entsprechenden Software-Implementierungen im Gesundheitswesen nicht in der gleichen Geschwindigkeit verbessert [Marques and Ferreira, 2020]. Die Zerstückelung des digitalen Gesundheitswesens in viele proprietäre Einzelsysteme erschwert den gewünschten ganzheitlichen Überblick und bremst technische Innovationen wie den Einsatz von mobilen Geräten zur Befundung. Dies spiegelt die Realität in deutschen Krankenhäusern und Universitätsklinika wider. In nur 60 % der hochschulmedizinischen Häusern stehen Patientendaten auf Mobilgeräten für eine direkte Dokumentation am Krankenbett zur Verfügung. In weit weniger als 40% aller anderen Krankenhäusern stehen dem medizinischen Personal solche technischen Möglichkeiten zur Verfügung. Nötig für eine flächendeckende mobile Befundung ist die Öffnung der Systeme mit neuen Schnittstellen und eine damit verbundene Integration aller Datenquellen. Dadurch ergeben sich zahlreiche weitere Anwendungen wie die personalisierte Medizin und insbesondere die sekundäre Nutzung der Routinedaten in klinischen Studien. Doch werden in deutschen Krankenhäusern und Universitätsklinika die Routinedaten kaum weiter genutzt [Hübner et al., 2020]. Um Innovationen zu forcieren, hat der deutsche Gesetzgeber

eine Vielzahl an neuen Gesetzen und Initiativen vorangetrieben, die das bestehende Gesundheitssystem nachträglich verändern werden: die Einführung einer verpflichtenden elektronischen Patientenakte, ein digitales eRezept, die Einführung von digitalen Gesundheitsanwendungen auf Rezept, und nicht zuletzt auch das Patientendaten-Schutz-Gesetz zur datenschutzrechtlichen Verbesserung in der intersektoralen Kommunikation. Die neuen Gesetze werden unser Gesundheitssystem zwangsläufig digitalisieren, aber das Problem der Datenintegration und Wiederverwendung im Gleichzug nicht lösen. Ein Blick nach Österreich zeigt, dass selbst mit der Einführung einer landesweiten elektronischen Gesundheitsakte nicht mehr Routinedaten in der klinischen Forschung genutzt werden. Die Daten hierzulande liegen in Tabellenform auf Klinikarbeitsplätzen und könnten doch in einem gemeinsamen gesicherten System einen höheren Mehrwert erzielen. Dies spiegelt sich auch in der Novelle des Krankenhausgesetzes für das Land Schleswig-Holstein aus 2020 wider. War es Ärzten zuvor möglich selbst erhobene Daten aus der Routineversorgung für eigene wissenschaftliche Forschung zu nutzen, wurde dies mit der Novelle in Kombination der Datenschutz-Grundverordnung erschwert. Jegliche Forschung muss unter strenger Wahrung des Datenschutzes und der Datensicherheit erfolgen. Daraus folgt ein direkter Bedarf zur Integration von Daten aus vorangegangen Forschungsprojekten in sichere und datenschutzkonforme Systeme, um dem Datenschutz gerecht zu werden und eine weitere Forschung zu ermöglichen. Die klinische Datenintegration ist ein eminentes Thema der medizininformatischen Forschung, da die Datenfusion ein wichtiger Grundstein für die weiterführende Verarbeitung der Daten ist. Im Sinne des *Data Trusts* [O'hara, 2019] muss eine Integration der Daten immer nachvollziehbar sein, da bei medizinischen Daten eine Doppelung von Datensätzen fatale Folgen haben könnte: fälschliche doppelte Medikamentengabe oder ein Tumor, der auf zwei Kontrolluntersuchungen nicht wächst.

Datenintegration im Allgemeinen ist ein großes und vielseitig erforschtes Gebiet der Informationstechnologien [Batini et al., 1986]. Doch birgt die Verarbeitung von Daten aus dem Gesundheitswesen neue Herausforderungen, sodass klassische Ansätze wie Schemaintegration [Bellahsene et al., 2011] oder moderne *Ontology Alignment*-Verfahren [Mate et al., 2015] nur schwer anwendbar sind. Die zu integrierenden Daten stammen nicht aus einer kompletten Datenbank eines Krankenhausinformationssystem (KIS)s, sondern es müssen eine Vielzahl von heterogenen Anwendungssystemen betrachtet werden. Die einzelnen Informationssysteme der Krankenhauslandschaft sind meist proprietär und aufgrund von marktwirtschaftlichen Interessen sind die Schemata nicht verfügbar. Ohne

den Einblick in das zugrundeliegendes Schema ist eine Datenintegration im klassischen Sinne erschwert, bzw. unmöglich. Zudem unterscheidet sich die verwendete Datenhaltung von Anwendungsfall zu Anwendungsfall. Immer wiederkehrende Informationsangaben zur Anamnese wie Name und Geschlecht des Patienten kann in klassischen und starren relationalen Datenbanken gespeichert werden. Bei Studienformularen hingegen ist der Inhalt volatiler. Die Form und Aufteilung in Fragen und Kontextgruppen ist gleichbleibend, aber der Inhalt, Anzahl und Verschachtelungstiefe in jedem Formular anders. Dies wäre mit *constrainten* relationalen Datenbanken schwerlich abbildbar. Alternativ denkbar wäre eine atomisierte Abbildung im Sinne des *Resource Description Framework* (RDF). Die technischen Möglichkeiten, wie die Atomisierung und die starke Vernetzung, könnten viele Hürden überwinden [Hammad et al., 2020]. Doch finden die Methoden des *Semantic Web* wenig Anwendung im (deutschen) Gesundheitswesen, da kein verwendetes KIS diese unterstützt. So wird für volatile Studienformulare oft ein Mittelweg eingeschlagen: das Entity–Attribute–Value-Modell [Dinu and Nadkarni, 2007]. Das Modell beschreibt, ähnlich zu RDF, alle Daten in einem Informationstripel, bestehend aus dem Objekt, einem Attribut und dem dazugehörigen Wert: Der Patientenblutdruck (*Entity*) ist (*Attribute*) 150 zu 80 (*Value*). Durch die gewonnene Flexibilität lassen sich die Studieninhalte besser beschreiben. Gleichzeitig muss die Vielzahl der neuen Modelle maschinenverarbeitbar und nachvollziehbar bleiben, um die Daten für eine sekundäre Nutzung verfügbar zu machen. Dadurch entsteht ein direkter Bedarf nach einem flexiblen Werkzeug für die Beschreibung der Datensätzen: ein geeigneter Kandidat sind Metadaten und diese sind das zentrale Thema dieser Arbeit. Es soll untersucht werden, ob sie sich als Werkzeug eignen, um einerseits die Datenintegration klinischer Daten zu unterstützen und anderseits die Sekundärnutzung der Daten zu fördern.

Metadaten sind in der Lage, die vielfältigen Charakteristiken von Informationsobjekten präzise zu beschreiben. Sie bilden neben inhaltlichen und administrativen Informationen auch die Struktur und - mithilfe von Annotationen - die Semantik geeignet ab. Ihre Bedeutung für die Beschreibung von Datensätzen wird durch die Adressierung in den renommierten FAIR-Kriterien klargestellt [Wilkinson et al., 2016]. Metadaten sind ein wichtiges und ausdrucksstarkes Werkzeug für die Beschreibung von Informationen und bieten zudem die Möglichkeit die beschriebenen Daten auf Grundlage der Metadaten auf ihre Qualität zu überprüfen [Kapsner et al., 2021, Schmidt et al., 2021]. Doch die Definition und Verwendung von Metadaten folgt keinem Selbstzweck. Sie sind ein Werkzeug, welches, falsch genutzt, keinen Mehrwert für die Datenintegration erzeugt.

Entscheidend ist dabei, dass sie langfristig nachvollziehbar und verfügbar sind, auch wenn der dazugehörige Datensatz selbst nicht mehr verfügbar und dessen Inhalt folglich auch nicht mehr nachvollziehbar ist. Ist das nicht gegeben, läuft man Gefahr die Integrationsherausforderung der Daten zu verschieben und erhält inkompatible Sammlungen von Metadaten. Dann müssen die Metadaten, welche die wertvollen Informationen aus dem Gesundheitswesen integrieren und verfügbar machen sollten, selbst erst integriert und verfügbar gemacht werden, um ihren Verwendungszweck zu erfüllen.

1.1 Aufbau und wissenschaftliche Beiträge

Das übergeordnete Ziel dieser Arbeit ist die Bereitstellung von klinischen Projektdaten für eine weitere Forschung im Sinne der Sekundärnutzung. Metadaten werden dazu als sinnvolles Werkzeug betrachtet und sollen hierzu auf Tauglichkeit als Integrationsmittel untersucht werden. Die Arbeit ist thematisch in drei Kapitel unterteilt, die jeweils ein Themengebiet auf dem Weg zur Metadaten-getriebenen Integration behandeln. Die Abbildung 1.1 zeigt schematisch die Zusammenhänge zwischen den Instanzdaten, den dazugehörigen Metadaten und den darauf aufbauenden Schnittstellen und Diensten.

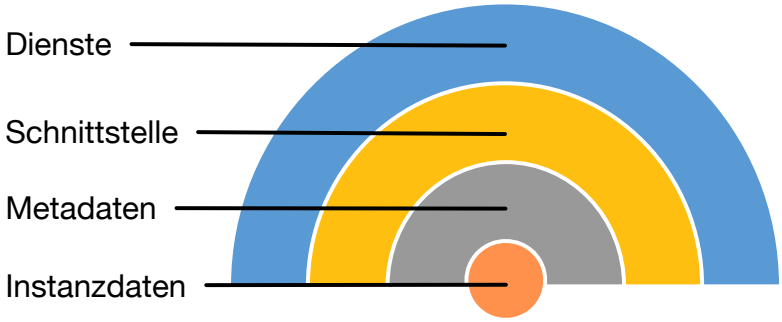


Abbildung 1.1: Die Grafik zeigt die schematischen Zusammenhänge zwischen Instanzdaten, Metadaten und aufbauend darauf die Protokolle und Dienste.

Kapitel 2 leitet in die grundsätzliche Thematik der Metadaten ein. Zu Beginn werden wichtige Standards eingeführt und die Vorgehensweise und Ergebnisse einer systematischen Literaturrecherche zum Thema *Metadaten* vorgestellt. Die Ergebnisse der Literaturrecherche konnten zudem hochrangig veröffentlicht werden [Ulrich et al., 2022]. Das Kapitel endet in einer Empfehlung für die semantische Annotierung von Metada-

ten, welche einer analytischen Betrachtung von semantischen Annotationen [Ulrich et al., 2017] zugrunde liegt.

Kapitel 3 behandelt die geringe Verfügbarkeit von Metadaten und den daraus entstehenden Anforderungen an eine Metadatenintegration. Dazu werden die formalen Anforderungen für eine mögliche föderale Lösung erörtert. Auf Grundlage der Anforderungen wurde eine Stakeholder-Umfrage erstellt und ausgewertet, zudem wurde eine Bestandsanalyse der *Metadata Repositories* angefertigt, welche in der medizininformatischen Forschung eingesetzt werden. Basierend auf diesen Vorarbeiten wird die Abfragesprache QL^{4}MDR vorgestellt [Ulrich et al., 2019b, Ulrich et al., 2020b], welche speziell für die Kommunikation zwischen Metadata Repositories konzipiert worden ist. Die Einsatz- und Integrationsfähigkeiten wurden anhand der zuvor analysierten Metadata Repositories untersucht [Ulrich et al., 2018b, Ulrich et al., 2019a] und die erfolgreiche Integration in zwei Systeme beschrieben.

Kapitel 4 beschreibt die Entwicklung von Mehrwertdiensten, welche auf Basis der zuvor eingeführten Abfragesprache QL^{4}MDR kommunizieren und die verfügbaren Metadaten gewinnbringend weiterverarbeiten können. Der übergeordnete Anwendungsfall fokussiert dabei die Metadaten-gestützte Datenintegration [Kock-Schoppenhauer et al., 2021]. Dazu wurden mehrere Dienste konzipiert und implementiert: ein Dienst für das Metadaten-Matching [Deppenwiese et al., 2019], ein Dienst für das Metadaten-Mapping [Ulrich et al., 2020a] und abschließend ein Dienst, welcher die Metadaten-gestützte Datentransformation unterstützt.

Kapitel 5 beinhaltet die Diskussion der vorangegangen Beiträge und die Zusammenfassung der beschriebenen Ansätze und Ergebnisse. In der Diskussion werden die Fragen adressiert, ob Metadaten für die Integration geeignet sind und ob klassische Metadata Repositories mit der voranschreitenden Verbreitung von HL7 FHIR obsolet werden. In der abschließenden Zusammenfassung werden die wichtigsten Schlussfolgerungen zu den Ergebnissen der Arbeit gezogen.

Kapitel 2

Metadaten

Metadaten dienen der detaillierten und eindeutigen Beschreibung von Informationen und sind das zentrale Thema dieser Arbeit. Im nachfolgenden Kapitel werden wichtige Standards eingeführt und die Struktur und Definition von Metadaten genauer betrachtet. Für die Untersuchung wurde eine systematische Übersichtsarbeit [Ulrich et al., 2022] angefertigt mit dem Ziel die Erstellung von Metadaten zu untersuchen und eine literaturgestützte Definition von Metadaten zu liefern. Hierbei wurde im Speziellen ein Fokus auf Datenintegration als Anwendungskontext der Metadaten gelegt. Zudem konnten Hindernisse und deren Lösungen aufzeigt werden, welche im Umgang und bei der Verarbeitung von Metadaten beschrieben wurden. Das Kapitel schließt mit der Betrachtung semantischer Annotation von Metadaten [Ulrich et al., 2017].

Metadaten werden für verschiedene Anwendungen in unterschiedlichen Forschungsbereichen wie der Bibliographie verwendet. Die W3C definiert Metadaten simpel als Information zur maschinenlesbaren Beschreibung von Inhalten [W3C, 2001]. Sie repräsentieren eine detaillierten Beschreibung des Objekts, welche für die Datenidentifikation und -klassifikation verwendet wird. Die US-amerikanische Standardisierungsorganisation National Information Standards Organization (NISO) unterteilt Metadaten zudem in drei verschiedene Arten: deskriptiv, strukturell und administrativ [NISO, 2017]. Die Kategorisierung in die drei Arten hilft, Metadaten besser zu verstehen.

- **Strukturelle** Metadaten beschreiben z. B. Datenmodelle und Referenzdaten.
- **Deskriptive** Metadaten beschreiben Ressourcen präzise zum Zweck der Identifizierung und Klassifikation.
- **Administrative** Metadaten liefern Informationen für die Verwaltung einer Ressource, wie Lizenzen die an die Informationen gekoppelt sind.

Beispielsweise kann ein Buch, wie in Abbildung 2.1 zu sehen, durch verschiedenen Arten von Metadaten beschrieben werden. Autor, Titel und Vorwort sind Beispiele für deskriptive Informationen, während die Aufteilung in Kapiteln und die Seitenanordnung strukturelle Metadaten sind. Informationen über das Publikationsdatum und Copyright-Informationen werden als administrative Metadaten klassifiziert. Entscheidend für die Verwendung von Metadaten ist jedoch eine klare Definition. Die Erfahrung mit der Verarbeitung und Verwaltung von Metadaten hat jedoch gezeigt, dass der Begriff *Metadaten* und ihre Verwendung trotz Klassifizierungen nicht immer eindeutig sind. Guerra et al. [Guerra and Fernandes, 2013] stellte diesen Umstand sehr treffend dar:

> »Metadaten sind ein überladener Begriff in der Informatik und können je nach Kontext unterschiedlich interpretiert werden.«

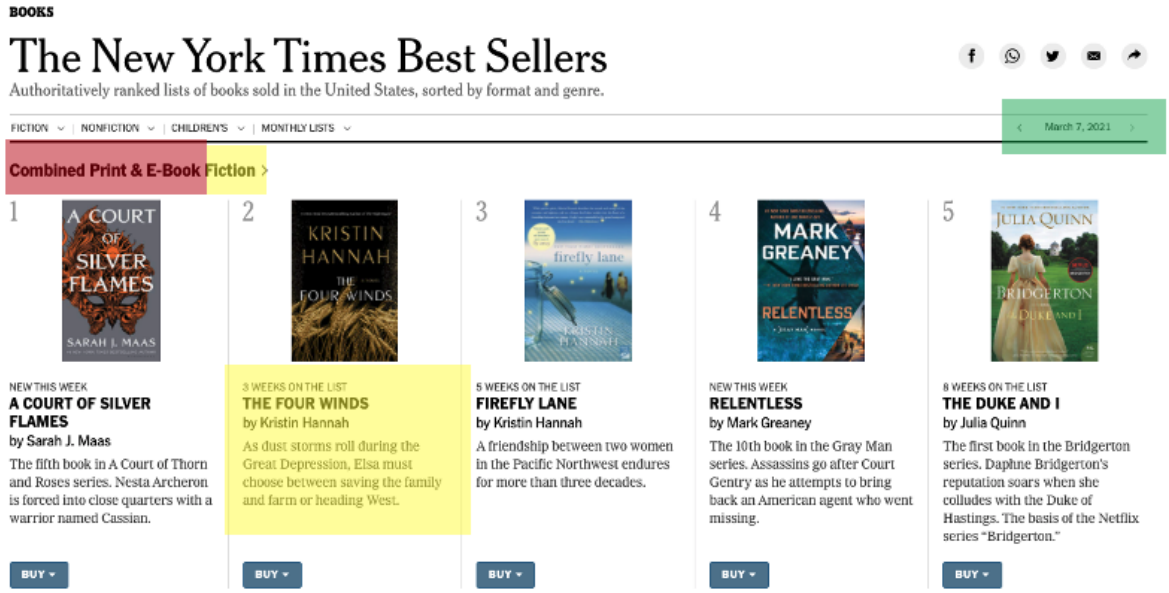


Abbildung 2.1: Für die Beschreibung eines Buches werden verschiedene Metadaten verwendet. Das Werk wird über deskriptive Metadaten (■) in verschiedenen Genre eingeteilt, zudem wird ein Titel und Autor angegeben. Die Art der Veröffentlichung (Print oder Ebook) wird über strukturelle Metadaten (■) beschrieben. Administrative Metadaten (■) zeigen an, wann die Liste erstellt worden ist. Sie dienen verwaltungstechnischen Zwecken.

Der thematische Rahmen dieser Arbeit beschäftigt sich weniger mit bibliographischen Metadaten sondern mit Metadaten im klinischen Umfeld. Hier werden die Metadaten der verschiedenen Kategorien für unterschiedliche Verwendungszwecke eingesetzt. Deskriptive Metadaten werden für die Beschreibung von klinischen Datensätze genutzt, um das Auffinden und die Wiederverwendung zu fördern oder administrative für die Nachvollziehbarkeit von klinischen Verarbeitungsschritten. Strukturelle Metadaten werden verstärkt zur Beschreibung einzelner Fragen in klinischen Studienformularen genutzt, wie in Abbildung 2.2 dargestellt ist. Die strukturellen Metadaten in Verbindung mit den deskriptiven Metadaten spielen für diese Arbeit eine entscheidende Rolle. Die Verbindung der Metadaten kann helfen den Kontext der klinischen Fragen maschinell zu verstehen und dadurch kann Sekundärverwendung ermöglicht werden. Eine genaue Aufarbeitung der Anatomie von Metadaten und eine klare Definition ist Ziel der folgenden Übersichtsarbeit (siehe Kapitel 2.2).

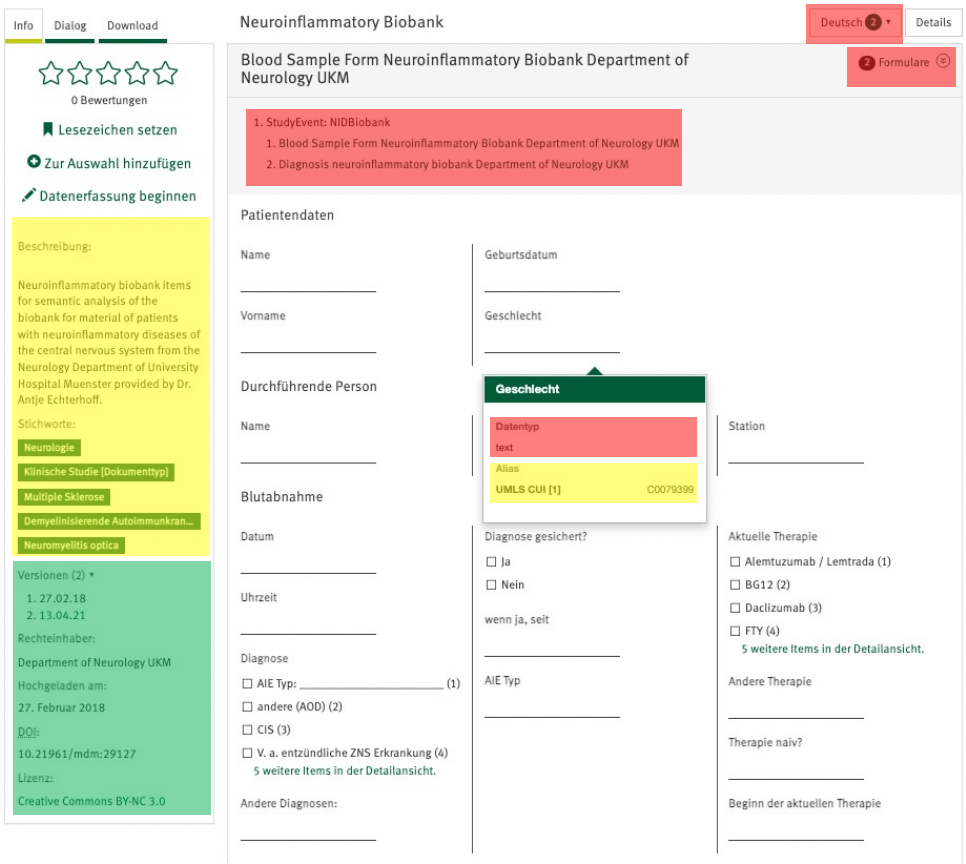


Abbildung 2.2: Dieses Beispiel zeigt eine farbliche Aufbereitung von Metadaten innerhalb eines Formulars des Universitätsklinikums Münster [Dugas, 2018]. Eine Beschreibung des Formulars wird über deskriptive Metadaten (■) angegeben, sowie Tags für eine Klassifizierung. Der Aufbau des Formulars mit verschiedene Unterformulare inklusive einer Datentypenangabe, hier exemplarisch für das Feld Geschlecht, ist durch strukturelle Metadaten (■) beschrieben. Administrative Metadaten (■) geben Versionsverlauf, Rechteinhaber und Veröffentlichungsdatum an.

2.1 Standards

Wichtig für eine sinnvolle Verwendung von Metadaten ist eine verständliche und nachvollziehbare Darstellung. Diese soll ein gemeinsames Verständnis der Daten und somit ihre Wiederverwendbarkeit sichern. Die Auswahl des verwendeten Standards ist dabei entscheidend. Geschuldet der Fülle an (Metadaten-)Standards kann das eigentliche Ziel der Verminderung der Heterogenität verfehlt werden. Im Nachfolgenden werden drei für diese Arbeit wichtige Strukturstandards vorgestellt.

2.1.1 ISO/IEC 11179

Der ISO 11179 - Information Technology – Metadata Registries (MDR) [ISO/IEC, 2013] ist ein weit verbreiteter Standard für die Darstellung von Metadaten in einem Metadatenregister. In Metadatenregistern werden die Daten zentral und organisiert zur Verfügung gestellt. In Abgrenzung zum reinen Metadata Repository, wo die Daten *nur* gespeichert werden, bietet ein Register zusätzliche administrative Informationen, wie Zuständigkeiten und Abstimmungsprozesse. Das Ziel des Standards ist die Vereinheitlichung der Metadatendarstellung, um einen interoperablen Austausch zwischen verschiedenen Registern bzw. Repositories zu sichern. Der ISO 11179 ist unterteilt in sieben zu unterschiedlichen Zeiten veröffentlichten Kapiteln:

- ISO/IEC 11179-1:2015 Framework
- ISO/IEC 11179-2:2005 Classification
- ISO/IEC 11179-3:2013 Registry metamodel and basic attributes
- ISO/IEC 11179-4:2004 Formulation of data definitions
- ISO/IEC 11179-5:2015 Naming and identification principles
- ISO/IEC 11179-6:2015 Registration
- ISO/IEC 11179-7:2019 Metamodel for data set registration

Der Standard unterteilt generell die Metadatenelemente nach ihrem Registierungsstatus im Metadatenregister und definiert welche Attribute für eine vollständige Registrierung beschrieben sein müssen. Es gibt drei Registrierungsstati: ein einfaches, unregistriertes Element, ein registriertes Element und ein administriertes Element. Je weiter

ein Element im Prozess fortgeschritten ist, umso mehr Attribute sind verpflichtend hinzufügen. Der Standard beschreibt für jedes Element allgemeine Attribute, welche in fünf verschiedene Kategorien unterteilt sind: Identifikation, Benennung, Definition, Administration und Relationen. Für eine eindeutige Identifizierung schreibt der Standard einen eindeutigen *Identifier* vor; zusätzlich kann die Version des Elements aufgeführt werden. Bei registrierten und administrierten Elementen kann zudem noch das verantwortliche Register angeben werden. Der Aufbau des Identifiers ist vergleichbar mit der Artefaktidentifikation in den bekannten HL7 Standards V3, bzw. Clinical Document Architecture [Aschhoff et al., 2013]. Für die Benennung erlaubt der Standard, dass ein Element verschiedene Namen mit maschinenlesbarer Angabe der Sprache erhält. So werden synonyme Bezeichnungen ermöglicht, vergleichbar dem Alphabetischen Verzeichnis der ICD-10-GM. Damit soll einer Redundanz von Metadatenelementen vorgebeugt werden, da verschiedene Benennung zugelassen werden und Nutzer die ihnen bekannten Bezeichnungen wiederfinden können. Sind mehrere Namen angegeben, so müssen zusätzlich noch Kontextnamen und Kontextidentifier definiert werden, welche eine Zuordnung der einzelnen Namen und dem beabsichtigten Verwendungszweck herstellt. Der Kontextname ist eine freitextliche Beschreibung des Verwendungszwecks, wohingegen der Kontextidentifier eindeutig für den Kontext sein muss. Der Kontext des Elements wird durch das definitorische Attribut *Definition* beschrieben. Der Standard erlaubt hier, dass es mehrere Definitionen pro Element geben kann, wenn sie semantisch den gleichen Inhalt ausdrücken.

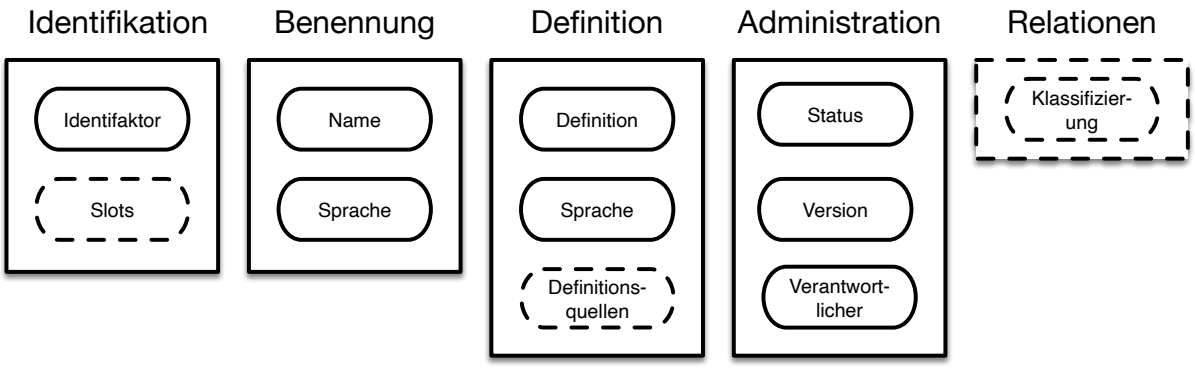


Abbildung 2.3: Die Grafik zeigt die Basisattribute eines Metadatenelements nach ISO 11179-3. Die Attribute sind ihrer Funktion zugeordnet. Mit steigenden Registrierungslevel müssen mehr Attribute verpflichtend angegeben werden. Die Attribute mit einer gestrichelten Umrandung sind optional anzugeben.

Zu jeder freitextlichen Definition kann die verwendete Sprache und eine Quellreferenz angegeben werden, wo der Kontext detaillierter beschrieben ist, bspw. ein Verweis zu kontrollierten Vokabularien oder Ontologien. Die administrativen Attribute sind nur für registrierte und administrierte Elemente verpflichtend vorgeschrieben. Es können der Registrierungsstatus, die Version, die verantwortliche und die erstellende Organisation, sowie ein freitextlicher Kommentar beschrieben werden. Über die Beziehungen lassen sich Elemente bezüglich definierter Konzepte klassifizieren. Das kann ein Klassifizierungsschema, eine Taxonomie, eine Ontologie oder ein anderes terminologisches System sein, bzw. kann auch nur eine Liste von kontrollierten Vokabeln sein. Eine Übersicht über die elementaren Attribute sind in Abbildung 2.3 dargestellt. Das dritte und für diese Arbeit wichtige Kapitel ISO 11179-3 führt die Darstellung von Metadatenelemenenten mit allen wesentlichen Attributen und Relationen ein. Im Kern der Darstellung steht ein Modell aus vier Metadatenentitäten, welche in zwei Schichten unterteilt werden, wie in 2.4 zu sehen. Es wird strikt unterteilt in zwei Ebenen: konzeptionell *was* die Information beschreibt und repräsentativ *wie* sie es beschreibt.

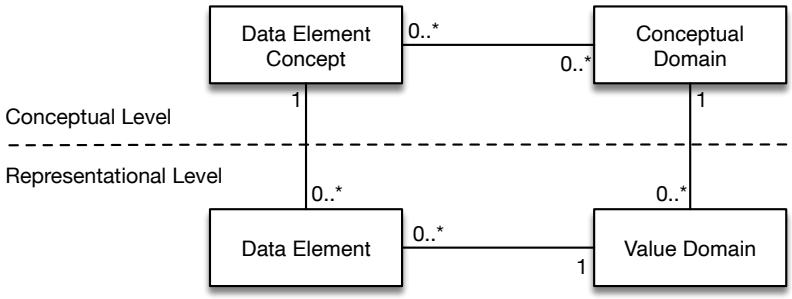


Abbildung 2.4: Schematische Darstellung des ISO/IEC 11179-3 Kernmodells mit der Unterscheidung in die Darstellungsebenen [ISO/IEC, 2013].

Die repräsentative Beschreibung der Informationen erfolgt über *Data Elements* (DE) und die *Value Domains* (VD). Ein Data Element stellt die grundlegende, atomare Struktur für die Definition und Beschreibung von Informationen dar, für die Definition, Identifizierung, Darstellung und zulässige Werte durch einen Satz von Attributen festgelegt werden. Die Value Domains beschreiben die möglichen Wertausprägungen. Sie bieten eine explizite Darstellung von Werten oder Messungen - ohne sie mit einem Kontext zu verbinden. Die Werte können einfache Messwerte oder Bezeichner mit vordefinierter Bedeutung sein, wie z.B. kontrollierte Vokabulare oder Codes aus terminologischen Systemen.

Die konzeptionelle Beschreibung wird über die zwei Entitäten *Data Element Concept* (DEC) und der übergeordneten *Conceptual Domain* (CD) dargestellt. Die DECs beschreiben Konzepte, welche die Datenelemente inhaltlich klassifizieren. Sie sind über die Kombination aus zwei Attributen eindeutig beschrieben: die Objektklasse und ein spezifisches Attribut. Die Objektklasse beschreibt dabei eine feste Menge von Ideen, Abstraktionen oder Dingen in der realen Welt, die mit expliziten Grenzen und Bedeutungen identifiziert werden. Das spezifische Attribut grenzt die Menge dann weiter ein, indem sie einen bestimmten Aspekt oder eine Eigenschaft hervorhebt. Die Conceptual Domain fasst verschiedene Konzepte unter einem gemeinsamen Bezug zusammen. Zudem verbindet sie die Konzepte mit den möglichen Werten aus den repräsentativen Value Domains. Eine Conceptual Domain kann auf zweierlei Wesen definiert werden, welche dann die VDs bestimmen: durch eine freitextliche Definition oder eingeschränkt auf einen bestimmten Wertebereich.

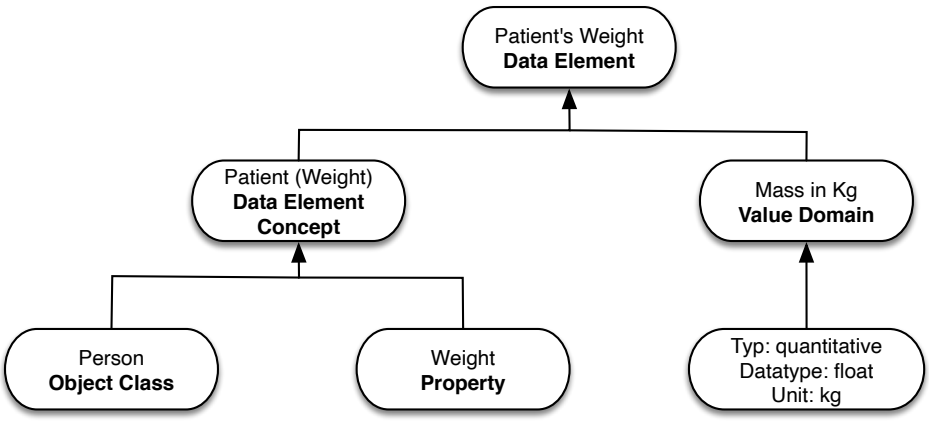


Abbildung 2.5: Die Grafik zeigt das Komposition eines Data Elements, welche das Patientengewicht erfassen soll. Es gehört einem Data Element Concept an (links), welches durch die Objektklasse *Patient* mit der Eigenschaft *Gewicht* beschrieben wird. Die Value Domain (rechts) gibt an, dass das Gewicht in *Kilogramm* gemessen wird.

Ein Data Element wird durch ein Data Element Concept und die Value Domain eindeutig definiert, wie in Abbildung 2.5 beispielhaft für die Gewichtsmessung eines Patienten zu sehen ist. Die definierten Data Elements sind meist für eine projektspezifische Identifikation in *Namespaces* gruppiert. Ein Studienformular wird beispielsweise durch viele Data Elementen gebündelt in einem Namespace dargestellt. Wichtig ist, dass Data Elements in verschiedenen Namespaces vorkommen können. Dadurch werden gleiche Datenerhebungen identifiziert. Zudem kann ein Data Element, welches einem Namespace

und dadurch einem spezifischen Verwendungszwecks zugeordnet wird, durch sogenannte *Slots* individuell erweitert werden. Slots sind Attribut-Werte-Paare und sollen durch ihre Flexibilität eine erneute Definition eines Metadatums aufgrund geringster Abweichungen verhindern.

2.1.2 ISO/TS 21526

Der Nachfolgestandard ISO/TS 21526 Health informatics - Metadata Repository Requirements (MetaRep) ist als Erweiterung und Präzisierung der ISO/IEC 11179 [ISO/IEC, 2019] konzipiert und wurde im Oktober 2019 veröffentlicht. Der neue Standard setzt auf die Wiederverwendung von etablierten und einsatzerprobten Standardkapiteln. Die Novelle zielt auf die Vereinfachung der Metadatendefinitionen trotz der strukturellen Komplexität ab, wie in Abbildung 2.6 dargestellt.

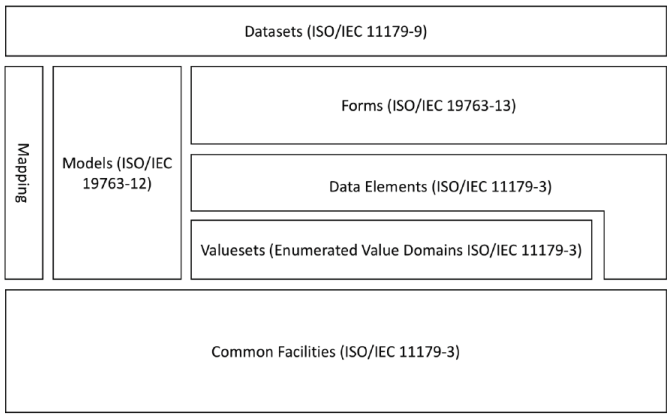


Abbildung 2.6: Die Metadaten-Standards ISO 11179 und ISO 19763 werden in der neuen ISO 21526 zusammengefasst, um Anforderungen an Metadaten-Repositories zu definieren. Zusätzlich wird ein Mapping-Paket eingeführt [ISO/IEC, 2019]

Er konzentriert sich auf die Erfassung der Zusammenhänge zwischen Datenmodellen, die zum Austausch von Informationen im Gesundheitswesen verwendet werden. Daher ist ein Mapping-Kapitel hinzugekommen, welches eine definierte Abbildung zwischen Metadaten erstmals im MDR selbst erlaubt. Die Speicherung dieser Zusammenhänge und deren Kontextinformationen sind für die spätere Interpretation und Wiederverwendung der ausgetauschten Daten entscheidend. Dafür wurden drei neue Entitäten eingeführt: MDRMapping, Map und ein enumerierter MapType. Das MDRMapping fungiert als Container und gruppiert verschiedene Maps zusammen. Die Gruppierungsart ist dabei

Tabelle 2.1: Die Tabelle zeigt die vom ISO 21526 definiertern MapTypes. Die Maptypes drücken die Beziehung zwischen Datenelementen aus.

Type (englisch)	Type	Bedeutung
broader	breiter	Das Zielobjekt hat eine allgemeinere Bedeutung als das Quellobjekt
narrower	beschränkter	Das Zielobjekt hat eine spezifischere Bedeutung als das Quellobjekt
related	verwandt	Das Ziel- und Quellobjekt besitzen einen verwandten Inhalt
same_as	gleich	Das Ziel- und Quellobjekt sind per Definition äquivalent
derived_from	abgeleitet	Das Quellobjekt wurde vom Zielobjekt abgeleitet

nicht vorgeschrieben, möglich sind alle Maps bezüglich zweier Namespaces zueinander oder alle Maps korrespondierend zu einem Datenelement *Patient_Geschlecht* gruppiert. Die Map als solches verbindet zwei Datenelemente unter der Angabe eines MapType miteinander. Eine Transformationsregel in Sinne der Matching-Definition (siehe 2.2.1) fehlt im Standard. Der MapType beschreibt die Art der Beziehung zwischen den beiden in der Map referenzierten Datenelementen. Es gibt fünf vordefinierte Beziehungen (siehe 2.1), aber der Standard erlaubt explizit weitere Arten, falls diese gebraucht werden.

Des Weiteren wurde das Konzept-Modell vereinfacht und nach dem *Simple Knowledge Organization System* (SKOS) [Miles and Bechhofer, 2009] modelliert, um die Implementierung von Konzeptrelationen zu erleichtern [Stöhr et al., 2021].

2.1.3 HL7 FHIR

Die Health Level 7 (HL7) Standards sind weit verbreitet im Gesundheitswesen und strukturieren einen Großteil der Kommunikation zwischen den verschiedenen Informationssystemen und fördern die Interoperabilität zwischen den Leistungserbringern des Gesundheitswesen [Benson and Grieve, 2016]. Ein Vertreter der HL7 Standardfamilie ist der erstmals 1989 veröffentliche HL7 v2, welcher bis dato einen Großteil der elektronischen Krankenhauskommunikation standardisiert [Smits et al., 2015]. 2011 begannen nach dem weniger erfolgreichen Nachfolger HL7 v3 die ersten Arbeiten an dem aktuellen Standardvertreter, Fast Healthcare Interoperability Resources (FHIR). Der neue Standard kombiniert moderne Webtechnologien wie *Restful Communication* [Fielding and Taylor, 2000] mit den Vorteilen der vorherigen Standards und kapselt die Inhalte

in vordefinierten Ressourcen. Die Ressourcen sind kleine abgegrenzte Informationsschemata, welche eine gemeinsame Struktur definieren, um von den meisten Implementierungen genutzt zu werden. Ein Fokus des Standards ist der Zugang und die einfache Implementierbarkeit, um eine möglichst weite Adaption zu ermöglichen. In der aktuellen FHIR Version R4 sind 148 verschiedene Ressourcen enthalten - diese reichen vom Patienten über Prozeduren, Biomaterialproben, Arzneimitteln bis hin zu ganzen Studienbeschreibungen inklusive Ein- und Auschlusskriterien. Die Schemata dienen dann als Vorlage für die einzelnen Instanzen, welche die individuellen Patienten- und Behandlungsinformationen beinhalten. Jede Ressourceninstanz (siehe Abb. 2.7) besteht dabei aus vier Teilen: Metadaten, einer menschenlesbaren Repräsentation, Dateninhalt und benutzerdefinierten Erweiterungen (*Extensions*). Diese Erweiterungen ermöglichen eine maschinenlesbare Flexibilität, welche im Gegensatz zu den vorherigen Standards nicht gegeben war: v2 ließ zu viele und V3 keine Freiräume [Smits et al., 2015]. Eine Extension dient dazu Länder- oder Usecase-spezifische Attribute abzubilden, beispielsweise die Angabe des Geburtsortes (siehe 2.7) oder zusätzliche komplexe genetische Informationen in einer einfachen Observation. Mehrere Extensions können in einem Profile zusammengefasst werden, welche die spezifischen Anforderungen bündelt. Die Medizininformatik-Initiative investiert viel in die Erstellunge eines Kerndatensatzes [Semler et al., 2018], welcher auf FHIR Profilen und Extensions basiert.

FHIR unterstützt in Abgrenzungen zu seinen Vorgängern v2 und v3 mehrere Kommunikationsparadigmen: *Restful Communication*, Nachrichten, Dokumente, Services und Datenbank. Bei der Restful Communication spezifiziert FHIR allein das Interface, nicht die darunterliegende Datenhaltung. Dadurch soll ein Umstieg auf die neuen Ressourcen vereinfacht werden. Nachrichten und Dokumente wurden von den beiden Vorgängern übernommen, da es dort erfolgreiche Anwendungen gab. Die Services sind definierte Funktionen, welche über das klassische REST-Pattern hinausgehen. Beispielsweise lässt sich über den Service *Patient $everything* jegliche Informationen, welche über einen Patienten auf einem Server gespeichert wurden, mit nur einem Aufruf abfragen. Das Database-Paradigma wurde im aktuellen Release aufgenommen und beschreibt Speicherung der Ressourcen direkt als Datenbankenschemata.
Die Entwicklung von FHIR wird maßgeblich durch die Entwickler der Anwendungssysteme getrieben. Durch den Bedarf im täglichen Entwicklungsalltag sind die einzelnen Ressourcen entstanden und zudem werden fortlaufend neue Ressourcen zum Testen in den Standard aufgenommen. FHIR folgt dabei einem sechs-stufigen Reifegradmodell, wel-

ches Ressourcen in der 6. normativen Endstufe im Standard verankert. Diese dürfen dann zukünftig nicht mehr verändert werden. In der aktuellen Version R4 haben 11 der 148 Ressourcen den normativen Stand erreicht, alle anderen Ressourcen können von Änderungen, bzw. sogar dem Löschen aus dem Standard betroffen sein. So beinhaltete FHIR in seiner zweiten Version DSTU 2. die Ressource DataElement, welches eine Abbildung des zuvor eingeführten ISO 11179 DataElements war. Die direkte Darstellung von standardkonformen Metadaten in FHIR war für die Verwaltung von klinischen Studien sehr effektiv, wie in einer früheren Arbeit gezeigt werden konnte [Ulrich et al., 2016]. Durch die DataElements konnten Fragebögen (*Questionnaire*) strukturiert und die einzelnen Fragen registert und dadurch wiederverwendet werden. Leider wurde aufgrund der mangelnden Verbreitung einer effektiven Metadatenverwaltung die Ressource DataElement in der nachfolgenden Standardveröffentlichung restlos gestrichen.

Abbildung 2.7: Dieses Beispiel zeigt den Aufbau einer FHIR Patienten-Ressource farblich aufgeteilt in die vier möglichen Sektionen. Die erste Sektion (■) enthält die Metadaten, wie die Version und die letzte Änderung. Weiterhin soll jede Instanz eine maschinenlesbare Repräsentation (■) haben. In Nachrichtenkörper (■) werden die Information darstellt. Fehlt im Schema der Ressource etwas, kann es durch *Extensions* (■) integriert werden. In diesem Beispiel wurde die Instanz um den Geburtsort erweitert, welcher nicht standardmäßig Teil der Patienten-Ressource ist.

2.2 Systematische Übersichtsarbeit zur Struktur von Metadaten

Eine klare Definition von Metadaten und der konsequente Einsatz von Standards ist entscheidend für die weitere Verwendung. Leider haben die umfangreichen Erfahrungen mit der Verarbeitung und Verwaltung von Metadaten gezeigt, dass der Begriff *Metadaten* und seine Verwendung nicht immer eindeutig sind. Daher wurde eine systematische Übersichtsarbeit [Ulrich et al., 2022] nach den PRISMA-Richtlinen [Liberati et al., 2009] durchgeführt. Letztere Richtlinien standardisieren den Erstellungsprozess der Übersichtsarbeit mit dem Ziel Vergleichbarkeit und Nachvollziehbarkeit der Ergebnisse zu gewährleisten. Für eine fundierte und unabhängige Auswertung der Literatur wurden insgesamt zehn Gutachter aus unterschiedlichen Feldern der Medizininformatik einbezogen. Der Gutachterkreis setzte sich aus verschiedenen relevanten Berufsgruppen zusammen: Ärzte, Medizininformatiker und Data Stewards - alle mit unterschiedlicher technischer Expertise. Die gesamte Übersichtsarbeit ist in zwei Prozessteile unterteilt: ein erster Harmonisierungschritt und die darauf folgende Literaturanalyse.

2.2.1 Harmonisierungsprozess

Um ein einheitliches Verständnis der Begriffe unter den Gutachtern zu gewährleisten, sowie Fehlinterpretationen und Fehlklassifikationen zu minimieren, ging der eigentlichen Übersichtsarbeit ein Harmonisierungsprozess voraus. Dazu wurde ein Fragebogen konzipiert, welcher insgesamt fünf Fragen und Aufgaben beinhaltete. Abgefragt wurden der wissenschaftliche Hintergrund der Gutachter auf dem Gebiet der Metadaten, Klassifikationen der Metadatenverarbeitung und deren Automatisierungsmöglichkeiten sowie Definitionsverständnis der zentralen Begriffe *metadata matching*, *mapping* and *transformation*. Die Ergebnisse zeigten eine starke Übereinstimmung bei allen in 2.8 aufgeführten Aufgaben, mit Ausnahme der fünften Aufgabe *Validierung der Konvertierungsregeln*. Für die Konformität in der Analyse wurde die Klassifizierung *Transformation* vereinbart. Bei der Klassifizierung der Automatisierbarkeit wurden die ersten beiden Aufgaben unterschiedlich angesehen, siehe Abbildung 2.9. Nach einer Ergebnisdiskussion wurde entschieden, dass Aufgabe 1 *Vollautomatisch* und Aufgabe 2 *Manuell* durchgeführt werden sollten.

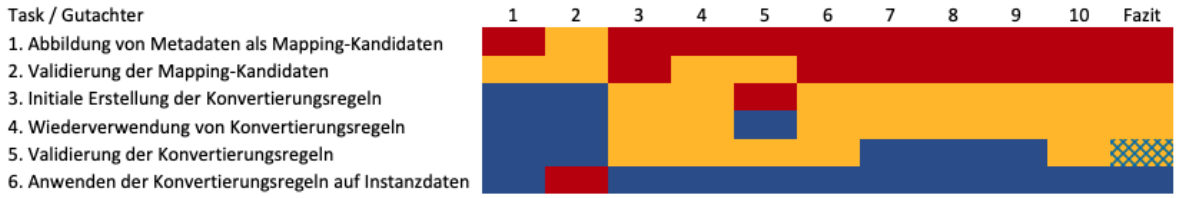


Abbildung 2.8: Die Gutachter klassifizierten sechs verschiedene Teilaufgaben aus dem Bereich der metadatengetriebenen Datenintegration in drei Kategorien: Matching ■, Mapping ■ und Transformation ■.

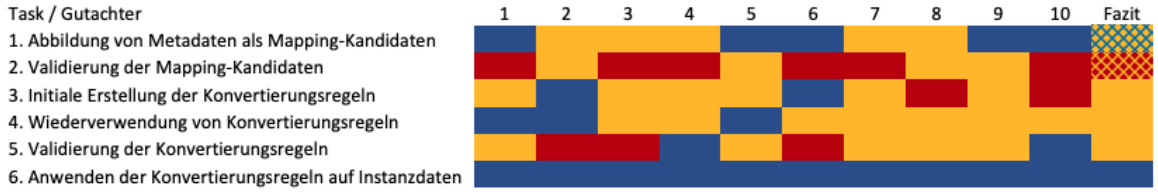


Abbildung 2.9: Die Gutachter klassifizierten sechs verschiedene Teilaufgaben aus dem Bereich der metadatengetriebenen Datenintegration hinsichtlich der Automatisierungsmöglichkeit in drei Kategorien: Manuell ■, Halbautomatisch ■ und vollautotmatisch ■.

Basierend auf den Ergebnissen wurden im Konsens aller Gutachter gemeinsame Definitionen für die drei Begriffe erstellt:

Matching Der Matching-Prozess beschreibt den Abgleich gegebener Datenstrukturen oder Metadaten und erstellt einen Matching-Vorschlag zwischen den einzelnen Datenelementen. Diese Matching-Kandidaten können von Fachexperten oder Matching-Algorithmen unter Verwendung von Äquivalenzklassen (z.B. äquivalent, schmaler, breiter) erstellt werden.

Mapping Beim Mapping-Prozess werden die Vorschläge des Matching-Prozesses zur Definition von Funktionen verwendet, dabei können ebenfalls externe Regelsätze (z.B. UCUM) genutzt werden, um die Quelldatenstruktur in eine Zieldatenstruktur zu transformieren. Die Umwandlungsfunktionen sind nicht zwingend symmetrisch.

Transformation Der Transformationsprozess kombiniert Metadaten und Instanzdaten. Er verwendet die im Mapping-Prozess definierten Konvertierungsregeln und Umwandlungsfunktionen, um die Instanzdaten entsprechend der Zieldatenstruktur zu transformieren.

2.2.2 Literaturanalyse

Nach dem Abschluss des Harmonisierungsprozess folgte die Analyse der Literatur. Der Prozess begann mit der Definition klarer und eindeutiger Forschungsfragen, die in der Literaturrecherche beantwortet werden sollten. Die tägliche Arbeit mit Metadaten für die klinische Datenintegration hat gezeigt, dass es bei den Anwendern und Experten kein klares Verständnis von Metadaten und ihren Anwendungsmöglichkeiten gibt. Als Beispiel: Matching kann auf verschiedene Weise verstanden werden. Metadaten *matchen* entweder auf Instanzdaten [Canakoglu et al., 2019] oder auf semantischen Attributen [Gonçalves and Musen, 2019] oder anderen Metadaten [Martínez-Romero et al., 2019]. Das allgemeine Verständnis ist mehrdeutig. Daher zielt die Studie darauf ab, eine akzeptable Definition von Metadaten (Q1) und, mit dem Fokus auf die Datenintegration, Definitionen für die Verarbeitung von Metadaten (Q2) zu finden, um das allgemeine Verständnis und die Kommunikation zu verbessern. Darüber hinaus soll ein Überblick über die Vielfalt der verwendeten Metadatenstandards (Q3) und die Erzeugung von Metadaten in anderen Forschungsbereichen (Q4) gegeben werden, um die damit verbundenen Probleme und deren Lösung zu verstehen (Q5). Die daraus resultierenden Fokusfragen lauten daher:

Q1. Wie wird der Begriff *Metadaten* in verschiedenen Forschungsbereichen definiert?

Q2. Wie werden die Begriffe *Matching, Mapping* und *Transformation* definiert?

Q3. Welche Metadatenstandards sind in Gebrauch?

Q4. Wie werden Metadaten in anderen Fachbereichen genutzt?

Q5. Was sind die beschriebenen Probleme und welche Lösungen werden genannt?

Die Recherche basierte auf einer umfangreichen Literaturauswahl, da die ausgewählten Publikationen äußerst entscheidend für Ergebnisse sind. Für die Recherche wurden *SCOPUS*[1] und *Web of Science*[2] verwendet. Im ersten Schritt wurde das sehr allgemeine Stichwort *metadata* verwendet, um eine möglichst breite Auswahl an wissenschaftlichen Arbeiten zu erhalten. Die Suchanfrage beschränkte sich auf Zeitschriftenartikel, Konferenzberichte und Buchkapitel aus den letzten zehn Jahren - 2010 bis 2019. Der erste

[1] https://www.scopus.com/
[2] https://webofknowledge.com/

Suchlauf ergab 23.233 Veröffentlichungen - 21.161 nach Entfernung der Duplikate. Etwa 10 % aus allen Publikationen wurden nach dem Zufallsprinzip ausgewählt und dann nach Titel und Kurzfassung analysiert. Entsprachen die Arbeiten dem Fokus der Forschungsfragen wurden die Schlüsselwörter der Arbeit extrahiert. Die Literatursuche wurde mit dem erweiterten Suchbegriff wiederholt (siehe 2.10), was 681 Papiere ergab, 551 Papiere nach der Entfernung der doppelten Einträge. Die Titel und Kurzfassungen der zweiten Literaturrecherche wurden dann wiederum analysiert, ob sie dem Aufgabengebiet entsprachen, sodass 81 Arbeiten für die Volltextanaylse des Gutachterkreises ausgewählt wurden. Die vorangegangen Screening-Schritte wurden vom Autor dieser Dissertation durchgeführt, der als einziger Reviewer alle 81 Arbeiten analysiert hat. Zudem wurden die Arbeiten randomisiert den Gutachtern zugewiesen, sodass jede Arbeit zusätzlich von zwei Gutachtern analysiert wurde. Der Begutachtungszeitraum erstreckte sich über acht Wochen. Die Beiträge kamen aus verschiedenen Disziplinen: Medizinische Informatik (41), Bibliographie (10), Bioinformatik (8), Informatik (8), Sozialwissenschaften (8), Geographie (4), Neuroinformatik (1), Energieinformatik (1) und Chemie (1).

Um den Review-Prozess zu standardisieren, wurde ein Erhebungsbogen mit acht Fragen erstellt: sechs Fragen, die dem Forschungsschwerpunkt entsprechen, und zwei Fragen, um zusätzliche Informationen über die ausgewählte Literatur zu erhalten. Die Forschungsfragen konzentrierten sich auf evtl. verfügbare Metadatendefinitionen (Q1), das Matching, das Mapping und/oder die Transformation von Metadaten (Q2), verwendete Standards (Q3), beschriebene Anwendungsfälle (Q4), aufgetretene Probleme und entsprechende Lösungen (Q5). Die zusätzlichen Fragen bezogen sich auf das Forschungsgebiet, aus dem das Papier stammt und welche Art von Metadaten beschrieben wird. Für die Klassifizierung der Metadaten wurde das zuvor eingeführte NISO Schema genutzt, siehe Abbildung 2.2. Die Ergebnisse der Literaturanalyse werden im Folgenden beschrieben - gruppiert nach den zentralen Forschungsfragen; darauf folgt die Diskussion der Ergebnisse.

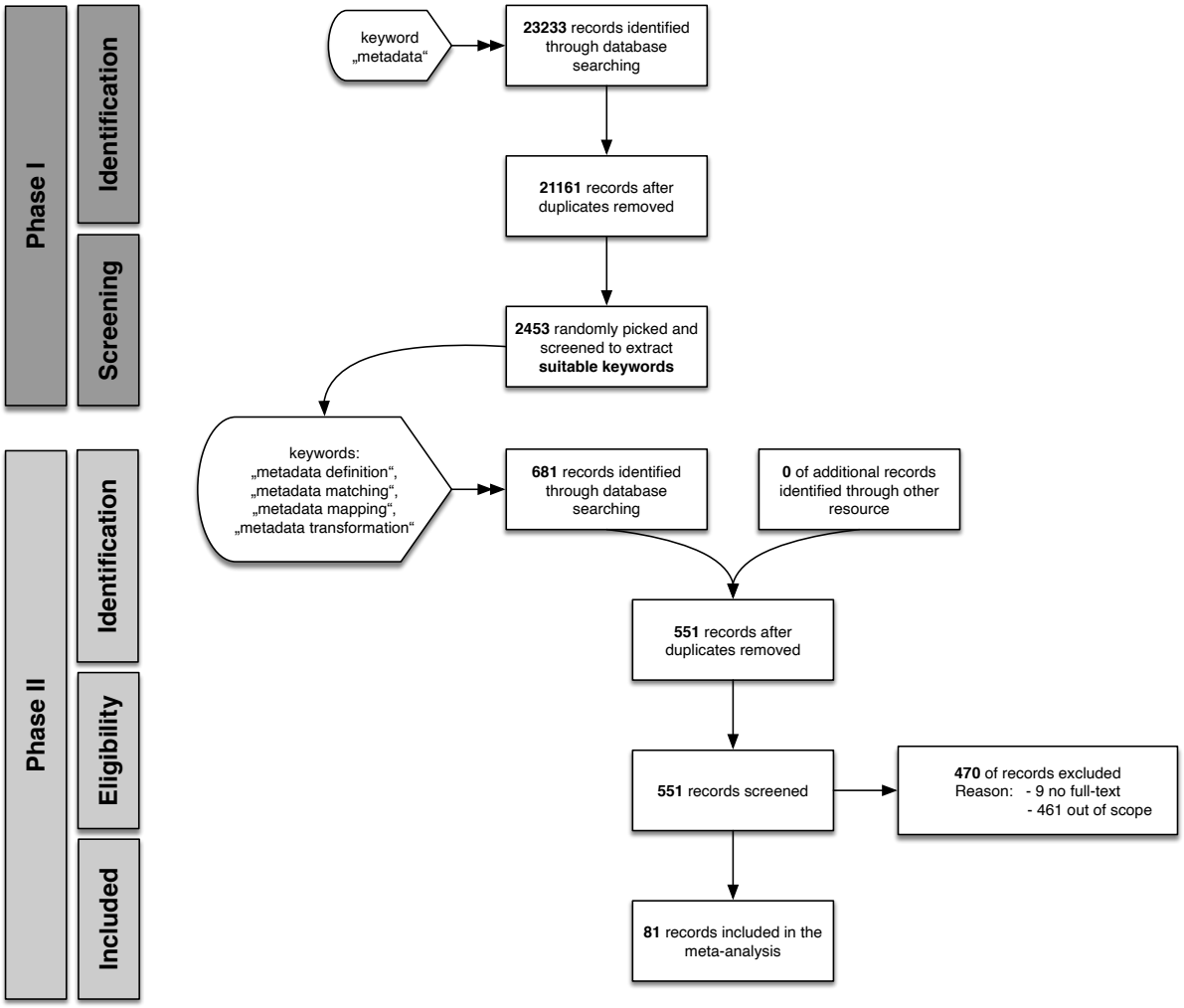


Abbildung 2.10: Die Übersicht stellt den Auswahlprozess der literarischen Übersichtsarbeit wieder. Der Prozess ist zweigeteilt, verdeutlicht durch die farbliche Abtrennung.

Q1. Die Definition und Klassifikation von Metadaten

Im Allgemeinen werden Metadaten als eine formale Repräsentation von Daten beschrieben, die Informationen in einer (vorzugsweise) standardisierten und zuverlässigen Weise definieren [Lopprich et al., 2014, Li et al., 2012]. In der Literatur wurden verschiedene Merkmale einer Metadatendefinition identifiziert:

- Kleine atomare Einheiten, die ein bestimmtes Informationsobjekt beschreiben und einschränken (Tabellenfelder, Attribut von Formularfragen, Datensätze)[Kim et al., 2019]
- Beschreibt Datentyp, Bereich oder Satz von möglichen Werten [Ashish et al., 2016, Breil et al., 2012]
- Einzelne Einheiten können zu komplexen Elementen zusammengesetzt werden [Corradi et al., 2012]
- Einzelne Einheiten werden oft in Anlehnung an die ISO 11179 als Datenelement bezeichnet [ISO/IEC, 2013].
- Metadaten können Bindungen zu Terminologien, kontrollierten Vokabularen und Taxonomien haben [Daniel et al., 2014, Gonzalez-Beltran et al., 2018]
- Trennung von Inhaltsinformationen von Strukturinformationen [Breil et al., 2012]
- Detaillierte, maschinenlesbare und verfolgbare Informationen, die eine Datenverarbeitung ohne menschliche Interaktion ermöglichen [Lyttleton et al., 2011, Martínez-Romero et al., 2019]
- Metadaten werden zentral in *Metadata Repositories* gespeichert [Ashish et al., 2016, Cunningham et al., 2016, Bruland et al., 2017, Kock-Schoppenhauer et al., 2018a]

Die Mehrzahl der Publikationen wurde von den Gutachtern als strukturell oder deskriptiv klassifiziert - Artikel mit reinem Fokus auf administrative Metadaten waren in der betrachteten Literaturauswahl kaum zu finden. Die Kategorisierung von Metadaten nach der NISO findet sich direkt in verschiedenen analysierten Arbeiten [Charles et al., 2013, Nadkarni and Marenco, 2013, Papež and Mouček, 2017, Song et al., 2014], aber

es wurden auch andere Definitionskriterien beschrieben. Chu et al. [Chu et al., 2018] führten die Trennung von Metadaten mit und ohne Abhängigkeiten vom Kontext ein. Ein wichtiges Unterscheidungsmerkmal ist hier: Einige Metadaten erfassen Informationen, die nicht von den Daten abhängig sind. Kontextunabhängige Metadaten könnten eher technische, herkunftsspezifische Datensätze beschreiben, während kontextabhängige Metadaten die Datensätze beschreibend definieren könnten, um die Identifizierung zu verbessern. Die Studie von Grewe et al. [Grewe et al., 2011] beschrieb ein neues Konzept zur Annotation neurophysiologischer Berichte, um möglichst viele Annotationen zu erfassen. Dabei unterschieden die Autoren zwischen harten und weichen Metadaten. Parameter und Informationen, die direkt gemessen und bewertet werden können, werden als harte Metadaten bezeichnet, während Kontextinformationen und Experimentaufbauten als weiche Metadaten bezeichnet werden. Li et al. [Li et al., 2013] entwarfen ein Datenverwaltungssystem für ein maritimes Observationsnetzwerk und unterscheiden dabei vier verschiedene Metadatentypen: Datenqualitätsinformationen zur Gewährleistung der Datenzuverlässigkeit, Referenzsysteminformationen zur Erfassung zeitlicher und regionaler Referenzdaten, Wartungsinformationen zur Anzeige von Aktualisierungen und Lebenszyklen sowie Identifikationsdaten. Ein anderer Kategorisierungsansatz wurde von Zozus et al. [Zozus and Bonner, 2017] gewählt, der das beschriebene Element differenzierte: Metadaten auf Datensatz- oder Wertebene. Sie stellten fest, dass insbesondere bei klinischen Studien eine feingranulare Definition auf Wertebene wünschenswert ist. Im Gegensatz zur Bibliographie, bei der der gesamte Datensatz für das Retrieval wichtig ist, sind bei klinischen Studien die Fragestellungen bezeichend und sollten so detailliert wie möglich definiert und eingeschränkt werden.

Q2. Definition von Matching, Mapping und Transformation

In der betrachteten Literatur wurden neben Metadatendefinitionen auch Beschreibungen von Matching und Mapping gefunden. Ashish et al. [Ashish et al., 2016] definierte einen Satz von Matching-Kandidaten als Vorschläge und ein Mapping als eine Eins-zu-Eins-Beziehung zwischen zwei Datenelementen. Rebai et al. [Rebaï et al., 2015] beschrieb, dass ein Mapping eine semantische Korrespondenzbeziehung zwischen zwei Metadatenschemata ist, die in einem Schema-Matching-Prozess identifiziert wurde. Mate et al. [Mate et al., 2019a] unterschied bei Definitionen ebenfalls zwischen Matching und Mapping: Mapping-Kandidaten sind das Ergebnis eines Matching-Prozesses, doch

sobald ein menschlicher Experte die Beziehung bestätigt, wird ein Mapping-Kandidat zu einem Mapping. In der Studie von Bernstein et al. [Bernstein et al., 2017] wurde eine neue Differenzierung eingeführt: explizite und abgeleitete Mappings. Eine explizite oder eher direkte Zuordnung wird zwischen zwei Metadatenelementen erstellt, während eine implizite Zuordnung die explizite Zuordnung verwendet, um neue Beziehungen zu erstellen, bspl. ein *metadata crosswalk* [Woodley, 2008].

Definitionen von metadatengetriebener Transformation von Instanzdaten wurden in den betrachteten Beiträgen nicht gefunden.

Q3. Metadatenstandards

Ein weiteres Ziel der Literaturstudie war eine Übersicht über die verwendeten Standards und Kerndatensätze zu erhalten. Im Ergebnis wurden 37 relevante Standards gefunden, die in den analysierten Arbeiten erwähnt oder verwendet wurden. Für die bessere Übersicht wurden sie anschließend in drei Kategorien gruppiert:

- Struktur: ISO 11179, ISO 15926, ISO 19101, ISO 19763, ISO 20943, ISO 21526, ISO 23081, openEHR, CDISC ODM, OMOP, IHE DEX, Dublin Core, ASTM CCR, CaDSR, EAD, GILS, VRA Core, CIMI, FHIR, CSDGM, ONIX, MARC, TMA DES, EXIF, INSPIRE, SKOS, DCAT, W3C PROV
- Technisch: XML, RDF, OWL, JSON-LD, ClaML
- Semantisch: ICD-10, UMLS, SNOMED CT, LOINC, MedDRA, RxNorm.

Q4. Usecases

Metadaten wurden für verschiedene Anwendungsfälle verwendet. Die in dieser Übersicht enthaltenen Beiträge zeigten, dass Metadaten hauptsächlich für vier Aufgaben verwendet werden: Information Retrieval (21 Beiträge), Datenintegration (19), Definition von Kerndatensätzen (10) und die sekundäre Nutzung von Daten (7). Beim Retrieval werden Metadaten und insbesondere semantische Annotationen verwendet, um die maschinelle Bearbeitung von Anfragen zu verbessern. Durch ein breiteres Spektrum an Beschreibungen können Anfragen feingranular erstellt werden und bessere Ergebnisse liefern. Die Prozesse der Datenintegration und Kerndatensatzdefinition verwenden Metadaten zur Beschreibung und Harmonisierung beliebiger zugrundeliegender Schemata, die

zur sekundären Verwendung von (klinischen) Daten verwendet werden können. Weitere angetroffene Anwendungsfälle waren eine automatische Datenqualitätsprüfung [Lyttleton et al., 2011, Park and Tosaka, 2010] oder die automatisierte Ontologiegenerierung [Huang et al., 2017].

Q5. Probleme und Lösungen

Die analysierten Beiträge befassten sich mit diversen Problemen bei der Verarbeitung und Nutzung von Metadaten in verschiedenen Forschungsbereichen, stellten aber auch Lösungen und neue Ansätze zur Bewältigung der Schwierigkeiten vor. Die Probleme und Lösungen wurden in fünf verschiedene Kategorien eingeteilt:

1. Strukturell bedingte Probleme
2. Semantisch bedingte Probleme
3. Probleme im Zusammenhang mit der menschlichen Interaktion
4. Probleme im Zusammenhang mit dem Lebenszyklus von Metadaten
5. Probleme bei der Verarbeitung von Metadaten.

Strukturell bedingte Probleme Der Analyse zufolge war die größte Anzahl von Problemen strukturell bedingt. Die Autoren der Artikel beschrieben eine mangelnde Anwendung von Standards, sie kritisieren eine entweder begrenzte oder zu umfangreiche Auswahl geeigneter Standards [Ivanschitz et al., 2018, Baek and Sugimoto, 2012]. Es wurde geschlussfolgert, dass sich die schwierige Wahl der Standards auf die Komplexität von Metadaten [Gonzalez-Beltran et al., 2018] und die Datenqualität [Bernstein et al., 2017] auswirkte, was zu einer mangelhaften Nutzung von Metadaten führt. Das Fehlen von Standards und damit ihre Nichtverwendung führte zu mehreren Problemen: Metadaten sind in Struktur und Format heterogen und enthalten schlechte oder fehlende Beschreibungen, die das Verständnis der vorhandenen Metadaten verhindern und wieder zu geringer Qualität führen [De Jong et al., 2019, Eichenlaub et al., 2014]. Die Verwendung unterschiedlicher Einheiten oder Genauigkeiten bei quantitativen Messungen erschwerte die Anwendung [Nadkarni, 2011] und die heterogenen Formate verhinderten die maschinelle Lesbarkeit und verschlechtern somit die Identifizierung [Ku et al., 2014, Urban,

2014], die Auffindbarkeit [Trani et al., 2018], das Retrieval [Maumet et al., 2016] und die Validierung [Charles et al., 2013]. Doch wurde auch betont, dass selbst die konstante Nutzung die Heterogenität nicht vermeidet: die derzeitigen Standards haben weder Erweiterungsfunktionen, um zukunftssicher zu sein [Chu et al., 2018] , noch bieten sie Modularität, um Metadaten aus verschiedenen Standards zusammen zu nutzen [Trani et al., 2018]. Der am weitesten verbreitete Standard ISO 11179 [ISO/IEC, 2013] ist ein Vertreter für diese problematischen Standards und den Schwachstellen: fehlende hierarchische oder zeitliche Abhängigkeiten [Lopprich et al., 2014], fehlende strukturelle [Milward, 2019, Park and Kim, 2010] oder semantische Erweiterungen [Ngouongo and Stausberg, 2011, Park et al., 2013]. In der Übersichtsarbeit wurden aber auch mehrere Lösungen in Bezug auf strukturelle Fragen gefunden: Reduzierung der ISO 11179-Entitäten, um sie zu rationalisieren und einfacher zu verwenden [Milward, 2019], Rekonstruktion der Basismodelle [Ngouongo et al., 2013] oder Erstellung eines Obermodels, das alle proprietären Erweiterungen und Anpassungen der ISO 11179 integriert [Park et al., 2013]. Eine breite Auswahl an Standards war auch nicht förderlich und begünstigt die Heterogenität der Metadaten [Cunningham et al., 2016]. Ein gutes Beispiel ist die Bibliographie, die zu viele konkurrierende Standards aufzeigt [Baek and Sugimoto, 2012]. Als möglicher Ausweg aus dem Standarddschungel wurden mehrere Möglichkeiten adressiert: ihre Anzahl zu reduzieren, indem nur Standards verwendet werden, die von der Forschungsgemeinschaft akzeptiert werden [Breil et al., 2012] oder vorhandene und validierte Datenelemente und Definitionen wieder zu verwenden [Lyttleton et al., 2011, Varghese et al., 2018a]. Wenn kein Standard geeignet ist oder die derzeitige Methode zur Definition von Standards nicht mehr angemessen ist, könnte zudem ein neuer konzeptioneller Ansatz hilfreich sein. Anstatt neue Standards zu schaffen, ermunterten Woodly et al. [Woodley, 2008] dazu, mehr Aufwand in Modellabstimmung und Modellharmonisierung zu investieren. Corradi et al. [Corradi et al., 2012] beschrieb die Verwendung eines ereignisgesteuerten Modells, um die fehlende Erweiterbarkeit anzugehen. Grewe et al. [Grewe et al., 2011] schlug einen generischen Metamodell-Ansatz vor, der auf fünf Merkmalen basiert: Erweiterbarkeit, Modularität, Verfeinerungen, Mehrsprachigkeit, maschinelle Verarbeitbarkeit.

Semantisch bedingte Probleme Die Semantik und deren Nutzung ist eine wichtige Voraussetzung für die Wiederverwendung von (Meta-)Daten, und laut den analysierten Arbeiten ist die fehlende Semantik ein schwer zu überwindendes Hindernis. Ein allgemeines Problem, das mit jeder standardisierten Datenerfassung zusammenhängt,

ist die Verwendung von Freitext [Berry and Edgar, 2019]. Metadatenelemente enthalten Beschreibungen und Definitionen, um den Zweck der Elemente zu verstehen. Mit der freitextlichen Beschreibung sind auch Synonyme und Rechtschreibvarianten möglich und oft vorhanden, welche Unstimmigkeiten und unterschiedliche Datenverständnisse verursachen [Eichenlaub et al., 2014]. Eine praktikable Lösung ist das Hinzufügen semantischer Codes zu den entsprechenden Datenelementen, die ein tieferes semantisches Verständnis darstellen, welches wichtig ist, wenn die Daten in einem anderen Kontext wiederverwendet werden sollen. Bestehende Thesauri decken jedoch (häufig) nicht alle benötigten Begriffe ab [Eichenlaub et al., 2014], oder die Verwendung proprietärer Codes führt zu semantischer Heterogenität [Breil et al., 2012]. Zur Verbesserung wurde ein optimierter Annotationsprozess vorgeschlagen, der entweder von Domänenexperten oder NLP-Tools [Bruland et al., 2017] durchgeführt werden sollte. Ergänzt durch den Prozess der Postkoordination und eine Überprüfung durch Experten, sollte eine konsistente Kodierung sichergestellt werden [Pathak et al., 2011]. Eine weitere wichtige Ergänzung wäre zudem der Zugriff auf und die Wiederverwendung von abgestimmten semantischen Annotationen [Daniel et al., 2014, Dugas et al., 2019] oder die Translation von proprietären Codes zu standardisierten Vokabularien [Berry and Edgar, 2019]. In den analysierten Arbeiten wurde noch eine weitere mögliche Lösung beschrieben: die Verwendung von Ontologien [Kim et al., 2019, Hall and McMahon, 2016]. Das Problem hier ist, dass eine solche Ontologie den Daten entsprechen muss, um anwendbar zu sein. Das heißt sie müssten spezifisch erstellt [Lunesu et al., 2011] oder anhand des Datenkörpers automatisch konstruiert werden [Ivanschitz et al., 2018]. Eine Wiederverwendung von bestehenden Ontologien, bzw. die Anpassung an den Datenkörper wurde hierbei als die bessere Wahl beschrieben, was im Falle der Anpassung aber wiederum Heterogeniätät bzw. Mapping-Aufwand mit sich bringt [Kim et al., 2019]. Bei der direkten Wiederverwendung von Ontologien entstehen jedoch Probleme aufgrund der notwendigen Konformität der Metadaten mit der Ontologiestruktur [Frosini et al., 2018].

Probleme im Zusammenhang mit der menschlichen Interaktion Menschliche Interaktion und Kollaboration ist ein weiterer Schwerpunkt, der in den betrachteten Publikationen beschrieben wurde. Gerade die Zusammenarbeit wurde nicht nur als Herausforderung, sondern ein notwendiger Schritt zur Überwindung von Problemen [Trani et al., 2018] verstanden. Doch ist die menschliche Mitwirkung zeit- und ressourcenaufwendig, da Software ungewohnt oder kompliziert ist und dem Benutzer wenig Un-

terstützung bietet [Breil et al., 2012], zudem überwiegend nur Domänenexperten eingesetzt werden können. Für medizinisches Fachpersonal ohne notwendige IT-Kenntnisse sind Metadatenmodelle oder die entsprechende Software [Pathak et al., 2011] zu kompliziert und werden daher selten eingesetzt [Li et al., 2013, Späth and Grimson, 2011]. Das Problem manifestiert sich zudem auf der konzeptionellen Ebene: Wenn die Beteiligten andere Anwendungsfälle im Sinn haben als der Ersteller der Metadaten, könnte das Modell nicht eindeutig verstanden werden [Eichenlaub et al., 2014]. Einfache Meinungsverschiedenheiten über Modellierungsentscheidungen führen zu unzureichenden Modellen [Varghese et al., 2018a]. Daher wird ein enger Feedback-Kreis zwischen Nutzern und Metadatenkuratoren empfohlen, um die erwarteten Ergebnisse und ein gemeinsames Verständnis der Metadatenelemente zu erreichen [Eichenlaub et al., 2014, Howarth, 2003]. Beispielsweise hilft die Erweiterung des Metadaten-Vokabulars durch natürliche Definitionen zur Unterstützung der Benutzer [Howarth, 2003]. Hall et al. [Hall and McMahon, 2016] beschrieben, dass Menschen ein schlechtes Langzeitgedächtnis für digitale Informationen haben und daher das Wissen *aussterben* wird. Daher sollten Vokabulare mit Einfachheit und Zweckmäßigkeit erstellt werden, anstatt eine erschöpfende Beschreibung zu verwenden sowie die Notwendigkeit für komplexes Tooling [Rodrigues et al., 2019]. In den analysierten Arbeiten wurden zwei Lösungen vorgeschlagen: Arbeitsteilung oder besseres Tooling. Verbesserte Werkzeuge würden medizinische Experten eine einfache Datenmodellierung und eine direkte Qualitätsvalidierung ermöglichen [Breil et al., 2012]. Der andere Ansatz sieht eine Aufgabenverteilung vor: ein Domänenexperte liefert das Wissen, welches von Data Stewards in enger Absprache und Rückkopplung in Metadaten zusammengestellt werden, was zu guten und wiederverwendbaren Metadaten führen würde [Kock-Schoppenhauer et al., 2019a].

Probleme im Zusammenhang mit dem Lebenszyklus von Metadaten Ein weiteres Kernproblem ist die Divergenz von Daten und den entsprechenden Metadaten [Li et al., 2012]. Die Metadaten *entsprechen* nicht den Daten und damit sind Informationen schwer wiederverwendbar. Die Gründe dafür sind vielfältig: die fehlende Transparenz der (Meta-)Datenherkunft [Maumet et al., 2016] oder die Grenze zwischen Daten und Metadaten ist unklar bzw. eher eine Frage der Perspektive [Papež and Mouček, 2017]. Ein gangbarer Lösungsansatz ist die Extraktion von Metadaten aus den primären IT-Systemen und ihre direkte Verwendung in klinischen Studien [Bruland et al., 2017].
Die Verteilung von Metadaten über mehrere Standorte ist einerseits wünschenswert, da

Metadaten in anderen Projekten wiederverwendet werden könnnen und der Kurationsaufwand verteilt werden kann [Haslhofer and Klas, 2010]. Doch anderseits besteht bei einer Verteilung die Gefahr von veralteten oder über die Zeit abweichenden Duplikaten, die aufgrund der hohen Kosten für die Pflege der Änderungensverfolung nicht synchronisiert sind [Ngouongo and Stausberg, 2011]. In den analysierten Arbeiten wurden verschiedene Maßnahmen gegen die *Alterung* genannt: kontinuierliche Anpassung und Kuratierung von Metadaten teilweise mit großen Ressourcenaufwand [De Jong et al., 2019], Verfolgung von Änderungen im Prozess der Metadatenerstellung [Park and Tosaka, 2010], Pflege und Verknüpfung von Provenance-Informationen [Francis et al., 2013] und Erstellung eines Lebenszyklusmodells für Metadaten [Baek and Sugimoto, 2012]. Vos et al. [Vos et al., 2012] wiesen auf einen entscheidenden Umstand hin: es gibt keinen aktuellen Standard für die Archivierung und Aufbewahrung, der den gesamten Lebenszyklus von Metadaten abdeckt. Gerade die Archivierung von Metadaten ist aber auch der Schlüssel für die Wiederverwendung archivierter Daten. Ohne die entsprechenden deskriptiven Metadaten sind die eigentlichen Forschungsdaten nur schwer auffindbar und interpretierbar.

Probleme bei der Verarbeitung von Metadaten Metadaten werden häufig zur Harmonisierung von Datensätzen verwendet, um den Abstimmungs- und Arbeitsaufwand zu reduzieren. Dennoch wird der Prozess der Metadatenharmonisierung häufig manuell durchgeführt [Ashish et al., 2016], was wiederum zeit- und ressourcenintensiv ist [Bruland et al., 2017]. Teilweise sind die Informationen maschinell verarbeitbar, so dass eine automatische Verarbeitung, insbesondere Matching und Mapping, möglich ist. Die Übersichtsarbeit fand jedoch Probleme schon bei der Metadatenakquise: heterogene Metadaten-Schnittstellen verursachen eine *Siloisierung* [Jeong et al., 2014], was die Erfassung und Wiederverwendung von Metadaten erschwert. Doch selbst wenn auf die Informationen zugegriffen werden kann, ist auch die Verarbeitung problematisch. Ein automatisches Matching von einer groben hin zu einer detaillierteren Beschreibung ist nahezu unmöglich [Daniel et al., 2014]. Selbst wenn die Matching-Ergebnisse vielversprechend sind, ist ein automatisches Mapping ohne menschliche Interaktion schwierig oder nicht verlässlich durchführbar [Ashish et al., 2016, Kock-Schoppenhauer et al., 2018a]. Um die Algorithmen zu verbessern, sind weiterhin mehr Testdaten erforderlich, die schwer zu erhalten sind [Deppenwiese et al., 2019].

Desweiteren ist die Fusion verschiedener Datensätze problematisch: Mappings können mehrdeutig sein [Song et al., 2014], die entsprechenden Elemente weisen unterschiedliche Kardinalität auf [Specka et al., 2019] oder vorgeschlagene Mappings sind nicht fehlerfrei, was dann zu Fehlinterpretationen von Informationen führt [Ngouongo et al., 2013]. Als Lösungsansatz wurde eine Fokussierung auf ein verbessertes Schema-Matching beschrieben, um ein breiteres Verständnis der Metadaten zu ermöglichen [Charles et al., 2013]. Die Verwendung lexikalischer und statistischer Methoden wurde als ausreichend für den Matching-Prozess beschrieben. Das anschließende manuelle Mapping [Pathak et al., 2011, Deppenwiese et al., 2019] sei unerlässlich, um angemessene Ergebnisse zu erzielen. Das Matching kann aber durch den Einsatz unbewachter Text-Mining-Techniken zur Berechnung von Ähnlichkeiten zwischen Datenelementen verfeinert werden [Ashish et al., 2016]. Um die Siloisierung von Metadaten zu überwinden, sollte die Verwendung standardisierter Softwareschnittstellen gefördert, bzw. existierende weiterentwickelt werden [Yuliant and Karna, 2017].

Diskussion

Ziel dieser Literaturanalyse war es, das Verständnis von Metadaten zu untersuchen und mögliche Probleme aufzuzeigen, welche im letzten Jahrzehnt analysiert und veröffentlicht wurden. Die folgende Diskussion behandelt zuerst das eigentliche Analyseverfahren und dann die inhaltlichen Ergebnisse.

Die erste Literatursuche war absichtlich offen angelegt, wobei das generische Suchwort *metadata* zu 21.161 Arbeiten führte. Nach mehreren Filterschritten blieben 81 Arbeiten übrig, die aber hauptsächlich aus dem Bereich der Medizinischen Informatik stammten. Dies könnte auf die Tatsache zurückzuführen sein, dass Metadaten gerade im letzten Jahrzehnt für diesen Bereich sehr relevant waren und daher dort wesentliche Arbeiten geleistet wurde. In der Volltextanalyse zeigte sich, dass der Begriff *metadata representation* als Synonym für die später verwendete Suchphrase *metadata definition* genutzt wird, was einen Einfluss auf die Literaturauswahl haben könnte. Die Volltextanalyse war so gestaltet, dass jede Arbeit von drei unterschiedlichen Gutachtern bewertet wurde. Leider wurden zehn Beiträge nur zweimal begutachtet, da ein Gutachter ausgefallen war. Um Konsistenz der eigentlich Volltextananalyse zu gewährleisten, ging ein Harmonisierungsprozess voraus. Es musste sichergestellt werden, dass alle beteiligten Gutachter das gleiche Verständnis der Definitionen hatten. Dieser Schritt erforderte zusätzlichen

Zeit- und Arbeitsaufwand, führte aber zu einem gemeinsamen Satz von Definitionen, die während der Analyse ausgewertet werden konnten. Die Definitionen und die Unterscheidung zwischen Matching und Mapping waren mit der Literatur kongruent. Dahingegen war das Verständnis von Transformation gegensätzlich. Die aufgestellte Definition konzentrierte sich auf die metadaten-getriebene Datenintegration: die Verwendung von Metadaten für die Transformation (klinischer) Instanzdaten. Die gefundenen Verwendungen standen aber im Zusammenhang mit der Transformation von Metadaten selbst. Bei der Aufstellung der Definitionen lag der Fokus auf der Verwendung von Metadaten, nicht deren Verarbeitung. Der Kontext und die Perspektive waren bei der Definitionserstellung entscheidend. Da es keine konsistente Unterscheidung zwischen der Transformation von Metadaten und der Transformation von Instanzdaten gab, kann die hier aufgestellte Definition als Abgrenzung und detaillierte Beschreibung verwendet werden.
Nach dem erfolgreichen Abschluss der Volltextanalyse, folgte die inhaltliche Analyse der Ergebnisse. Diese wurde vom Autor dieser Dissertation durchgeführt und danach allen Gutachter zur Validierung vorgelegt, um eine Missinterpretation zu verhindern. Dabei zeigte sich, dass die Verteilung der Metadatenkategorien unausgewogen war: Es gab kaum Arbeiten mit einer administrativen Ausrichtung in der Literaturauswahl. Die Herausforderungen einer nachvollziehbaren Datenerfassung und Nachvollziehbarkeit verschärfen sich mit der stark zunehmenden Digitalisierung und administrative Metadaten werden häufig zur Unterstützung von Managementprozessen verwendet. In der Literatur ist sie aber offenbar nicht stark vertreten. Dies ist erstaunlich, da gerade diese Informationen für die Dokumentation der Herkunft und Nachvollziehbarkeit von Datensätzen unverzichtbar sind. Es scheint, dass der Bereich der administrativen Metadaten, einschließlich der Provenance-Informationen, im letzten Jahrzehnt unterschätzt wurde.
Die vorgefundenen Anwendungsfälle entsprachen den Erwartungen: Metadaten werden hauptsächlich zur Verbesserung der Informationsgewinnung und Datenintegration eingesetzt. Kaum überraschend war die schiere Menge an Standards - siehe die vergleichende Arbeit von Baek et al. [Baek and Sugimoto, 2012]. Die Vielzahl unterschiedlicher Standards führt zu Übersättigung und Ablehnung, was eine wichtige Erkenntnis für die Medizininformatik ist. Folglich ist das Bewusstsein für eine begrenzte Anzahl von Standards, die von der Gemeinschaft unterstützt und verbessert und daher eingehalten werden, ein wichtiges Ziel.
Im Hinblick auf die Verarbeitung spielt die Siloisierung der Metadaten eine ausschlaggebende Rolle [Dugas et al., 2015]. Es ist für die Verarbeitung und daraus folgend für

die Wiederverwendung der Daten entscheidend, dass sie verfügbar gemacht werden. Im Bereich der Bibliographie hat sich das *Open Archives Protocol for Metadata Harvesting* (OAI-PMH) [Van de Sompel et al., 2004] durchgesetzt, in der Medizinischen Informatik gibt dafür keinerlei Ansätze. Es wäre aber förderlich, die kuratierten Metadaten abrufbar und in anderen Anwendungen nutzbar zu machen. Die wichtigste Erkenntnis der Literaturanalyse war die Abhängigkeit von Kontext und Perspektive bei der Definition und Auswertung von Metadaten. Konsistente Metadaten erfordern bei ihrer Erstellung ein hohes Maß an zuverlässiger Abstraktion, damit sie von den Benutzern allgemein verstanden werden. Dies würde eine inkonsistente und falsche (Wieder-)Verwendung der Metadaten und der entsprechenden Daten verhindern. Ein verwandtes Problem stammt aus dem Bereich des Terminologie-Engineering unter Verwendung unterschiedlicher Kodierungssysteme für die Postkoordination [Qamar et al., 2007]. Der Kontext beeinflusst die Erstellung von Informationen und verwischt die eigentlich präzise Grenze zwischen Struktur und Semantik. Die Informationen, die überhaupt erst universell anwendbar sein sollten, werden durch eine individuelle Sichtweise beeinflusst.

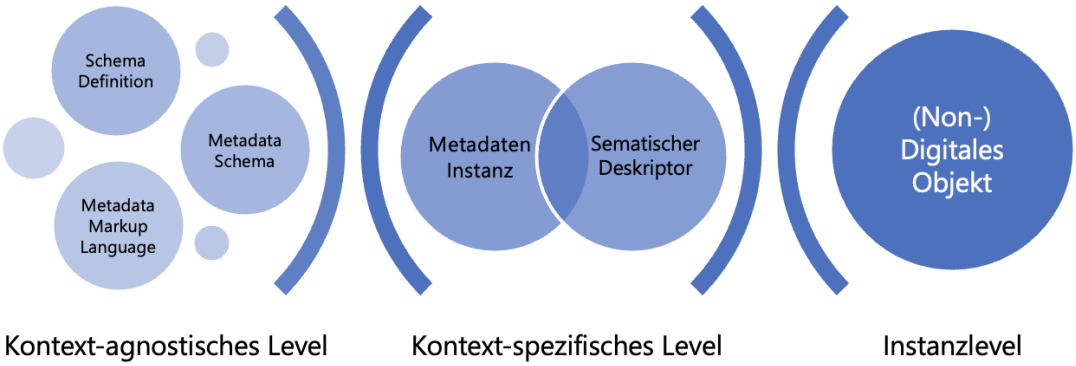


Abbildung 2.11: Die Abbildung stellt den Aufbau eines Metadatums dar. Es ist aufgeteilt nach der Abhängigkeit vom Kontext. Links, im Kontext-agnostischen Teil befinden sich die Schemadefinitionen, die Metadatenschemata und die technische *Markup Language*. Im kontextspezifischen, bzw. kontextabhängigen Teil befindet sich die Metadatensinstanz. Zusätzlich kann die Instanz mit semantischen Beschreibungen (Annotationen) angereichert werden. Ganz rechts befindet sich das eigentliche Objekt, welches mit Metadaten beschrieben wird.

Unter Berücksichtigung der entscheidenden Rolle des Kontextes für die Metadaten wurden ein neues Schema für die Klassifizierung von Komponenten für Metadaten abgeleitet. Hierbei wurde das von Haslhofer et al. [Haslhofer and Klas, 2010] vorgestellte Modell adaptiert. Wie in Abbildung 2.11 dargestellt, basiert das neue Schema auf der Identifizierung und Trennung des Kontexts im Hinblick auf die Erstellung von Metadaten. *Schema Definition*, *Metadata Schema* und eine technische *Metadata Markup Language* sind dabei kontext-agnostisch. Diese drei Bausteine bilden den technischen und semantischen Kontext, in dem ein Metadatum instanziiert wird. Die Metadateninstanz beschreibt dann die eigentliche Information, das (nicht-)digitale Objekt. Im Hinblick auf den Kontext kann die Metadateninstanz mit Hilfe eines annotierten semantischen Deskriptors unter Verwendung einer Vielzahl von Ontologien, Terminologien und Kodierungssystemen erweitert werden. Das instantiierte Metadatum kann selbst an die Stelle eines digitalen Objekts treten und durch weitere Metadaten näher beschrieben werden. Dieser Verkettungsmechanismus ermöglicht eine präzise Beschreibung der hochgradig vernetzten Natur von Metadaten. Darüber hinaus ermöglicht die Verkettung, dass Metadaten aus verschiedenen Systemen und Standards gemeinsam in einem einzigen verketteten Schema dargestellt werden können. Ein Beispiel ist die Anreicherung von Instanzdaten mit Provenance-Informationen, die die Herkunft der Metadaten beschreiben. Die Schemadefinition kann angeben, wie Metadatenmodelle aufgebaut sind. Bekannte Vertreter sind die Normen ISO 11179, ISO 15926, ISO 19763 und ISO 21526. Das Metadatenschema beschreibt die Metadatenobjekte mit allen erforderlichen Attributen und ist meist das Ergebnis einer Metadatenharmonisierung und der Erstellung von Kerndatensätzen, z.B. Dublin Core [ISO, 2017] oder Common Data Element (CDE) aus dem CaDSR [Warzel et al., 2003]. Für die technische Beschreibung des definierten Schemas werden Metadata Markup Languages wie XML, JSON , RDF oder OWL verwendet. Das Metadatenschema und die Markup Languages sind für die Metadateninstantiierung unerlässlich. Die übergeordnete Schemadefinition nach bspw. ISO 21526 ist nicht obligatorisch, wird aber für die Vergleichbarkeit und Interoperabilität dringend empfohlen.

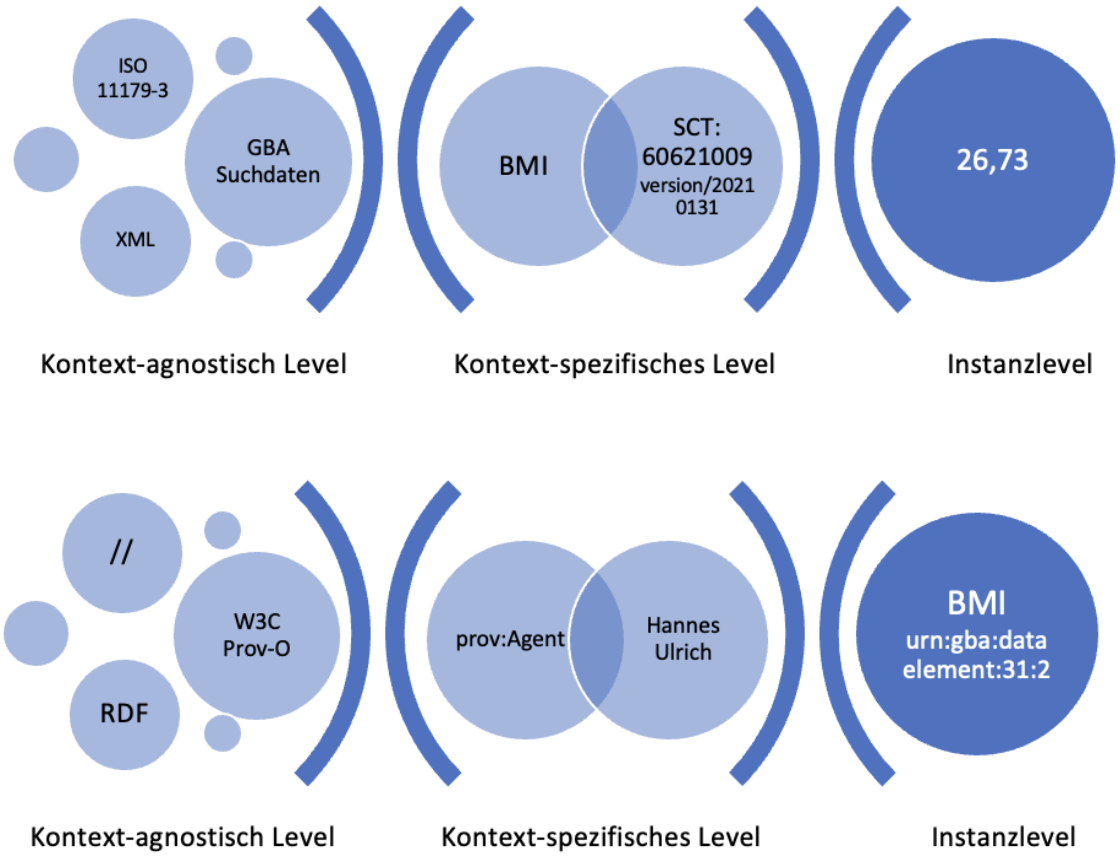


Abbildung 2.12: Die Abbildung zeigt den Aufbau eines Metadatums aufgeteilt in die identifizierten Komponenten. Das Metadatum ist dem German Biobank Alliance Suchdatensatz entnommen und mit einem SNOMED CT Code unter Angabe der Version annotiert. Die zweite Abbildung zeigt ein administratives Metadatum, welches die Annotierung des vorherigen beschreibt. Hierbei wird die annotierende Person durch die W3C Prov-Ontology beschrieben.

2.3 Semantische Annotierung

Metadaten helfen Instanzdaten besser aufzufinden und deren Inhalt zu abstrahieren. Durch den stetigen Anstieg an Instanzdaten in verschiedensten Formaten wächst auch die Anzahl an Metadaten. Um diese besser zu verarbeiten und sie selbst zu kodieren, werden Metadaten oft annotiert. Diese Annotationen oder *Tags* sollen den beschriebenen Inhalt verschlagworten und damit einfach gruppieren und die Auffindbarkeit steigern. Neben einfachen Tags, die meist für eine flache Klassifikation ausreichen, beispielsweise

ob ein Werk als Buch oder eBook verfügbar ist, bieten semantische Annotierungen einen weitergehenden Verwendungszweck. Die Metadaten werden mit Codes aus kontrollierten Vokabularien oder Terminologien versehen, um dann die hierarchische Bedeutungsstruktur der Kodiersysteme zu nutzen. In der Übersichtsarbeit konnten sechs semantische Standards gefunden werden (siehe Absatz 2.2.2) und ihre Bedeutung für die Metadaten spiegelte sich auch in aufgestellten Modell für die Metadatenklassifizierung wieder. Im Nachfolgenden sollen die sechs gefundenen Systeme aus der Übersichtsarbeit näher beschrieben werden.

Das *Unified Medical Language System* (UMLS) wird seit 1980 von der US-amerikanischen *National Library of Medicine* (NLM) herausgegeben und hat das Ziel die Vielzahl von (bio)medizinischen Vokabularien und Ontologien zu verbinden. Damit unterscheidet sich UMLS von den folgenden fünf Systeme, da es sich eher um ein übergeordnetes mediierendes Ordnungssystem handelt. Dafür wurde ein Metathesaurus erstellt, welcher mehr als 100 unterschiedliche medizinische Nomenklaturen (*Source Vocabularies*) integriert. Um die einzelnen Vokabularien in Relation zu setzen, werden harmonisierte Strukturen bereitgestellt. Der Metathesaurus identifiziert dabei einzelne Konzepte, die dann mit den entsprechenden Konzeptnamen in den verschiedenen Quellvokabularien verknüpft sind und bietet Multilingualität durch die sprachlichen Übersetzungen in derzeit 15 Sprachen. Alle Konzepte des Metathesaurus sind in einem semantischen Netz eingebunden, welches eine semantische Kategorisierung (*Semantic Types*) der Konzepte in verschiedenen Domänen ermöglicht. Es gibt 133 semantische Typen und 54 semantische Beziehungen. Die NLM aktualisiert das UMLS zweimal pro Jahr, jeweils im Mai und November.

Die Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (ICD-10) ist weltweit die weitverbreitetste Klassifizierung von Krankheiten und wird von der Weltgesundheitsorganisation (WHO) herausgegeben. Sie dient in erster Linie der statistischen Erfassung von Todesursachen und wird gerade im deutschen Raum für die Abrechnung von medizinischen Leistungen genutzt. Die Klassifikation der Krankheiten und verwandter Gesundheitsprobleme - Deutsche Modifikation (ICD-10-GM) wird vom Bundesinstitut für Arzneimittel und Medizinprodukte (BfArM) im Auftrag des Bundesministeriums für Gesundheit übersetzt und herausgegeben und beinhaltet derzeit ca. 13.300 Klassen und 76.400 alphabetische Einträge. Die ICD-10-GM wird jährlich erweitert und basiert auf der ICD-10-Variante von 2019. 2018 wurde zudem der Wechsel auf die ICD-11 von der WHO beschlossen, sodass am 1. Januar 2022 die ICD-11 in Kraft tritt.

Aber es wurde auch eine flexible Übergangszeit von fünf Jahren eingeräumt, sodass dann alle Todesursachen ausschließlich mit der ICD-11 kodiert werden müssen.

Die *Systematized Nomenclature of Medicine - Clinical Terms* (SNOMED CT) gilt als die umfassendste mehrsprachige klinische Terminologie des Gesundheitswesens weltweit [Benson and Grieve, 2016]. Stand 2021 umfasste die Terminologie 354.448 Konzepte und wird in einem halbjährlichen Zyklus von *SNOMED International* erneuert. SNOMED CT ist kompositionell und polyhierachisch in 19 Hauptachsen (*Top Level Hierachies*) organisiert und bietet über die ontologische Basis formal-logische Definitionen der einzelnen Konzepte. Dadurch ist eine maschinelle Ableitung von Schlussfolgerungen möglich. Zudem bietet SNOMED CT zwei Anwendungssprachen, bzw. Grammatiken: die *Expression Constraint Language* (ECL) und die *Compositional Grammar*. Mit ECL lassen sich maschinenauswertbare Ausdrücke erzeugen, die eine bestimmte Konzeptmenge beinhalten. Dadurch können beispielsweise Wertelisten auf eine bestimmte Konzeptuntermenge beschränkt werden; hier alle Subkonzepte eines gebrochenen Beines:

$$<< 71620000|\textit{Fracture of femur (disorder)}|.$$

Mit Hilfe der Compositional Grammar können die vorhandenen Konzepte kombiniert werden, man spricht dabei von postkoordinierten Ausdrücken (PCE). Dadurch können Konzepte ausgedrückt werden, die noch nicht vordefiniert sind. So kann der obige Ausdruck des gebrochenen Beines weiter spezifiziert werden, sodass der Bruch des linken Beines beschrieben wird:

$$\begin{aligned} &71620000|\textit{Fracture of femur (disorder)}| : \\ &363698007|\textit{Finding site}| \\ &= (71341001|Femur| : 272741003|Laterality| = 7771000|Left|). \end{aligned}$$

Die *Logical Observation Identifiers Names and Codes* (LOINC) repräsentieren eine der bekanntesten Terminologien im elektronischen Gesundheitswesen und werden verbreitet für die standardisierte Kodierung von Labordaten in elektronischen Gesundheitsakten verwendet. LOINC wird vom Regenstrief Institute gepflegt und kostenfrei für die medizi-

nische Forschung bereitgestellt. Sie spezifizieren jeden Labortest durch einen eindeutigen LOINC-Code und einen *Fully Specified Name* (FSN). Hinter jedem Code steht eine eindeutige Kombination aus verschiedenen Definitionsachsen, den LOINC *Parts*. Diese Achsen oder *Parts* sind Komponente, Eigenschaft, Zeit, (Proben-)Material und Skala. Zudem gibt es die optionale Achse Messmethode. Diese Parts werden auch für die Erstellung des FSN genutzt. Folgend wird die Namenszusammensetzung und ein Beispiel aus den ca. 80.000 Codes genannt:

<Analyt>:	<Eigenschaft>:	<Zeit>:	<Material>:	<Skala>:	<Methode>
Physical findings:	Find:	Pt:	Abdomen:	Nar:	Observed

Zudem wird die LOINC Document Ontology in HL7 V3 Clinical Document Architecture für die Kodierung klinischer Dokumente genutzt. Dabei werden die Dokumenttypen und auch Dokumentsabschnitte mit LOINC Codes spezifiziert. Damit sind die Codes neben der Beschreibung von spezifischen Labortests auch für die genaue Beschreibung von klinischen Dokumenten geeignet, wie im folgenden Beispiel ein Einwilligungsdokument mit dem 59284-0 beschrieben wird:

<Analyt>:	<Eigenschaft>:	<Zeit>:	<Material>:	<Skala>:	<Methode>
Consent:	Find:	Pt:	{Setting}:	Doc:	Patient

Das *Medical Dictionary for Regulatory Activities* (MedDRA) ist eine medizinische Terminologie, welche vom International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use seit 1999 herausgegeben wird und den internationalen Austausch von regulatorischen Informationen für Menschen zugelassene Medizinprodukte erleichtern soll. MedDRA wird bei Arzneimittelstudien im Rahmen einer möglichen Zulassung maßgeblich verwendet und soll die Registrierung, Dokumentation und Sicherheitsüberwachung von Medizinprodukten harmonisieren, dabei vor als auch nach der Zulassung eines Produkts verwendet werden. Zudem können *adverse event* mit MedDRA kodiert werden, was gerade im Bereich der klinischen Studie eine wichtige Rolle im Zulassungsverfahren spielt.

RxNorm ist ein standardisiertes kontrolliertes Vokabular für die Beschreibung von Arzneimitteln, welches speziell für den US-amerikanischen Markt von der NLM gepflegt wird. RxNorm bietet dabei standardisierte und eindeutige Bezeichner für Medikamente an mit dem Ziel, arzneimittelbezogene Informationen maschinenlesbar effizient zu kommunizieren. Hier werden monatliche Updates herausgegeben.

Semantische Annotierung am Beispiel

Die semantische Kodierung von Metadaten ermöglicht ein besseres Verständnis der beschriebenen Daten, ist aber nur dann sinnvoll, wenn die Annotation umfangreich durchgeführt wird und sie nachhaltig ist. Es muss sichergestellt werden, dass Codes bzw. das Kodiersystem auch zum beschriebenen Inhalt des Metadatums passen und die Annotationen konsistent durchgeführt werden. Inkonsistente Annotationen verschlechtern die Datenqualität sogar [Varghese et al., 2018b]. Eine händische Annotierung ist zeitintensiv und es braucht gerade im medizinischen Bereich geschulte Domänenexperten, die die Annotierungen vornehmen und unabhängig validieren. Die Auswahl der richtigen Kodiersysteme ist dabei entscheidend. Mit Blick auf die gefundenen Kodiersysteme, stellt sich die Frage, ob eher ein allumfassendes Kodiersysteme wie UMLS oder eine Vielzahl von spezialisierten Systemen für bestimmte fachliche Aspekte genutzt werden sollte. Diese Frage soll im Folgenden untersucht werden und zum Ende eine Kodierempfehlung erarbeitet werden.

Das Medical-Data-Models-Portal (MDM) der Universität Münster stellt seit 2013 klinische Formulare für die medizininformatische Forschung bereit und ist die größte Plattform dieser Art. Es wird verwendet, um Heterogenitätsprobleme durch Suchen, Vermitteln, Wiederverwenden und Bewerten von i.Allg. Formularberichten aus klinischer und Studiendokumentation zu lösen, z. B. die semi-interaktive Erstellung von Kerndatensätzen in speziellen medizinischen Fachbereichen [Holz et al., 2019]. Außerdem kann es als Benchmark für die Bewertung von Algorithmen verwendet werden, die strukturierte Patientendaten erstellen, transformieren, annotieren und analysieren. Die Daten werden im *Operational Data Model* (ODM) des *Clinical Data Interchange Standards Consortium* (CDISC) [Hume et al., 2016] abgebildet, einem Datenformat für die Darstellung klinischer Studiendaten, der den Daten- und Metadatenaustausch zwischen heterogenen Systemen unterstützen soll. Das Format ist in der Lage die stark verschachtelten Studienformulare abzubilden. Dabei nutzt es *items* und *itemsgroups* zur Definition und Gruppierung von Datenelementen. Über *codelists* und *codelistitems* werden Antwortoptionen definiert. Ein Export aus dem MDM in eine Vielzahl von anderen Formaten wird zudem angeboten [Riepenhausen et al., 2019]. Zudem werden die Formulare bzw. die einzelnen Datenelemente im Portal per Hand mit UMLS Konzepten von medizinsichem

Fachpersonal annotiert. Das Portal und der dahinterliegende Datensatz stellt eine einzigartige Quelle für die Untersuchung von annotierten Daten dar und der Datensatz ist (glücklicherweise) frei verfügbar[3]. Eine Analyse soll einen Überblick über die Annotationen des Datensatzes geben und am Ende sollen Empfehlungen für die Annotierung von Datenelementen mit Konzepten aus medizinischen Terminologiesystemen entstehen. In der Analyse sollen drei Fragen untersucht werden:

Q1. Wie viele der fünf gefundenen Vokabularien lassen sich hinter UMLS Konzepten wiederfinden und reichen sie für eine vollständige Annotation aus?

Q2. Gibt es einen Unterschied in der Annotierung von englischen und deutschen Elementen?

   a) Wie ist die Verteilung der UMLS Semantic Types?

   b) Welche der SNOMED CT Top Level Hierarchies werden genutzt?

   c) Wie wurde LOINC verwendet?

Q3. Wie qualitativ nachvollziehbar sind die Annotationen im Datensatz für eine Weiterverwendung?

Der Datensatz enthält insgesamt 427.106 Elemente in 53 Sprachen. Für jede Sprache sind vier CSV-Dateien im Datensatz vorhanden, aufgeteilt nach Elementen (*items*), Elementengruppen (*itemgroups*), Antwortlisten (*codelists*) und die dazugehörigen Optionen (*codelistitems*). Für die Analyse werden insbesondere die englischen und deutschen Elemente betrachtet. Die Analyse erfolgte in mehreren Schritten (siehe 2.13).

Abbildung 2.13: Die Grafik zeigt eine schematische Übersicht über die einzelnen Schritte der Auswertung des MDM-Datensatz.

[3]http://static.uni-muenster.de/mdm/models.zip

Tabelle 2.2: In der Tabelle werden vier verschiedene Problemtypen darstellt, welche bei der Untersuchung des kodierten Datensatzes gefunden wurden.

1.	Unbekannte Codes:	CL428482
2.	Platzhalter:	Code finden
3.	Codeeinträge:	männlich oder M
4.	Komplexe Codes:	C0085532 C1532338 C0010055 oder C0005771†C0310367

Bevor der Datensatz für die Analyse verwendet werden kann, musste er vorverarbeitet werden. Bei einer ersten Analyse des rohen Datensatzes wurden verschiedene Probleme identifiziert, die eine direkte Analyse unmöglich machten. So war die CSV-Datei durch die Verwendung von Umbrüchen im Fragetext nicht mehr syntaktisch valide, sodass 2,56 % der deutschen, bzw. 2,90% der englischen Elemente unbrauchbar waren. Der Datensatz wurde hauptsächlich mit UMLS CUI kodiert, doch wurden auch UMLS-fremde Codes (1.), Platzhalter (2.) oder einfache Codelist-Einträge (3.) gefunden. Zudem wiesen die Annotationen eine Besonderheit auf. Durch die nachträgliche Annotation von zusammengesetzten Elementen, verwendeten die medizinischen Kodierer mehrere Codes um den komplexen Sachverhalt abzubilden (4.). Doch durch die reine Aneinanderreihung war der Zusammenhang schwer nachvollziehbar, bzw. es fehlten Informationen.

Need for unplanned coronary angiography, PCI, or CABG,

C0085532 C1532338 C0010055

Die einzelnen Codes bedeuten:

C0085532 - Coronary angiography

C1532338 - Percutaneous Coronary Intervention

C0010055 - Coronary Artery Bypass Surgery

Abbildung 2.14: Das Beispiel aus (4.) zeigt, wie ein komplexes Studienelement benannt und codiert wurde. Inhaltliche Teile des Elements sind wiedererkennbar, aber der Kontext *Need for unplanned* ist nicht abgebildet.

Die bereinigten UMLS CUIs wurden über die UMLS API abgefragt. Zu den Konzepten wurden einerseits deren SemanticType erfragt und zudem alle weiteren assoziier-

ten Konzepte abgefragt. Die assoziierten Konzepte wurden dann auf die Zugehörigekeit zu den sechs zuvor genannten Vokabularien überprüft und gespeichert. Im Falle eines SNOMED CT-Codes wurden die übergeordnete Top Level Hierachie, im Falle eines LOINC-Konzepts die Parts über einen lokalen FHIR Terminology Server bestimmt.

Q1. Alle fünf gefundenen Vokabularien waren als *Source Vocabulary* in UMLS integriert und konnten den annotierten Codes zugeordnet werden. Jedoch reichen die fünf Vokabularien nicht aus, um alle Elemente zu beschreiben, wie in Tabelle 2.4 zu sehen ist. Die Abdeckung betrug bei den deutschen Elementen 72,93 %, bei den englischen Elementen 67,35 %. Im Weiteren wurden dann die nicht abgedeckten Konzepte betrachtet. Es wurden viele verschiedene, teils ältere Vokabularien gefunden: MEDCIN, Medical Entities Dictionary (CPM) und aber auch oft der NCI Thesaurus (NCIt). Dieser Thesaurus ist eine Referenzterminologie für die onkologische Forschung, welcher krebsbedingte Krankheiten, Befunde und Anomalien einschließt. Die relative Häufigkeit des NCIt lässt sich mit der Ausrichtung des MDM erklären. Ein großer Teil der annotierten Daten stammt aus der Formularen klinischer Studien, ein Gros davon aus onkologischen Studien.

Tabelle 2.3: Die Tabelle zeigt die allgemeinen Angaben des Datensatzes und die Ergebnisse der UMLS-Analyse. Es wurden dabei deutsche und englische Elemente aus dem Datensatz des Münsteraner MDM-Portals betrachtet. Neben ungültigen Annotationen, wie *Code finden*, konnten einige Codes auf Grund von Versionswechel nicht aufgelöst werden.

	Elemente	Elemente nach Reinigung	Elemente mit CUI	Ungültige Annotationen	Unique CUI	Nicht auflösbar
DE	81809	79716	70818	81	17118	573
EN	328708	319174	293161	288	30937	661

Tabelle 2.4: Die Tabelle zeigt die Auswertung bezüglich der untersuchten Kodierungssysteme. Die gefundenen UMLS-Konzepte wurden hinsichtlich der in der Literaturübersichtsarbeit gefundenen Kodiersysteme analysiert. 39% der deutschen und 43% der englischen Elemente waren nicht mit den fünf Kodiersystemen assoziiert.

	SNOMED CT	LOINC	RxNorm	ICD 10 WHO	MedDRA	Keine Annotation
DE	9674 (94%)	2868 (27%)	873 (8%)	1053 (10%)	3376 (32%)	6836
EN	14165 (80%)	3819 (21%)	1321 (7%)	1711 (9%)	5905 (33%)	13371

Q2. Der Datensatz bietet eine ähnlich hohe UMLS-Annotationsdichte für die deutschen (88%) und englischen Elemente (91%), siehe Tabelle 2.3. Da viermal so viele englische Elemente vorhanden sind, spricht das für den außerordentlichen Arbeitsaufwand der Münsteraner Kodierer. Bei den UMLS SemanticTypes war *Finding* in beiden Sprachen sehr dominant (16%, 15%). Die Verteilung der verwendeten SemanticTypes war aber unterschiedlich. Dem Pareto-Prinzip folgend, brauchte man 15 der 133 Types um 80% der deutschen Elemente abzudecken und für die englischen 16 Types. Doch für die komplette Annotierung wurden im Deutschen nur 38 Types benötigt, wohin beim Englischen 51 Types verwendet worden sind. Zudem unterscheiden sich die zehn meist genutzten UMLS CUIs auch in beiden Sprachen. Im Englischen machen die Top 10 fast 71% aller Annotationen aus, angeführt vom UMLS-Konzept *Alter* mit 5301 Verwendungen (17,1%). Im Deutschen hingegen bilden die Top 10 nur knapp 22 % ab und das strukturelle Konzept *Zeitliches Datum* führt mit 651 Verwendungen (3,8 %). Die höhere Anzahl der SemanticTypes deutet daraufhin, dass der englische Datensatz themenspezifisch breiter aufgestellt ist, da er auch viermal größer oder die Kodierung über die Sprachen hinweg nicht konsistent ist. Dafür würde sprechen, dass die Verteilung der anderen verwendeten Vokabularien trotz des Größenunterschieds relativ ähnlich verteilt ist, siehe Tabelle 2.4, und dass die Top 10 CUIs sehr unterschiedlich sind - nur fünf Konzepte sind in beiden Listen. Im Vergleich war die Dichte an SNOMED CT-Codes bei allen Elementen stark erhöht und zudem waren auch alle 19 Top Level Hierachien vertreten. Von den ca. 10.200 Elementen, welche eine Annotierung aus den fünf zuvor untersuchten Vokabularien hatte, war SNOMED CT mit 94% sehr flächendeckend vertreten, wobei es bei den Englischen *nur* 80% waren. Über den gesamten Datensatz gerechnet waren es im deutschen dann 56% und im englischen dahingegen nur 45%. SNOMED CT bietet durch seinen enormen Inhalt eine wirkliche breite Vielfalt zur Beschreibung von medizinschen Inhalten, gerade wenn Beobachtungen in klinischen Studien dokumentiert werden müssen. Am zweithäufigsten war MedDRA vertreten, was durch die regulatorische Prägung gerade bei klinischen Studien zu erwarten war. Ebenfalls zu erwarten war eine Vielzahl von Laboruntersuchungen und daher eine dichte Kodierung mit LOINC, welche aber im Vergleich zu SNOMED CT eher gering ausfiel. Bei der eingehenden Untersuchung der LOINC-Codes fiel auf, dass die UMLS Konzepte zum großen Teil nur einen LOINC-Part beschreiben. Dies bedeutet, dass es kein vollständiger LOINC-Code ist, da mindestens vier weitere Parts fehlen. Desweiteren fiel auf, dass bei 2868 bzw. 3819 LOINC Part Codes im Datensatz nur 1000 bzw. 1119 CUIs mit dem Semantic Type *Laboratory Procedure* beschrieben waren, d.h. mehr

als die Hälfte der annotierten CUIs waren zwar mit LOINC assoziiert, aber nicht als laborspezifische Prozedur klassifiziert. Es ist anzunehmen, dass diese fehlende Majorität als Strukturannotation im Sinne der HL7 Clinical Document Architecture Levels verwendet wurden, doch keiner der LOINC-Parts gehörte der LOINC Document Ontology an. Dies ist eine verpasste Möglichkeiten neben der inhaltlichen, deskriptiven Annotation, auch die Struktur der Formulare semantisch sinnvoll anzugeben. Dahingegen beschrieben viele der gefundenen Annotationen organisch-chemische, pharmakologische oder andere biochemische Substanzen. ICD-10 und RxNorm waren jeweils nicht sehr stark vertreten. Aber es ist erstaunlich, dass die pharmakologischen Substanzen, welche mit LOINC kodiert waren, keine Assoziation zu RxNorm hatten.

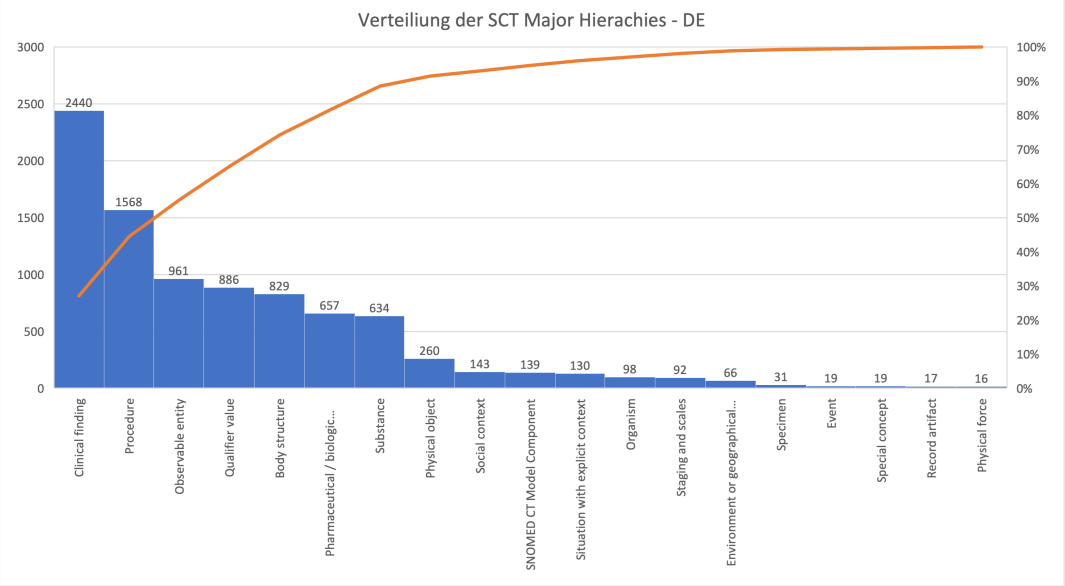


Abbildung 2.15: Die Paretografik zeigt die Verteilung der SNOMED CT Top Level Hierachies über die deutschen Datenelementen. Mit den ersten fünf Hierachies können 80 % aller Datenelemente abgedeckt werden.

Q3. Bei der Wiederverwendung der Annotationen traten zwei Probleme auf, zum einen die Annotationen selbst und zum anderen bzgl. der hinterlegten Kodierungen. Die Annotationen selbst waren teilweise einfach nicht auflösbar bzw. nachvollziehbar, siehe Tabelle 2.2. Gerade die Abbildung komplexer Inhalte sollte nicht einfach durch eine Aneinanderreihung von Codes bestehen, sondern in einem nachvollziehbaren und am besten maschinell verarbeitbaren Format. Denkbar wäre hier die Verwendung von postkoordinierten SNOMED CT Ausdrücken. Durch die ontologische Basis lassen sich sehr

ausdrucksstarke und maschinell-berechenbare Konzepte erstellen. Dafür müssen die jeweiligen Teilausdrücke auch in SNOMED CT darstellbar sein. Um dies zu verdeutlichen, wurden die gefundenen komplexen CUI-Ketten untersucht und als postkoordinierter SNOMED CT Ausdruck abgebildet, siehe Tabelle 2.5. Die ersten drei Ketten konnten per Postkoordinierten Ausdrücke (PCE) dargestellt werden, wobei der erste Ausdruck etwas an Informationen verloren hat. Es gab zwar Substanzen mit Reverse-Transkriptase-Inhibitor in SNOMED CT, aber nicht die Möglichkeit, sie näher als nicht-nukleoside zu beschreiben. Dafür war erstaunlicherweise kein Konzept in SNOMED CT vorhanden. Für den vierten Ausdruck war es nicht möglich einen validen PCE zuerstellen. Die Teilkonzepte waren zwar vorhanden, doch gab es nach dem SNOMED CT Concept Model, keine Möglichkeit diese konform zu verbinden. Die letzte Kette war eher eine Aneinanderreihung von Konzepten, welche nicht durch einen PCE auszudrücken war, da es in der Compositional Grammar keine logische Veroderung gibt. Dahingegen konnte der Ausdruck als ECL-Term dargestellt werden, da hier durch die Veroderung eher eine Menge an Konzepten beschrieben wurde. Es folgt daraus, dass eine Postkoordination der CUI-Ketten möglich ist, aber nicht alle beschriebenen komplexen Zusammenhänge durch SNOMED CT beschrieben werden können.

Tabelle 2.5: Die Tabelle zeigt die gefundenen komplexen CUI-Ketten und die korrespondierenden SNOMED CT Postkoordinationsausdrücke. Die ersten drei Ketten könnten übertragen werden. Für die letzte und eher einfachere Kette konnte kein PCE gefunden werden, da zwar die Teilkonzepte vorhanden sind, aber es keine Möglichkeit gibt, sie geeignet zu verbinden.

	CUI Ketten	SNOMED CT PCE	Freitextliche Beschreibung
1.	C0205225 C1373120 C0013203	31438003: 47429007=372531002, 719722006=63161005	Erstes Auftreten einer Medikamentenresistenz gegen (nicht-nukleoside) Reverse-Transkriptase-Inhibitor-Mittel
2.	C0042313 C0150349	386439008: 424361007=42082003	Topische Hautpflege mit einem Vancomycin enthaltendes Produkt
3.	C0598463 C0332162	364665006: 370130000=77374008	Beginn des Funktionsstatus
4.	C0011008 C3544287		Tag der Transplantation einer Leber entnommen eines Lebendspenders
5.	C0019004 C0031139	<<302497006 OR 71192002	Peritonealdialyse oder Hämodialyse

Das zweite große Problemfeld waren die hinterlegten Kodiersysteme. Die reine Angabe eines UMLS CUI ist nicht ausreichend, da die entscheidende Versionierung fehlt. In der Literaturanalyse wurde als ein Problem bei der Metadatenprozessierung (siehe 2.2.2) das Veralten von Informationen beschrieben, welchem man durch händische Kuratierung oder eine semantische Kodierung entgegentreten kann. Doch es zeigt sich, dass die Annotierungen selbst veralten und durch die fehlende UMLS Versionsangaben nicht mehr nachvollziehbar sind. In der Analyse konnten die Annotierungen von 573 deutschen und 661 englischen Elementen nicht mehr nachvollzogen werden. Dies sind nur verschwindend geringe 3,34 % bzw. 2,12 % der gesamten Annotationen, doch stetig wachsend. Das bedeutet, dass die händische Annotierung von diesen Elementen mittlerweile wertlos ist und wiederholt werden müsste. Aus der vorangegangenen Analyse des MDM-Datensatzes lassen sich Empfehlungen für die Annotierung von Datenelementen mit Konzepten aus medizinischen Terminologiesystemen ableiten. Prinzipiell ist ein Generalist wie UMLS durch seinen extrem breiten Wortschatz gut für die nachträgliche Annotierung von Metadaten geeignet, da die zu annotierenden Datenelemente und deren Ausdruckstiefe sehr variabel sind. Die Analyse zeigte, dass die fünf spezialisierten Terminologiesysteme nicht ausreichend sind, um alle Elemente abzudecken. Doch gleichwohl muss bei der Annotation mit UMLS mit Vorsicht kodiert werden und auf Assoziationen zwischen Codes geachtet werden. Dadurch kann die Ausdrucksvielfalt der spezialisierten Terminologiesysteme besser genutzt werden. Ein Beispiel dafür ist die Verwendung von LOINC Document Ontology Codes, welche bei der Analyse nicht gefunden worden sind. Dabei ergibt die Nutzung der Document Ontology gerade für die Kodierung von Elementgruppen und Dokumentsektionen Sinn, um eine Gruppierung geeignet semantisch zu beschreiben. SNOMED CT hatte unter allen fünf Terminologiesystemen die höchste Abdeckung und bietet durch die Verwendung von PCEs und ECL mächtige Beschreibungswerkzeuge. Doch wie in Tabelle 2.5 gezeigt sind nicht alle Inhalte dadurch zu beschreiben. Aber der Einsatz von PCEs ist für die Annotierung von Elementen mit einer einzelnen Bedeutung sinnvoll, wohingegen ECL für Elemente genommen werden sollte, wo eine bestimmte Auswahl besteht. Dies ist beispielsweise bei der 5. CUI-Kette in Tabelle 2.5 zu sehen oder aber auch in Auswahlelementen, die eine definierte und endliche Menge an Antwortoptionen vorsehen. Zuletzt bleibt eine dringende Empfehlung, egal ob Generalisten oder Spezialisten verwendet werden: die Versionsangabe des genutzten Terminologiesysteme sollte verpflichtend dem Code zugeordnet werden. Sonst droht wie im Falle des MDM-Datensatz der Verlust von Annotationen und damit verbundenen eine Gefährdung der Wiederverwendbarkeit der klinischen Metadaten.

Kapitel 3

Metadatenintegration - Konzept und Implementierung

In diesem Kapitel werden zuerst die Anforderungen an föderierte Metadatenstrukturen untersucht. Aufbauend auf der Anforderungsanalyse wird der Bedarf nach solchen Strukturen durch eine strukturierte Umfrage abgeklärt und bestehende MDR-Systeme untersucht. Zudem wurde aus den Anforderungen eine neue Kommunikationsschnittstelle speziell für die Kommunikation von Metadaten Repositories und föderierter Metadatenstrukturen konzipiert [Ulrich et al., 2019b, Ulrich et al., 2020b] und Integrationsszenarien für die bestehende MDR-Systeme beschrieben [Ulrich et al., 2018a, Ulrich et al., 2018b, Ulrich et al., 2019a].

Der oft gesetzlich geforderte standortübergreifende Austausch von Gesundheitsdaten wird durch die Heterogenität der eingesetzten IT-Systeme mit proprietären Datenformaten und Schnittstellen erheblich erschwert. Eine Vielzahl an Standards im Gesundheitswesen, wie HL7 v2 für klinische Daten oder DICOM für Bilder, stellen auf technischer Ebene nützliche Nachrichtenprotokolle zur Verfügung. Diese Standards sind aber aufgrund unterspezifizierter Nachrichtenschemata und vernachlässigten Standardterminologien kaum geeignet die Anforderungen an semantische Interoperabilität zu adressieren [Benson and Grieve, 2016]. In jüngerer Zeit werden international fortgeschrittene Standards wie HL7 FHIR oder openEHR verwendet, um die enorme Vielfalt und Granularität medizinischer Inhalte durch eingeschränkende Informationsmodelle mit Verknüpfungen zu Standardterminologien wie SNOMED CT zu spezifizieren. Zur Validierung der kommunizierten Inhalte gemäß den Spezifikationen stehen Werkzeuge inklusive leistungsfähiger Terminologieserver zur Verfügung [Benson and Grieve, 2016].

Vor diesem Hintergrund lassen sich zwei Ansätze zur Harmonisierung medizinischer Daten unterscheiden, die beide ihre Daseinsberechtigung haben: Top-Down (TD) versus Bottom-Up (BU). Der zentrale Bestandteil des Top-Down-Ansatzes ist der Standardisierungsprozess zur Erstellung von Kerndatenspezifikationen. Der Erstellungsprozess ist zeit- und ressourcenintensiv und beinhaltet aufgrund des aufwändigen Abstimmungsprozesses und der daraus resultierenden Kommunikation diverse Rückkopplungsschleifen, um die Anwendbarkeit an allen Standorten und Organisationsarbeiten sicherzustellen. Diesem Aufwand steht jedoch die gesicherte Qualität eines Kerndatensatzes durch die iterativen Feedback- und Korrekturschleifen gegenüber. Alle Beteiligten müssen sich auf den gemeinsamen Minimaldatensatz, bzw. ein gemeinsames Schema einigen, was zu einem möglichen Informationsverlust gegenüber der eigenen lokalen Datenschemata führt. Um den enormen Aufwand zu reduzieren, werden die harmonisierten Datensätze oft an bestehende Standards gebunden, um internationale Vergleichbarkeit, gesicherte Funktionalität und Adaption zu gewährleisten. Prominente Beispiele sind die stetig wachsende Verbreitung von HL7 FHIR und openEHR – national wie international. Ein wesentlicher Vorteil der Übernahme etablierter Standards ist die Verwendung bewährter Modellierungswerkzeuge und -methoden, wie z.B. der webbasierte *Clinical Knowledge Manager* (CKM) aus der openEHR-Community [Wulff et al., 2018].

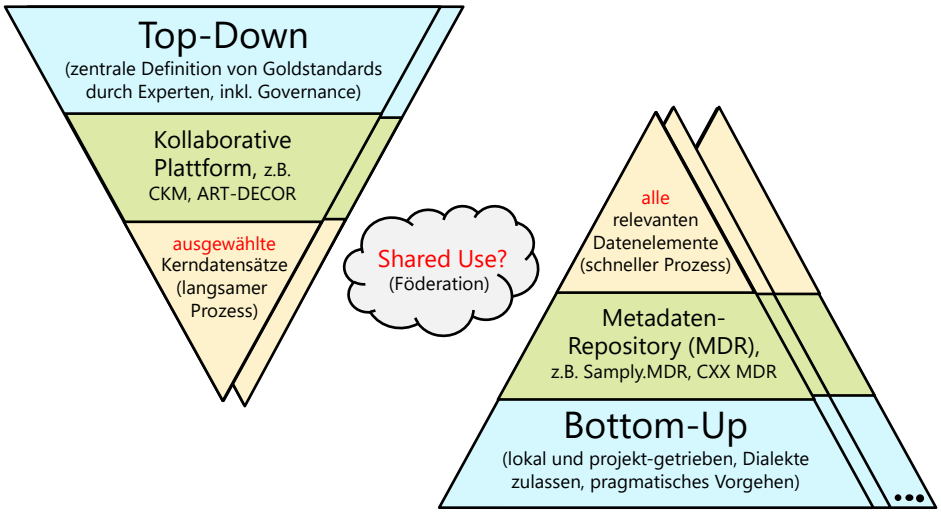


Abbildung 3.1: Schematische Unterschiede zwischen dem Governance-getriebenen **Top-Down**- und dem forschungsgetriebenen **Bottom-Up**-Ansatz. Beide ermöglichen die vernetzte Nutzung und Föderation von Metadaten auf unterschiedlich effektive Weise.

Dem Top-Down-Ansatz steht ein dynamischer Bottom-Up-Ansatz gegenüber, der alle relevanten Datenelemente einbezieht, um deren Metadaten projektspezifisch zu definieren. Im Gegensatz zu den in Abbildung 3.1 skizzierten Top-Down-Verfahren werden in zahlreichen Forschungsprojekten die den Forschungsdaten entsprechenden Metadaten auf lokaler Ebene erstellt und gepflegt. So werden in Kohorten- und Registerstudien die Datenelemente für eigene Zwecke definiert [Stausberg and Harkener, 2019], z. B. mittels *Data Dictonaries* in Tabellenkalkulationsprogrammen oder in projektspezifischen Metadaten-Repository-Systemen. So stehen sich harmonisierte Schemata im TD und lokale atomisierte Dataelemente im BU gegenüber. Insbesondere ISO 11179-konforme MDRs stellen beim BU sicher, dass Metadaten strukturell und semantisch auf Basis eines lokal vereinbarten Schemas beschrieben und dargestellt werden [Ulrich et al., 2016]. Die Lücke zwischen den von Domänenexperten bevorzugten Schemata gegenüber dem MDR kann mit speziellen Nutzer-orientierten Verfahren abgedeckt werden [Kock-Schoppenhauer et al., 2019b]. Da beim Bottom-Up-Ansatz die zeitaufwändige, aber auch qualitätssichernde Harmonisierung des Top-Down-Ansatzes entfällt, muss der Erhebungskontext lückenlos dokumentiert werden. Der Vorteil des Bottom-Up-Ansatzes ist die individuelle Abbildung der vorhandenen, projektspezifischen Metadaten und der entsprechenden Datensätze [Mate et al., 2019b]. Der Informationsverlust gegenüber dem Minimaldatensatz-Prinzip wird reduziert, und durch die direkte Integration der Einzeldaten wird eine höhere Informationsabbildung erreicht. Gegenstand dieser Arbeit ist eine Forschungsdaten-getriebene Integration in Sinne des Bottom-Up-Konzept.

3.1 Anforderungen an föderierte Strukturen für Metadaten

Ziel der Metadatenintegration sind lokale, projektspezifische Schemata, welche an verschiedenen Standorten in MDRs erstellt und verwaltet werden. Im Gegensatz zu den Top-Down-definierten Schemata sind die Informationen in lokalen MDRs viel näher an den projektspezifischen Datensätzen. Sie dienen betreibenden Standorten in erster Linie als Werkzeug für die Datenverarbeitung, wie automatische Instanzdatenvalidierung oder Formulargenerierung für klinische Studien. Daher sind die lokal gespeicherten Metadaten für einen datenintegrationsgetriebenen Ansatz von hohem Wert. Um diese Metadaten und damit auch die dazugehörigen Instanzdaten in großem Rahmen verfügbar und

verarbeitbar zu machen, müssen die Informationen zuerst lokal identifiziert und dann verfügbar gemacht werden.

Der Aufbau eines nationalen Metadata Repositories gehört seit einigen Jahren zu den strategischen Zielen der medizinformatischen Gemeinschaft und wurde dementsprechend häufig in deren Arbeitsgruppen und Tagungen thematisiert. Für den datengetriebenen Informationsaustausch wie in der Medizininformatik-Initiative [Semler et al., 2018] oder anderen Verbundsprojekten ist der Einsatz von MDRs unerlässlich. Die Vorbereitungen zur Etablierung eines MDRs für die klinische und epidemiologische Forschung begannen in der Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. (TMF) bereits 2008. Aus einer grundsätzlichen Bedarfs- und Strategiediskussion heraus wurde ein erstes TMF-Projekt „Spezifikation eines Metadata Repository der Telematikplattform für Medizinische Forschungsnetze (TMF) e. V.“ erfolgreich durchgeführt. Die Erkenntnisse des Projektes führten zu weiteren Projekten, aber nach Auswertung und Dissemination der Projektergebnisse verstetigte sich der Fokus auf eine übergreifende nationale Metadaten-Erfassung und es wurde deutlich, dass ein Paradigmenwechsel erforderlich ist. Es stellt sich als unrealistisch heraus, dass alle Standorte ihre lokal gepflegten Metadaten in einem nationalen Metadaten Repository bereitstellen [Dugas et al., 2015]. Mit Blick auf den Brückenkopf-Ansatz [Lablans et al., 2015], der sich für das Teilen von Forschungsdaten (im Sinne von Instanzdaten) etabliert hat und die Daten dabei in der Hoheit lokaler Forscher belässt, ist erkennbar, dass auch für die Metadaten-Ebene verteilte Ansätze gefunden werden müssen. Die gepflegten Datenelemente müssen auf den lokalen MDRs heraus verfügbar sein, da nur so die dazugehörigen Instanzdaten auffindbar und verfügbar gemacht werden können. Die heutigen MDRs sind aber technisch heterogen: Sie sind unterschiedlich aufgebaut und die Schnittstellen heterogen strukturiert, falls überhaupt vorhanden. Daher müssen für einen verteilten Ansatz neue Konzepte erarbeitet werden. Um allen Anforderungen und möglichen Einschränkungen eines föderierten Metadatenansatzes gerecht zu werden, soll eine Anforderungsanalyse [Lightsey, 2001] detailliert die Voraussetzungen herausarbeiten. Die Analyse ist dabei in drei Gruppen mit weiteren Subeigenschaften unterteilt: operative, funktionale und architektonische Anforderungen, welche im Nachfolgenden ausführlich beschrieben werden.

Operative Anforderungen

Die operativen Anforderungen beschreiben welches übergeordnete Ziel die Anwendung hat und in welchem Einsatzgebiet, bzw. mit welchen Einschränkungen, das System diese erfüllen und leisten muss.

O.1 **Aufgabenprofil:** Ein föderierter Ansatz muss primär sicherstellen, dass die technisch sehr diversen MDR-Systeme abfragbar werden und das themenspezifische Metadaten für die Anwendung von Diensten und Auswertungen bereitgestellt werden. Ziel muss es sein die Siloisierung der abgeschottenen MDRs zu überwinden und die Verarbeitung der korrespondieren Datensätze nachhaltig zu vereinfachen. Änderungen, Anreicherungen oder Verarbeitungen der Metadaten sollen bestmöglich dem Quellsystem zurückkommuniziert werden.

O.2 **Operative Umgebung:** Die lokalen MDR-Systeme werden meist in Forschungseinrichtungen betrieben und dienen in erster Linie der Verwaltungshilfe von klinischen Daten für die dortige Forschung. Die Metadaten verbessern die Klassifizierung und Identifikation der korrespondierenden Forschungsdaten im Sinne der Wiederverwendbarkeit, siehe Kapitel 2.2.2. Das Einsatzgebiet ist in erster Linie forschungsgetrieben, der Einsatz in der Patientenversorgung ist eher zweitrangig. Die Systeme werden von verschiedenen Nutzergruppen verwendet. Man muss dabei technisch versierte *Data Stewards* und anwendungsorientierte Domänenexperten unterscheiden - wie in Kapitel 2.2.2 beschrieben. Die Metadaten am Standort weisen eine hohe Nähe zu den korrespondierenden Forschungsdaten auf, d.h. die Metadaten sind aktuell und im Einklang zum Datenkörper. Dies ist entscheidend, da die lokalen Schemata sich gegenüber nationalen TD-Formaten schneller ändern und die Integrität der Metadaten für die Datenintegration höchst entscheidend ist. Technisch gesehen sind die lokalen MDR-Systeme sehr heterogen. Zudem sind lokale MDRs eigenständige Instanzen mit selbstbestimmter Organisation. Ein Ausfall oder permanenter Austritt aus einem Verbund ist jederzeit möglich.

O.3 **Beschränkungen:** Es gilt immer die lokale Datenhoheit zu wahren. So muss an jedem Standort gesondert die Urheberrechtslage der Daten abgeschätzt werden, da die MDR-Betreiber und domänenspezifischen Lizenzhalter nicht immer der gleiche Personenkreis sind. Um eine gewinnbringende Erschließung und Weiterverarbeitung

zu sichern, ist eine Bereitstellung der Daten unter einer Creative Commons Lizenz wünschenswert. Ein Verfahren, welches an das MDM (siehe Kapitel 2.3) angelehnt ist, sollte in Betracht gezogen werden: Die Daten werden dort unter Nennung des Lizenzgebers für die Forschung frei zur Verfügung gestellt. Somit ist ein weiterer wichtiger Aspekt die eigentliche Bereitschaft zum Teilen und Bereitstellen der Metadaten von Seiten der Lizenzhalter. Die Metadaten sind abgeleitet von den Forschungdaten, sodass die im MDR gespeicherten Informationen auch Rückschlüsse auf die betriebene Forschungsaktivitäten geben können. Beispielsweise könnte ein Studienformular mit spezifischen Messparametern auf die Forschungshypothese Aufschluss geben. Problematisch ist dabei das *Zurückspielen* von extern geänderten oder (semantisch) angereicherten Metadaten, da hier die Datenhoheit ganz klar beim Lizenzhalter liegt und im Zweifelfall eine inhaltiche Änderung der Daten vorliegen könnte. Damit einhergehend gibt es auch kein einheitliches Verständnis über Nutzen und Bedeutung an den Standorten - gerade unter den Domänenexperten und Lizenzhaltern. Neben den rechtlichen ist ein föderierter Ansatz von weiteren organisatorischen Gegegenheiten abhängig: das lokale Authorizations- und Authentifikationssystem muss technisch ansprechbar und integrierbar sein.

Funktionale Anforderungen

Die funktionalen Anforderungen beschreiben, welchen Funktionsumfang das profilierte System erfüllen muss. Es werden zudem Anforderungen an die Schnittstellen und die Leistungsfähigkeit des Systems gestellt.

F.1 **Systemfunktionen:**

Ein föderiertes System soll in erster Linie Metadaten aus den lokalen *Silos* auslösen und nutzbar machen. Dafür müssen die Daten zuerst identifizierbar sein und dann auch abrufbar. Die Identifikation ist einerseits technisch über eindeutige Kennziffern, beispielsweise UUIDs zu realisieren. In aktuellen Systemen findet eine Identifikation nur auf Elementebene statt. Wünschenswert ist aber eine tiefergehende Identifikation bis hin zur Wertebene. Anderseits müssen die Daten auch inhaltlich über semantische Kodierungen klassifizierbar sein. Durch die lokale Verfügbarkeit an den verschiedenen Standorten, können die Daten auch in einem externen föderierten Konzept verarbeitet und zusammengesetzt werden. Dafür ist gerade die inhaltli-

che Identifikation entscheidend, da so inhaltliche Korrespondenzen oder Dubletten erkannt werden können. Ergebnisse aus der föderierten Verarbeitung soll an die lokalen Standorte zurückgegeben werden, um auch dort die (Meta)Datenqualität zu erhöhen - gleichwohl hier die lokale Datenhoheit gewahrt bleiben muss.

F.2 **Systemperformance:** Die Abfrage der lokalen Systeme sollte automatisch und ohne menschliche Interaktion erfolgen. Zudem sollte die Integration in einen föderativen Verbund nicht die lokalen Prozesse behindern, verändern oder die Performance verschlechtern.

F.3 **Systemschnittstellen:** Das System soll eine Schnittstelle definieren, welche dann von allen lokalen MDR-Systemen integriert und implementiert wird. Diese Schnittstelle soll auch in einer föderierten Umgebung Metadaten abfragbar machen. Die Schnittstelle soll keine technischen Annahmen über das darunterliegenden lokale MDR-System machen. Sie soll die Kommunikation vereinheitlichen, nicht aber Vorschriften zur Datenhaltung machen. Vorteilhaft ist in dem Sinne ganz klar eine ISO-kompatible Datenhaltung, aber auch nicht standardisierte MDR-Systeme sollen integriert werden. Um dies zu unterstützen, soll es eine maschinenlesbare Angabe über den implementierten Funktionsumfang geben, da es wahrscheinlich ist, dass nicht alle lokalen MDR-Systeme den gleichen Funktionsumfang anbieten können.

F.4 **Systembeschränkungen:**

Die lokalen MDR-Systeme müssen direkt erreichbar sein, um die Anfragen entgegen zu nehmen. Eine indirekte Kommunikation über *Message Broker* ist aufgrund der geringeren Datenschutzanforderungen nicht nötig.

Architektonische Anforderungen

Die Systemarchitektur des föderierten Verbunds ist maßgeblich vom Zusammenspiel der lokalen MDR-Systeme abhängig. Es wird ein hierarchisches und dezentrales Client-Server-Modell angestrebt, wobei alle MDR-Systeme als Server über die gleiche Schnittstelle die Metadaten bereitstellen und die Clients ebenfalls über diese Schnittstelle die Daten konsumieren und produzierte Verarbeitungsergebnisse zurückkommunizieren können. Durch die lose Kopplung aller beteiligten Komponenten kann ein Ausfall

von MDR-Systemen besser abgefangen werden, welcher durch die Autonomie und Datenhoheit der MDR-Systeme immer in Betracht gezogen werden muss. Zudem können sich MDR-Systeme dann an verschiedenen Verbünden beteiligen, um beispielweise Forschungsdaten in Projektkonsortien mit Partnern besser teilen zu können. Die Kommunikation soll netzwerkbasiert über etablierte Web Technologien erfolgen, vornehmlich REST oder ähnliche Maschine-zu-Maschine-Kommunikationsparadigmen. Die technologische Basis ist stark heterogen, da die lokalen MDR-Syteme schon im produktiven Einsatz sind und man dadurch keinen Einfluss auf technische Aspekte hat. Daher sollte das Verbundssystem in Sinne der losen Kopplung plattform-unabhängig und sprachagonstisch sein, bzw. Paradigmen und Tools nutzen, welche in allen großen Programmiersprachen verfügbar sind.

3.2 Metadaten-Akquise

Der Nutzen der angestrebten föderierte Metadatenabfragen ist abhängig von der Bereitschaft der standortspezifischen MDR-Betreiber, sich im Sinne eines föderierten Verbunds anzuschließen. Um diesbezügliche Bedarfe und potenzielle Hürden zu ermitteln, wurde eine strukturierte Umfrage durchgeführt. Die Umfrage beinhaltete 12 Fragen, gruppiert nach drei Themenbereichen: Nutzen von Metadaten im lokaler Einsatz, Bedarf und Nutzen in der Wiederverwendung von bestehenden Metadaten und zuletzt die Bereitschaft zur Beteiligung, bzw. dem Teilen der eigenen Metadaten. Zudem wurden verschiedene Verwendungsmöglichkeiten skizziert und nach weiteren Einsatzfeldern gefragt. Zielgruppe der Befragung waren MDR-Betreiber aus dem DACH-Raum, welche mit Hilfe der TMF kontaktiert wurden. Die Befragung erreichte 121 Einzelpersonen, wovon 21 die Umfrage vollständig abschlossen.

Die Ergebnisse fielen durchweg positiv aus und es zeigte sich, dass in der medizinischen bzw. medizin-informatischen Community ein steigender Bedarf für die Vernetzung von MDRs bestand. Es wurde beschrieben, dass Metadaten schon konsequent definiert werden und zugleich einen hohen Nutzen in den Projekten brachten, siehe Abbildung 3.2. Der weitere Nutzen und Einsatz von vernetzten, bzw. föderierten Metadaten wurde ebenfalls als *hoch* beschrieben, im Speziellen auch der Einsatz der FAIR-Prinzipien [Wilkinson et al., 2016]. Die Prinzipien empfehlen den Einsatz von erreichbaren Metadaten zur Identifikation und Beschreibung von wissenschaftlichen Datensätzen zur Förderung

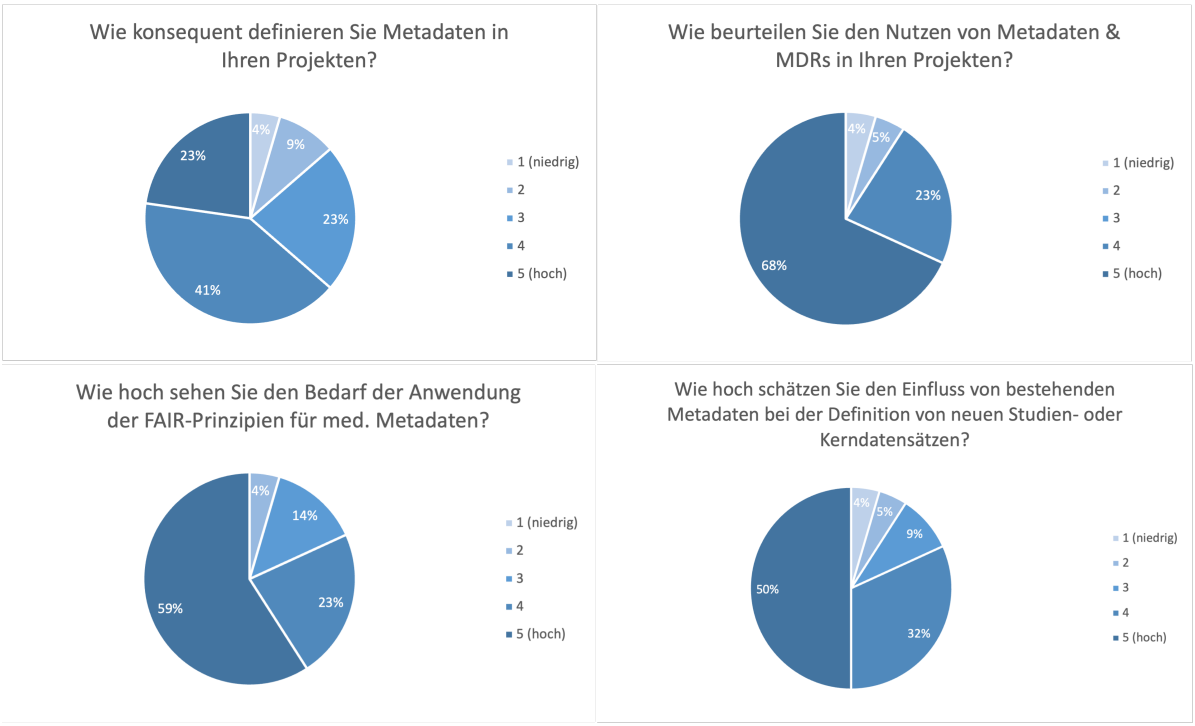


Abbildung 3.2: Die vier Grafiken zeigen die Auswertung der Betreiberumfrage. Die ersten vier Fragen zielte auf die Relevanz von Metadaten und deren Einsatz in standortspezifischen Projekten.

der Reproduzierbarkeit. Neben der Signifikanz von Metadaten in lokalen Projekten wurde auch die Bereitschaft zur Vernetzung lokaler MDRs und die Anzahl der von einer solchen Vernetzung betroffenen Projekte erfragt, siehe Abbildung 3.3. Die Umfrage ergab darüber hinaus, dass Informationen aus ca. 1.500 Einzelprojekten durch die Integration in eine föderierte Metadaten-Abfragen nutzbar werden könnten.

Alle vorgeschlagenen Verwendungsmöglichkeiten (Qualitätsanalyse, Harmonisierung und Annotation) fanden positiven Anklang bei den Befragten, zugleich wurden noch zwei weitere Anwendungsfelder angebracht: *shared use* und generische ETL-Prozesse. Die gemeinsame Nutzung von Metadaten ist im Konzept der föderierten Bereitstellung zentral, sodass dieser Anwendungsfall ein definitives Ziel war. Die Unterstützung generischer ETL-Prozesse findet seinen Einsatz in der Datenintegration und war somit ebenfalls Zielstellung der Arbeiten.

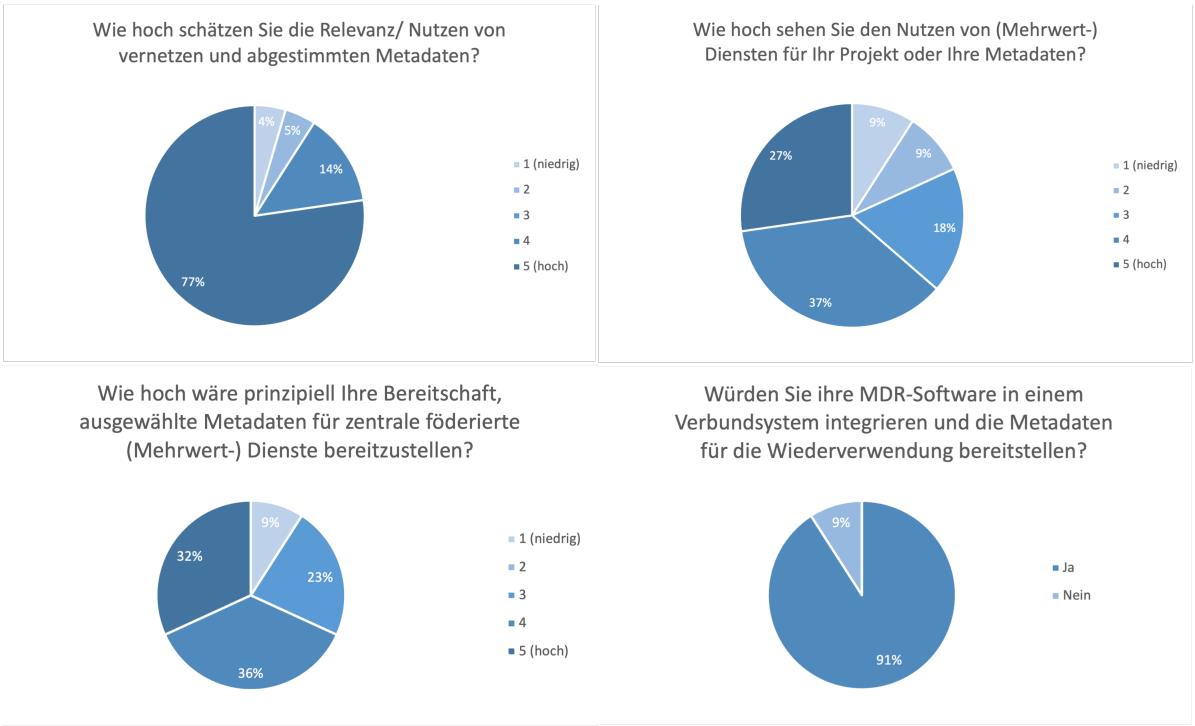


Abbildung 3.3: Die vier Grafiken zeigen die zweite Auswertung der Betreiberumfrage. Diese Fragen sollten die Bereitschaft zur Partizipation der MDR-Betreiber erfragen.

3.3 Bestandsanalyse von MDR Systeme

Für die Übersicht über die aktiven MDRs wurde eine systematische Literatursuche durchgeführt [Ulrich et al., 2019a]. Als Suchmaschine wurde PubMed verwendet und mit dem Term *metadata repository AND healthcare* Arbeiten aus den letzten sechs Jahren (2015-2021) gesucht, zudem wurden bekannte kommerzielle Systeme hinzugefügt. Die Arbeiten, bzw. die Systeme wurden dann nach folgenden Auswahlkriterien gewählt: medizininformatisch relevant, frei zugänglich und aktiv genutzt und gewartet. Es kamen nur Systeme in Betracht, welche relevante Informationen aus dem Gesundheitswesen beinhalten könnten, daher wurde nach einer medizininformatischen Verwendung aussortiert. Zudem sollten die Systeme auch noch aktiv genutzt und die Projekte - im Speziellen Open Source Systeme - in fortlaufender Entwicklung und damit verbunden Wartung sein. Abschließend sollten die Systeme zugänglich sein, d.h. für die nachfolgende Untersuchung auch überhaupt verfügbar sein. Dies war gerade im Hinblick auf kommerzielle Produkte ein mögliches Auschlusskriterium. Die Ergebnisse der Litera-

tursuche sind in Tabelle 3.1 dargestellt. Neun Systeme wurden in der Literatursuche gefunden, wobei zwei anhand der Kriterien aussortiert wurden. Die *Metadata Online Registry* (METeOR) [Australian Institute of Health and Welfare, 2021] des australischen Gesundheitsministerium war zwar verfügbar, doch waren die Daten stark veraltet und daher aufgrund der fehlenden Nutzung ausgeschlossen. Das SemanticMDR [Daniel et al., 2014], welches im SALUS-Projekt entwickelt wurde, war als Open Source-Variante zwar verfügbar, doch das Projekt wird nach Aussage der Entwickler nicht mehr aktiv gepflegt. Die kommerzielle Variante war aber leider nicht verfügbar. Dahingegen konnten sieben MDRs eingeschlossen werden. Der *Common Data Element Browser* [Covitz et al., 2003] wird vom US-amerikanischen *National Cancer Institute* getrieben und bietet Zugang zu den Metadaten aus dem Cancer Data Standards Registry and Repository (CaDSR). Dieses Repository stellt einen standardisierten Datensatz für die Krebsforschung im US-amerikanischen Gesundheitssystem bereit. Das Aristole MDR ist der Nachfolger des veralteten METeOR und bietet eine Vielzahl von kuratierten Metadaten aus dem australischen Gesundheitssystem. Das MDM Portal der Universität Münster sammelt medizinische Formulare und stellt diese der nicht-kommerziellen medizinischen Forschung bereit. Die Formulare werden händisch mit UMLS-Codes annotiert und in der Universitätsbibliothek der Universität Münster archiviert, zudem kann zu jedem Formular eine permanente DOI beantragt werden. ART-DECOR ist ein kollaboratives MDR, welches gerade im DACH-Raum für die Abstimmung von medizischen Formaten genutzt wird. So wurde die österreichische elektronische Gesundheitsakte (ELGA) über ART-DECOR harmonisiert und abgestimmt [Ott et al., 2019]. Zudem nutzt die Medizininformatik Initiative (MII) ART-DECOR für die Abstimmung der Kerndatensätze und das MII-Konsortium SMITH das System für ihr Metadatenmanagement. Das Samply.MDR ist ebenfalls im DACH-Raum in Forschungsprojekten verbreitet und findet in der nationalen, wie auch europäischen Biobankenvernetzung, sowie im MII-Konsortium MIRACUM Verwendung. Das kommerzielle CXX MDR von Kairos war im Rahmen einer Forschungskooperation in einer Teststellung vom Unternehmen für einen Evaluationszeitraum bereitgestellt worden und war seit Dezember 2020 für die Untersuchung verfügbar. Das CXX MDR wird im MII-Konsortium DIFUTURE für die Metadatenverwaltung eingesetzt. Der Clinical Knowledge Manager (CKM) ist ein System für die gemeinsame Entwicklung, Verwaltung und Veröffentlichung von openEHR-Archetypen. Zu den verwalteten Ressourcen gehören neben den Archetypen, auch Templates und Metadaten zu klinischen Modellen.

Tabelle 3.1: Die Suche im PubMed untersuchte MDR Systeme, welche im Zeitraum von 2015-2021 in der Literatur erwähnt worden sind. Neun Systeme wurden gefunden und nach drei Kriteriern für die weitere Untersuchung im Rahmen dieser Arbeit ausgewählt. Die beiden grau hinterlegten Systemen wurden auf Grund der fehlenden aktiven Nutzung nicht weiter betrachet.

	Medizininformatisch relevant	**frei zugänglich**	**aktiv genutzt / gewartet**
Aristotle	✓	✓	✓
MDM Portal	✓	✓	✓
ART-DECOR	✓	✓	(✓)
Samply.MDR	✓	✓	✓
Clinical Knowledge Manager	✓	✓	✓
Common Data Element Browser	✓	✓	✓
Kairos MDR	✓	(✗)	✓
METeOR	✓	✓	✗
Semantic MDR	✓	(✓)	✗

Authentifizierungskonzepte der MDRs

In der Anforderungsanalyse für Metadaten im föderierten Kontext wurden als organisatorische Beschränkungen die Verwendungsmöglichkeiten lokaler Authorizations- und Authentifikationssysteme herausgearbeitet. Daher wurden die zuvor gefundenen MDR-Systeme auf die eingesetzten Authorizations- und Authentifikationsmechanismen untersucht und die technischen Verwendungsmöglichkeiten näher betrachtet.

- Aristotle basiert auf dem Python-Web-Framework Django und nutzt die internen Authentifikationmechanismen für das User Management. Die erstellten Metadaten sind über eine API per API-Token abrufbar. Es können verschiedene Token mit differenzierten Berechtigungen (Lesen vs. Schreiben) individuell vergeben werden. Es gibt keine Möglichkeit das User Management von außen abzufragen. Öffentliche Metadaten sind ohne Anmeldung abrufbar.

- Das MDM-Portal implementiert ein eigenes User Management auf Basis von Spring Security. Es gibt eine API zur Abfrage der Metadaten, welche ebenfalls einen Token zur Authentifizierung nutzt. Der Token ist aber nicht selber erstellbar und wird nur

von den Betreibern vergeben. Ein Abrufen der Metadaten ist ohne eine vorherige Anmeldung nicht möglich.

- ART-DECOR setzt laut der eigenen Dokumentation JSON Web Token ein. Es ist jedoch nicht ersichtlich, wie man sich dort registrieren kann.

- Samply.MDR nutzt eine externe Authentifizierung über OpenID Connect (OID), welches auf standardisierten Token basiert, die im Payload die Berechtigungen der Nutzer enthalten. Öffentliche Metadaten sind ohne Anmeldung abrufbar, die Erstellung oder Änderung ist über die Rest-API nicht möglich.

- Der Clinical Knowledge Manager implementiert ein eigenes User Management. Die API ist entweder über Basic Auth oder einen Token abrufbar. Öffentliche Metadaten sind ohne vorheriges Anmelden abrufbar. Für die Erstellung oder Veränderung ist eine Authentifizierung nötig.

- Der Common Data Element Browser nutzt die Benutzerverwaltung von UMLS, welches wiederum die Verwaltung an verschiedene Identity Provider wie Google, Facebook, Microsoft delegiert und auch ein Institutslogin anbietet. Dahinter verbirgt sich der eduGAIN-Föderationsdienst, der die Identitätsverbände für Forschung und Bildung weltweit miteinander verbindet. Für die Authentifizierung von API-Anfragen wird das Kerberos-Protokoll eingesetzt [Miller et al., 1988].

- Die Nutzerverwaltung des Kairos MDR ist über OAuth2 implementiert, wobei das MDR als Authentication Server und Resource Server gleichzeitig fungiert. Dadurch können externe Programme als Clients registriert werden und auf die Daten zugreifen. Den Nutzern können dabei sehr granulär Lese-, bzw. Schreibrechte zugeteilt werden.

Die Analyse zeigte, dass 6 von 7 Systemen eine Authentifizierung per API-Token oder durch OAuth2 per *JSON Web Token* (JWT) von den MDRs implementiert, bzw. delegiert wurde. Einzig der Common Data Element Browser setzt auf eine Authentifizierung über das Kerberos-Protokoll [Miller et al., 1988]. Ein API-Token dient der Ersetzung des klassischen User/Passwort-Logins und werden entweder als URL Parameter oder im HTTP Authentification Header übertragen. Sie dienen der reinen Authentifizierung und haben keinen weiteren Payload, der für die Autorisierung genutzt wird. Einen Payload hingegen besitzen die JWT, siehe Kapitel 3.1 und werden bei einer Anfrage im

Authentification Header mitgeschickt. Im Payload des Token können beispielsweise der Benutzername, E-Mail-Adresse und vorher definierte Berechtigungen übergeben werden. Die Systeme verwenden zwar Tokens, aber sind diese nicht direkt überführbar und somit nicht direkt kompatibel. Als Beispiel: ein Nutzer besitzt Zugang zu zwei Systemen, eines per API-Token, eines per JWT. Den API-Token könnte im Payload des JWT verpackt sein, sodass das angefragte System diesen dann auswerten kann. Dahingegen besitzt ein einfacher API-Token keine Möglichkeit, die Informationen aus dem JWT abzubilden. Dafür muss eine Middleware die Token vermitteln und zwischen die Anfragen MDR-spezifisch authentifizieren.

```
{
  "exp": 1623934265,
  "iss": "https://register.itcr.uni-luebeck.de/auth/realms/
     mdr-itcr",
  "aud": "account",
  ...
  "allowed-origins": [
    "https://mdr.itcr.uni-luebeck.de"
  ],
    "mdr-itcr": {
      "roles": [
        "admin",
        "createNamespace",
        "importExport",
        "createCatalog"
      ]
    }
  },
  "scope": "mdr",
  "name": "Hannes Ulrich",
  "preferred_username": "h.ulrich",
  "locale": "de",
  "given_name": "Hannes",
  "family_name": "Ulrich",
  "email": "h.ulrich@uni-luebeck.de"
  }
```

Listing 3.1: Dieses Beispiel zeigt den decodierten Inhalt eines JWT für die Anmeldung am lokalen MDR. Es beinhaltet den Aussteller (*issuer*), die erlaubten Anfragenherkünfte, Nutzerinformationen und die Berechtigungen, die der Nutzer im MDR hat.

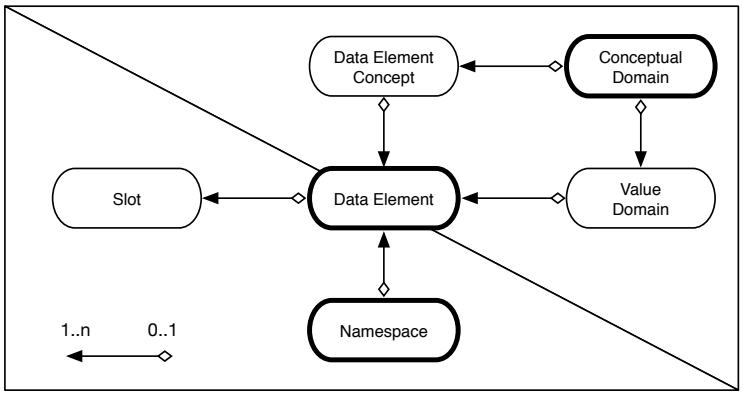
3.4 QL^{4}MDR

Während die Verarbeitung von Datenelementen innerhalb eines MDRs ein gut erforschtes Thema ist [Mate et al., 2019a, Kock-Schoppenhauer et al., 2019a, Deppenwiese et al., 2019], ist der Austausch zwischen mehreren MDRs - Voraussetzung für den Austausch- und Integrationsprozess über verschiedene Standorte hinweg - weitaus weniger untersucht worden. Die Anforderungsanalyse zeigte dabei wichtige Voraussetzungen auf und war ausschlaggebend für die Entwicklung eines uniformen Schnittstelle für MDRs. Die meisten MDR-Systeme sind teilweise konform zum ISO-Standard 11179, so dass Metadaten prinzipiell zwischen MDRs ausgetauscht werden können. Allerdings definiert ISO 11179-3 zwar, wie in 2.1.1 beschrieben, ein Metamodell und grundlegende Attribute zur Beschreibung von Metadaten bietet, aber keine Implementierung. Nach der Untersuchung der zuvor genannten Systeme wurde festgestellt, dass einige Systeme entweder überhaupt keine externen Anfragen zulassen oder die vorhandenen Schnittstellen veraltet waren. Zudem sind bestehende Systeme nicht für Metadatenaustausch konform nach ISO 11179 ausgerichtet und konnten daher nicht verwendet werden. Dies hatte zur Folge, dass selbst bestmöglich kuratierte Metadaten aufgrund der technischen oder syntaktischen Heterogenität nicht verfügbar waren.

Um die Siloisierung der Metadaten aufzubrechen und mehrere MDRs geeignet anzufragen, entstand das Konzept einer einheitliche Schnittstelle für klinische Metadaten. Die Idee einer einheitlichen Schnittstelle (für klinische Systeme) hat ein prominentes Beispiel durch HL7 FHIR [Bender and Sartipi, 2013]. Es werden klinisch relevante, standardisierte Austauschformate unter der Verwendung von modernen Werkzeugen wie JSON und REST zur Verfügung gestellt. Ähnlich zum ISO 11179 gibt es hier eine Trennung des Informationsaustauschs und der darunter liegenden technischen Realisierung. Es wird einzig und allein die Schnittstelle als uniformer Abfrageendpunkt definiert, es wird keine Implementierung beschränkt. Dadurch soll die Integration in bestehende Systeme vereinfacht und gefördert werden. Neben den vielen Vorteilen des REST-Standards gibt es hier aber einen entscheidenden Nachteil: tief gehende, strukturierte Ressourcen sind schwerer zu verarbeiten. Da es sich bei Metadaten überwiegend um tief verschachtelte Informationen handelt, ist man auf technischer Ebene dringend darauf angewiesen, eine effektive Abfrage auf produktive MDR-Systeme zu realisieren. Das ursprünglich von Facebook entwickelte GraphQL wurde als Abfragesprache-Framework gewählt, da

es sich besonders für stark verknüpfte Datenmodelle eignet [Buna, 2016] und bspw. von GitHub, Twitter oder der Deutschen Bahn verwendet wird [GraphQL Foundation, 2021]. FHIR selbst führte GraphQL als Abfragealternative zu den REST-APIs mit dem aktuellen Release R4 ein. Technisch gesehen fungiert GraphQL als eine Abstraktionsschicht für die darunterliegende Datenhaltung und stellt einen einzigen API-Endpunkt, sowohl für Abfragen als auch für Modifikationen, bereit. Die Informationsobjekte werden in einem Schema definiert, welches verbreitete Software-Pattern wie Vererbung, Interfaces, benutzerdefinierte Typen und Attributbeschränkungen wie nicht-nullbare Attribute unterstützt. Ein GraphQL-Schema besitzt drei übergeordnete Objekttypen: *objects*, *query* und *mutation*. Durch die *objects* und dazugehörigen Attributen werden die abfragbaren Informationen dargestellt, welche durch *queries* und beschriebene Filter abfragbar gemacht werden. Die *mutations* dienen der Informationserfassung und -veränderung. Für die Bereitstellung der im Schema definierten Objekte müssen *Data Fetchers* und *Resource Resolvers* implementiert werden, die die abgefragten Ressourcen sammeln und im definierten Format bereitstellen. Neben der Schnittstellenspezifikation unterstützt GraphQL die Introspektion auf Basis des zugrundeliegenden Schemas, so dass die Schnittstelleninformationen maschinenlesbar zur Verfügung stehen. Damit wird die Interaktion mit Clients vereinfacht und kann zur automatischen Generierung von Kommunikationsbibliotheken genutzt werden. Außerdem bietet es Referenzimplementierungen und Softwarebibliotheken in verschiedenen Programmiersprachen, wie JavaScript, Erlang, C# und Java.

Für die Verwendung als Schnittstelle in MDRs wurde das ISO 11179-3 Metamodell als Grundlage für GraphQL Schema genutzt, welches nachfolgend *Query Language for MDR* (QL⁴MDR) genannt wird. Aus den 26 Klassen des Metamodells sind 13 Objekttypen mit sechs Einstiegspunkten im QL⁴MDR-Schema enthalten. Das zentrale ISO 11179-3-Kernmodell wird mit vier Objekttypen dargestellt: *Data Element, Value Domain, Data Element Concept und Conceptual Domain*. Das Schema umfasst auch Namespace und die anpassbaren Slots als Strukturen für die Identifikation von Metadaten. Die ISO 11179-3 spezifiziert diese grundlegenden Objekttypen durch Attribute detaillierter, daher wurden diese Attribute in GraphQL-Felder übersetzt, die zum Filtern und Einschränken der Abfrage verwendet werden können. Um die Filterfunktionalität zu verbessern, werden Objekttypen mit weniger als zwei Attributen in verwandten Objekten als Felder aufgenommen. Zum Beispiel erlaubt die ISO 11179 Property Class die String-Darstellung Property bezogen auf das Data Element Concept. Die GraphQL-Abfragen beginnen an




```
{
    dataElement(slot: "SNOMED-CT",
                value: "723232008") {
        identification {
            urn
            status
        }
        definitions {
            definition
            designation
        }
    }
}
```

Abbildung 3.4: Die sechs definierten Einstiegspunkte, getrennt in die identifizierenden Metadaten (unterer Teil) und die formale Beschreibung der Metadaten (oberer Teil). Die drei fettgedruckten Entitäten sind geeignete Einstiegspunkte für Mutationen. Der rechte Kasten zeigt eine Beispielabfrage zur Abfrage aller Datenelemente, die einen Slot mit dem Namen 'SNOMED-CT' und dem Wert '723232008' (durchschnittlicher Blutdruck) enthalten. Die Abfrage definiert die Darstellung der Antwort: jedes entsprechende Datenelement soll mit seiner Identifikation und seinen Definitionen zurückgegeben werden.

einem Einstiegspunkt (*entrypoint*) und traversieren durch den Datengraphen. QL^{4}MDR bietet sechs Einstiegspunkte: Datenelement als zentrales Informationselement, Value Domain, Data Element Concepts und Conceptual Domain, Namespaces und Slot. Jeder *entrypoint* bietet einen bestimmten Satz von Filtern, um die angefragten Informationen zu spezifizieren, z. B. alle Konzepte bezüglich Person und deren Masse. Da Slots benutzerdefinierte Informationen zu jedem Datenelement enthalten können, ermöglichen sie zusätzliche Parameter für eine reichhaltige Abfrage. Wie in Abbildung 3.4 zu sehen, wurden von den sechs verfügbaren Einstiegspunkten in QL^{4}MDR drei als gültige Ausgangspunkte für die Informationsveränderung ausgewählt: Namespace, Conceptual Domain und das zentrale Datenelement. Diese Auswahl bietet zwei wichtige Garantien: Erstens kann jede Entität erstellt, geändert oder gelöscht werden, da es einen garantierten Pfad gibt. Zweitens ist durch die gerichteten Kanten nicht möglich, zyklische Mutationen zu definieren. Diese würden sonst zu Redundanzen bei der Speicherung führen.

3.4.1 Vorteil und Nachteile von QL⁴MDR

Die vorgestellte Schittstelle QL⁴MDR und ihr technisches Design erfüllt wichtige funktionale Schnittstellenvoraussetzungen, welche in der Anforderungsanalyse für föderierte Strukturen herausgearbeitet wurden, siehe Kapitel 3.1. Das Schnittstellenkonzept fußt auf zwei wichtigen Entscheidungen, die zu Vorteilen in Bezug auf die Interoperabilität führen: die Wahl von GraphQL anstelle von REST- oder einer serviceorientierten Schnittstelle und die Verwendung des ISO 11179-3 Standards anstelle eines proprietären Metadatenmodels.

GraphQL kann als eine Variante des weit verbreiteten RESTful-Design-Patterns betrachtet werden, unterscheidet sich aber in spezifischen Merkmalen und bringt sowohl Vorteile als auch Einschränkungen mit sich: Als GraphQL-basierte API kann QL⁴MDR auch komplexe Fragen beantworten, die durch die verschiedenen Entitäten des ISO 11179-Standards navigieren, und so die Anzahl der erforderlichen Abfragen reduzieren. Mit anderen Worten: Die RESTful- oder serviceorientierten Schnittstellen benötigen wesentlich mehr Anfragen, um die gleichen Informationen zu liefern. Die Anzahl der Abfragen gegen eine RESTful-Schnittstelle hängt von der Anzahl der angefragten Datenelemente und deren Verschachtelung ab. Betrachten wir zum Beispiel eine elektronische Datenerfassung, die Validierungsregeln für alle in einem bestimmten Namespace vorhandenen Datenelemente anfordert, wie in Abb. 3.5 dargestellt. Der anfragende Client erhält neben den gewollten auch eine Vielzahl an redundanten Informationen, da er das Antwortformat des Servers nicht, bzw. nur geringfügig beeinflussen kann. Der Client ist gezwungen, Datenelemente mit allen Eigenschaften abzufragen und muss die verwerfen, die keinen weiteren Nutzen haben [Buna, 2016].

Der Server könnte zwar maßgeschneiderte Anfrage-Routen implementieren, ist aber in der Gegenüberstellung von Nutzen und Aufwand für die Wartung nicht praktikabel. In GraphQL können Clients bei jeder Anfrage das gewünschte Antwortformat definieren, was die Schnittstelle zukunftssicher für neue Client-Anforderungen macht. Dies verlagert die Arbeitslast vom Client zurück auf den Server, um mit einer größeren Anzahl von Client-Implementierungen konform zu sein, bringt aber technische Einschränkungen gegenüber REST mit sich. Einerseits sind die tief verschachtelten Abfragen gut für die stark vernetzten Metadaten geeignet, da sie mehr Informationen enthalten und somit die *Request Roundtrips* reduzieren. Auf der anderen Seite verursachen sie eine höhere Last auf den Datenbankensystemen. Noch nachteiliger ist, dass GraphQL nicht

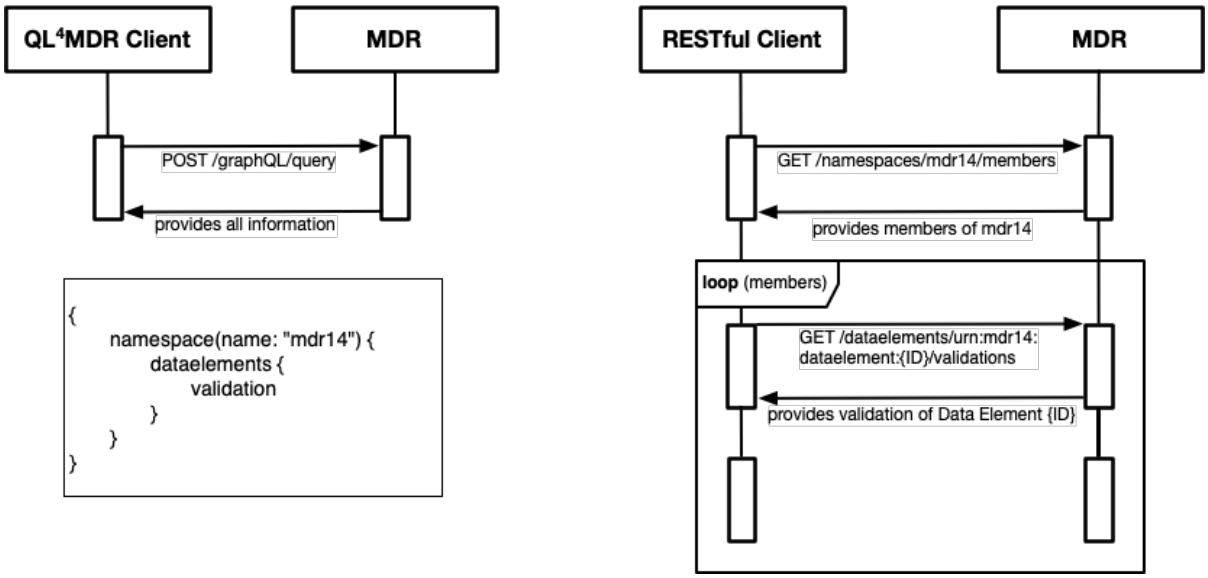


Abbildung 3.5: Dieses Sequenzdiagramm zeigt die erforderlichen Nachrichten zwischen dem GraphQL-Client (links) einschließlich der verwendeten Abfrage (Box), dem RESTful-Client (rechts) und dem MDR-Server, um die Validierungsregeln für jedes Datenelement in einem bestimmten Namespace zu erhalten. Der GraphQL-Client benötigt nur die in der Box gezeigte Abfrage, während die Nachrichtenmenge des RESTful-Clients von der Anzahl der mit dem gewählten Namespace verbundenen Datenelemente abhängt.

auf die Standard-HTTP-Mechanik setzt und daher nicht von den ausgereiften Caching-Mechanismen moderner Webbrowser und Client-Bibliotheken profitiert, was die Datenbanklast bei wiederholten Abfragen weiter verstärkt [Kern et al., 2018]. Aber Facebook war sich dieses Nachteils bewusst und stellt dafür eine JavaScript-Bibliothek fürs Caching zur Verfügung, um dieses Problem zu heben [Buna, 2016]. Ein weiterer Vorteil von GraphQL liegt in der automatischen Erstellung einer implementierungsnahen Dokumentation. Insbesondere können GraphQL-Implementierungen wie *graphql-java* sowohl eine menschen- als auch maschinenlesbare Dokumentation aus dem definierten Schema generieren. Die *Introspektion* ermöglicht nicht nur Anwendern und Entwicklern ein leichteres Verständnis der Schnittstelle, sondern die maschinenlesbare Darstellung ermöglicht eine dynamische und lose Kopplung zwischen Server und Clients und erleichtert so die Föderation verschiedener, technisch unterschiedlicher ISO 11179-basierter MDRs. Bisherige Standards wie der WS-MetadataExchange [Ballinger et al., 2004] können diese

Flexibilität und lose Kopplung aufgrund ihrer schwergewichtigen serviceorientierten Architektur [Kumari and Rath, 2015] nicht bieten.

Der zweite Design-Aspekt war der Fokus auf Metadaten-Standards anstelle ihrer technischen Implementierungen. QL⁴MDR ist dabei nicht auf eine bestimmte Repository-Implementierung zugeschnitten, sondern streng nach dem ISO 11179-3 modelliert. Dieser Ansatz bringt sowohl Vorteile als auch Einschränkungen mit sich. Einerseits gewährleistet das Festhalten an der ISO 11179-3 als gemeinsames Metadatenmodell wiederverwendbare Abfragen, die gegen verschiedene MDR-Implementierungen ausgeführt werden können, solange diese der ISO 11179-3 folgen und QL⁴MDR implementieren. Auf der anderen Seite werden MDR-Systeme für spezielle Anwendungsfälle und Spezifikationen angepasst, die über das hinausgehen, was ISO 11179-3 definiert. Zum Beispiel implementiert Samply.MDR [Kadioglu et al., 2018] die sogenannte *DataElementGroup*, um bestimmte Datenelemente vereinfacht zu gruppieren - diese ist im Standard aber nicht enthalten. Das Design von QL⁴MDR als gemeinsame Schnittstelle ist der erste Schritt auf dem Weg zu einer einfachen Föderation von heterogenen MDRs über eine einheitliche und standardisierte Schnittstelle und damit zur Wiederverwendung von Metadaten. Eine Schnittstelle allein kann jedoch nicht die allgemeinen Probleme des Umgangs mit Metadaten in einem verteilten Kontext lösen, wie z. B. die Konsolidierung von Datensätzen und/oder die Vermittlung zwischen bestehenden Sätzen, das Matching und Mapping von Datenelementen und den Schutz des geistigen Eigentums (Studiendesigns usw.). Außerdem ergeben sich bei der Föderation verschiedener MDR-Instanzen die üblichen Probleme verteilter Informationssysteme wie Replikation, Konsistenz- und Dublettenerkennung, Adressierung und betriebliche Verfügbarkeit und Versionierung.

3.4.2 Erweiterung auf ISO 21526

ISO 21526 ist der designierte Nachfolger des weit verbreiteten ISO 11179, erbt aber auch einige Probleme. Daher war es notwendig, beide Standards strukturell zu vergleichen und die neuen konstruktiven Erweiterungen sowie die inhaltlichen Änderungen hinsichtlich der Auswirkungen auf die beschriebenen Probleme zu untersuchen. Frühere Arbeiten beschrieben [Ngouongo et al., 2013, Park and Kim, 2010] verschiedene Probleme der ISO/IEC 11179: das Fehlen einer semantischen oder syntaktischen Verknüpfung von gemeinsamen Konzepten zwischen Datenelementen und die fehlende Strukturmechanismen entweder für die Metadatenerweiterung oder die Abbildung eines Nutzungsmodells [Sol-

brig, 2000]. Die Umgestaltung des Konzeptpakets in Richtung SKOS und die Einführung des Mapping-Pakets eröffnen die Möglichkeit, das Problem der fehlenden Verknüpfung zwischen Konzepten zu lösen. Das strukturelle Mapping durch das *MDRMapping* und die semantische Annotation mit Konzepten gemäß SKOS ermöglichen direkte Verknüpfungen zwischen jedem verwalteten Element. SKOS ist eine W3C-Spezifikation zur *einfachen* Wissensbeschreibung und wird vornehmlich zur Kodierung von Thesauri und Klassifikationen genutzt. Konzepte können durch simple Relationen wie *related, broader* und *narrow* in Beziehungen gesetzt werden.
Dennoch ist die fehlende Erweiterbarkeit ein strukturelles Problem und hätte im neuen ISO 21526 adressiert werden müssen. Maschinenlesbare oder maschinenverarbeitbare Erweiterungen sind sehr nützlich, wie der häufig verwendete FHIR-Erweiterungs- und Profilierung zeigt und die renommierten FAIR-Prinzipien [Wilkinson et al., 2016] empfehlen. Eine Struktur für ein Nutzungsmodell der Metadaten wurde nicht direkt adressiert, aber die neu hinzugekommenen semantischen Möglichkeiten erlauben Annotationen. Die konzeptuelle Darstellung erfordert, dass jeder Begriff im Vokabular eine einzige, kohärente Bedeutung hat - auch wenn seine Bedeutung je nach seinem Auftreten in einem Kontext variieren kann [Cimino, 1998]. Im Gegensatz zu domänenspezifischen Ontologien wie SNOMED CT ist SKOS in der Lage diese kontextabhängige Beziehung pragmatisch darzustellen und damit ein ausreichendes Nutzungsmodell abzubilden. Für die Erweite-

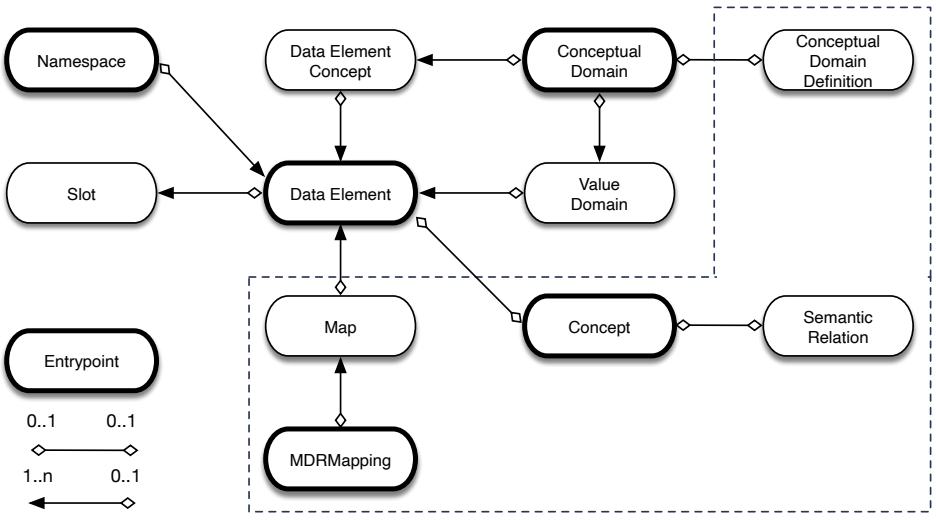


Abbildung 3.6: Die neu hinzugefügten Objekte sind in dem gestrichelten Kasten dargestellt. Neben den bestehenden drei Einstiegspunkten Conceptual Domain, Namespace und Data Element wurden zwei neue hinzugefügt: MDRMapping und Concept.

rung von QL^{4}MDR wurde das entsprechen Schema von der ISO 21526 abgeleitet und die neu eingeführten Mapping-Klassen und Konzeptklassen aufgenommen. Wie in Abbildung 3.6 dargestellt, umfasste das ursprüngliche QL^{4}MDR Schema sechs Objekte und wurde nun um fünf weitere ergänzt: *Concept*, *Semantic Relation*, *Map*, *MDRMapping* und *Conceptual Domain Definition*. Zudem wurden Concept und MDRMapping als neue Einstiegspunkte eingeführt, um eine Abfrage an diesen Objekten zu starten. Durch die Erweiterung des QL^{4}MDR Schemas wird ein Datenaustausch zwischen MDR beider ISO-Standards ermöglicht, da das zugrunde liegende Metamodell nicht verändert wurde. Dadurch können Metadaten standardübergreifend ausgetauscht und verarbeitet werden. Zudem müssen die Information im Sinne einer Altdatenübernahme transformiert werden. Die Implementierung der Mapping-Klasse ist für die föderierte Metadatenverarbeitung zudem von entscheidendem Vorteil. Die Abbildung von standardisierten Mappings zwischen Metadatenelementen ermöglicht Schema-Crosswalks zwischen verschiedenen Elementen in verschiedenen Systemen und fördert deren Wiederverwendung und gemeinsame Nutzung aufgrund ihrer Auffindbarkeit und Verarbeitbarkeit.

3.4.3 Integration in bestehende MDRs

Die Verwendung von QL^{4}MDR setzt eine Integration in bestehende Systeme voraus, welche in Kapitel 3.3 identifiziert wurden. Dafür muss zuerst untersucht werden wie eine Integration von QL^{4}MDR möglich ist. Für die Integrationsstudie [Ulrich et al., 2019a] wurden verschiedenen Kriterien für eine erfolgreiche technische Integration identifiziert:

I. ISO 11179/21526-Konformität,

II. Zugriff auf den Datenkörper,

III. Zugriff auf den Quellcode (Open Source),

IV. vergleichbare Kommunikationsschnittstellen.

QL^{4}MDR und das zugrundeliegende Abfrageschema basieren auf den Standards ISO 1179 und 21526, daher ist eine entsprechende Konformität (I.) der eingebundenen MDR-Systemen vorteilhaft. Eine Integration von nicht-ISO-basierten Systemen ist dennoch möglich, wenn auch sie komplexer ist. Dort muss zuerst geprüft werden, ob das vorliegende Metadaten-Datenmodell auch in das ISO 11179-3 Modell überführt werden

kann. Ein direkter Zugriff auf den Datenkörper (II.) ist dabei vorteilhaft, da dann die Möglichkeit besteht, den Datensatz geeignet auszuleiten und beliebig zu modifizieren. Dadurch können die Daten in eine Standard-konforme Repräsentation überführt werden, sodass eine Integration effektiv realisierbar wäre, obwohl dann auf eine kontinuierliche Datensynchronität geachtet werden muss. Des Weiteren ist ein Zugriff auf den Quellcode (III.) und dessen Erweiterung vorteilhaft und erleichtert die Integration ungemein. Dabei können vorhandenen Strukturen wiederverwendet werden, was die spätere Wartung erleichtert. Zudem ist eine Anknüpfung an die bestehenden Authentifizierungsmechanismen wünschenswert, um die Datenhoheit der MDRs zu gewährleisten. Wenn weder auf Daten noch auf den Quellcode zugegriffen werden kann, sollte das MDR über vergleichbare Kommunikationsparadigmen verfügen (IV.), um die Informationen über eine Schnittstelle bereitzustellen. Das heißt eine *RESTful* Programmierschnittstellen oder gar eine GraphQL-Schnittstelle ist erforderlich. Ist keiner der vier Integrationsfaktoren gegeben ist eine erfolgreiche und nachhaltige Integration nicht gewährleistet. Die verbliebenen sechs Systeme wurden dann auf die zuvor aufstellten Integrationsfaktoren untersucht, die Ergebnisse sind in Tabelle 3.2 dargestellt.

Tabelle 3.2: Die Studie betrachtet die zuvor gefundenen MDR Systeme (siehe Kapitel 3.3) und untersucht sie auf die Integrationsmöglichkeiten. Dabei werden die MDRs auf vier Kriterien untersucht: eine Standardkompatiblität, Zugriff auf den Quellcode, Zugriff auf den Datensatz und ob Kommunikationsschnittstellen im Sinne einer API vorhanden sind.

	ISO 11179/ ISO 21526	**Zugriff Quellcode**	**Zugriff Datenkörper**	**Kommunikations-schnittstellen**
Aristotle	✓/ ✗	(✓)	✗	✓
MDM Portal	✗/ ✗	✓	✓	✓
ART-DECOR	(✓) / ✗	✗	✗	✓
Samply.MDR	(✓/ ✓)	✓	✗	✓
Clinical Knowledge Manager	✗/ ✗	✗	✗	✓
Common Data Element Browser	✓/ ✗	✗	✓	✓
Kairos MDR	✗/ ✗	✗	✗	✓
METeOR	✓/ ✗	✓	✓	✗
Semantic MDR	✓/ ✗	✓	✗	✓

Die Analyse zeigte, dass es vier primäre Möglichkeiten gibt, QL⁴MDR in MDRs zu integrieren: entweder auf Daten- oder Abfrageebene, wie in Abbildung 3.7 zu sehen ist. Während die Datenintegration zumindest teilweise den Zugriff auf Quellcode oder Datenquellen erfordert, findet die Query-Integration auf einer höheren Ebene statt. Da weder auf Daten noch auf Quellcode zugegriffen werden kann, muss die Query-Integration außerhalb des MDR etabliert werden und dafür eine anfragbar Schnittstelle bereitgestellt werden. Bei vier der fünf betrachteten MDR-Systemen war der Quellcode vorhanden, sodass eine Integration auf Datenebene analysiert werden konnte. Die Datenintegration kann mit drei verschiedenen Ansätzen (1-3) unterteilt werden, der letzte Ansatz (4) zielt auf eine Übersetzung der Anfrage.

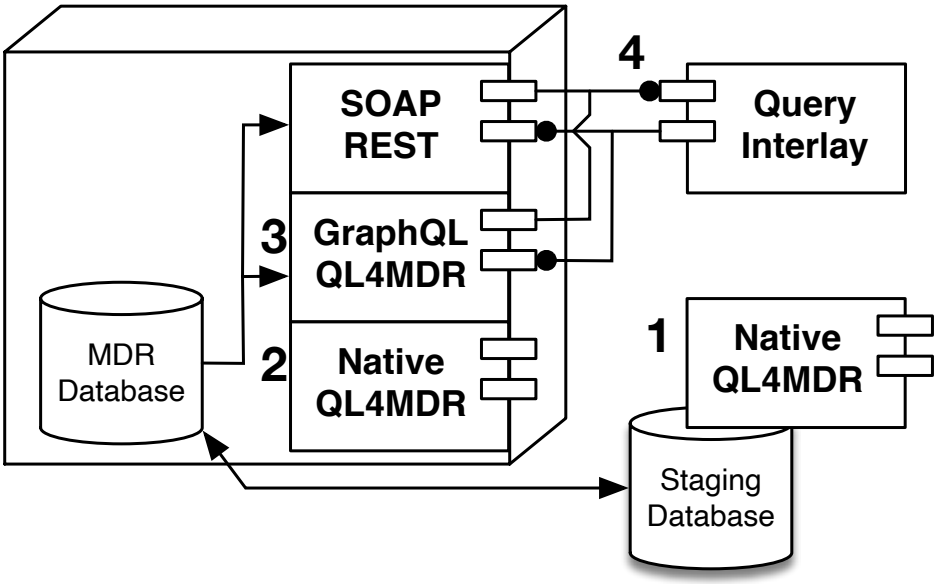


Abbildung 3.7: Die verschiedenen Integrationspunkte, um einen MDR für die Verwendung von QL⁴MDR zu ermöglichen. Die Ziffern entsprechen den genannten Integrationsszenarien.

Transformation auf Datenebene (1)

Wenn die vorhandenen Daten nicht oder nur teilweise ISO-konform sind, kann eine Datentranformation vorgenommenn werden, wodurch der Datenbestand neu formatiert und in eine neue Ausleitungsdatenbank transformiert wird. Ein möglicher Ansatz ist ein eigener Webdienst mit angeschlossener Ausleitungsdatenbank, der die transformierten Daten direkt über eine QL⁴MDR ausleitet, im weiteren wird diese Applikation QL⁴MDR-Pod genannt. Ein anderer, verwandter Ansatz [Ulrich et al., 2018a, Ulrich et al., 2018b] ist die

Verwendung von Neo4j [Webber and Robinson, 2018], einem graphbasierten System mit einem optionalen GraphQL-Plugin. Die Datentransformation sollte in beiden Fällen auf dem QL^{4}MDR-Schema basieren und ist dementsprechend ISO 11179-3-konform. Dies sind gangbare Szenarien für das MDM-Portal. Das Portal stellt klinische Studienformulare in verschiedenen Formaten zur Verfügung, ist aber auf das CDISC ODM-Format fokussiert. In der Literatur findet sich eine Transformation von CDISC ODM nach ISO 11179 [Ngouongo et al., 2013], welche genutzt werden kann, um die Informationen zu konvertieren und über den bereitgestellten GraphQL-Endpunkt verfügbar zu machen.

Quellcodeanpassung (2)
Eine direkte Implementierung in ein ISO 11179/21526-basiertes MDR erfordert die Verwendung von GraphQL-Datenfetchern, um die Daten abzurufen und über eine Schnittstelle bereitzustellen, sowie den Funktionsumfang des HTTP-basierten Abfrageendpunkts sicherzustellen. Hierbei kann die bereits vorhandene Infrastruktur (z. B. Authentifizierung oder Datentransferobjekte) wiederverwendet werden. Diese Kombination aus der GraphQL-Schnittstelle und QL^{4}MDR ist einfach anwendbar für die Verwendung in beliebigen ISO 11179/21526-basierten MDRs, setzt aber den freien Zugriff auf den Quellcode voraus.

Schnittstellenmanipulation (3)
Verfügt das MDR-System bereits über eine bestehende GraphQL-Schnittstelle, kann diese um das Abfrageformat von QL^{4}MDR erweitert werden. Das sogenannte Schema Stitching erlaubt es, verschiedene Schemata zu kombinieren, um alle Informationen mit einer einzigen Abfrage abzurufen [Frisendal, 2018]. Der Resource Resolver muss in das bestehende System integriert werden. Dies bedeutet jedoch, dass der Quellcode der bestehenden GraphQL-Schnittstelle geändert werden muss, was ein denkbares Szenario für die Aristoteles-Metadaten-Registry ist.

Query Interlay (4)
Wenn der Quellcode nicht verfügbar ist oder eine Änderung des Quellcodes der MDR-Implementierung nicht möglich ist (z.B. aufgrund ihrer Lizenz), muss die Integration auf der Abfrageebene erfolgen. Eine zusätzliche externe Komponente wird benötigt, um die eingehenden Anfragen zu transformieren und zu übersetzen. Das *Query Interlay* ist eine Komponente, die mit einem lokal installierten Kommunikationsserver vergleich-

bar ist und als Query Proxy [Braunstein and Detmer, 2016] arbeitet. Es empfängt die Anfrage von einem Client in einem freigegebenen Format, z. B. QL⁴MDR, übersetzt sie in das spezifische Anfrageformat des MDR (z. B. REST oder auch ein anderes GraphQL-Schema) und leitet sie an die Schnittstelle des MDRs weiter. Dementsprechend empfängt das Query-Interlay die Ergebnisse, aggregiert sie bei Bedarf, transformiert sie in das vereinbarte Ergebnisformat und antwortet dann dem Client. Diese Lösung ändert die Schnittstelle des MDRs selbst nicht und kann je nach MDR möglicherweise nicht alle Funktionalitäten bereitstellen, je nachdem, wie gut das MDR konform zum ISO 11179-3-Standard ist. Geeignete Kandidaten sind der Clinical Knowledge Manager und ART-DECOR, welche nicht, bzw. nur partiell ISO-konform sind oder der Common Data Element Browser, da sie alle eine RESTful-Schnittstelle anbieten.

Integrationsszenarien im Vergleich

Die vorgestellten Integrationsszenarien sind allesamt valide Ansätze, um die Integration von QL⁴MDR in bestehende MDRs zu adressieren. Sie unterscheiden sich in den Voraussetzungen und der Komplexität. Es wurden vier Hauptaspekten identifiziert, die eine mögliche Szenariowahl beeinflussen sollten.

Ein **Entwicklungsaufwand** entsteht in jedem Integrationsszenario, ob auf Daten- oder auf Abfrageebene. Die Modifikation eines bestehenden Systems (3) verlangt eine Einarbeitungszeit in ein automon entwickeltes System - es ist aufwändiger das bestehende System zu analysieren und richtig einzubinden. Die Implementierung eines Query-Interlays (4) ist dennoch wesentlich aufwändiger als die Modifikation einer gegebenen Schnittstelle, da die gesamte Kommunikation zum MDR inkl. Client-Anfragen bearbeitet werden muss. Die Transformation des Datenkörpers (1) ist abhängig von zugrundeliegenden System: ist die Datenhaltung bereit ISO-konform, so ist der Entwicklungsaufwand auch geringer. Die Quellcodeanpassung (2) scheint aufgrund des direkten Datenzugriffs den geringsten Entwicklungsaufwand zu erfordern.

Änderungsmanagement und **Datensynchronisation** sind unvermeidlich, da QL⁴MDR Methoden zur Änderung von Informationen bereitstellt und somit eine Synchronisationsstrategie der geänderten Datenelemente mit den MDR-Systemen erforderlich ist. Die Vermeidung von Inkonsistenzen ist zwingend nötig, um die Funktionsfähigkeit des Systems und die Integrität des Datenkörpers zu sichern. Daher gilt es, entweder Datenänderungen über nur eine Schnittstelle zuzulassen oder eine Semaphore zu implemementieren, falls das darunterliegende System nicht über nicht-blockierende Synchronisationsme-

chanismen verfügt. Die Datenintegrationsszenarien (1-3) scheinen unter diesem Aspekt die größte Herausforderung darzustellen. Die Implementierung der Transformation (1) auf Datenebene erfordert eine zweite, graphenbasierte Datenbank, die mit den Quelldaten synchron gehalten wird, sobald eine Transaktion in der primären MDR-Datenbank stattfindet. Eine mögliche Lösung wäre die Synchronisation zu festen Zeitpunkten (vergleichbar mit *nightly builds*), dadurch wären die Metadaten aber nur tagesaktuell. Die Datenmanipulation (2) und die Schnittstellenmanipulation (3) sind unter diesem Aspekt einfacher zu handhaben, da es sich um Änderungen der Metadaten handelt und die bestehende Infrastruktur wie die Datenzugriffsschicht inkl. Datentransferobjekte nutzen kann. Das Query-Interlay (4) wird nur dann beeinflusst, wenn das System Datenänderungen über die API überhaupt unterstützt. In dem Falle ist anzunehmen, dass das System eine Synchronisationsstrategie vorhält, da es eine Datenänderung zulässt.
Wenn eine neue Schnittstelle oder eine neue Komponente in die bestehende MDR-Infrastruktur eingeführt wird, muss sie die bereits vorhandenen **Authentifizierungsprozesse** implementieren. Andernfalls besteht die Gefahr, dass die Sicherheit des MDR und somit die Datenhoheit verletzt wird. Die Datenintegrationen (1-3) erfordert in diesem Punkt wenig Aufwand, da die Authentifizierungsmechanismen des MDR genutzt werden können. Das Query Interlay (4) sollte als neue Komponente am bestehenden Authentifizierungsverfahren teilnehmen. Dies ist nur möglich, wenn das System eine Authentifizierungsschnittstelle nach außen anbietet, bspw. OpenID Connect. Ist diese Möglichkeit nicht gegeben, muss das Query Interlay eine eigene Benutzerverwaltung implementieren, die die Nutzer des MDRs spiegelt - was zu einem höheren Sicherheits- und Arbeitsaufwand führt.
Mit Blick auf die Akzeptanz und Entwicklungsübernahme von QL^{4}MDR müssen sich die Integrationen durch einfache **Wartung** und **Nachhaltigkeit** auszeichnen. Die QL^{4}MDR-Komponente muss leicht an Änderungen angepasst werden können, wenn sich die MDR-Schnittstelle oder das Datenformat ändert. Die Transformation auf Datenebene (1) bringt durch die zweite Datenbank weniger Wartungsaufwand am eigentlichen System mit, doch verbraucht sie durch die doppelte Datenhaltung mehr Resourcen. Die weiteren Integrationen auf Datenebene (2, 3) modifizieren die bestehenden Systeme und erzeugen somit einen höheren Wartungsaufwand, der von den Entwicklern des Systems getragen werden muss. Die Verwendung von vorhandenen Strukturen erleichert die Wartung und die Nachhaltigkeit, da keine zusätzlichen Komponenten gewartet werden müssen. Das Query Interlay (4) erhöht als neue externe Komponente den Wartungsaufwand am Sys-

tem, da jegliche Änderungen an der MDR-Schnittstelle Auswirkungen auf den Funktionsumfang des Query Interlays haben. Aufgrund des Aufwands der neu implementierten Komponente muss es einen Kompromiss zwischen dem Entwicklungsaufwand und der Nachhaltigkeit geben, da letztere durch das Softwaredesign adressiert werden kann. Ein hoher Entwicklungsaufwand kann durch gut konzipierte und implementierte Strukturen den Wartungsaufwand gering halten und damit die Akzeptanz der Komponente fördern. Ein Standardszenario kann nicht definiert werden, da es immer an den lokal vorhandenen IT-Komponenten liegt. Angesichts dieser Kriterien ist es unumgänglich, das zu integrierende MDR sehr genau zu analysieren, um die am besten geeignete Methode der Metadatenföderation zu bestimmen. Im nachfolgenden wird die Integration QL⁴MDR in zwei bestehende Systeme beschrieben. Das Samply.MDR und das Pragmatic MDR wurden ausgewählt, da sie die breiteste Verwendung und größten Datenbestand boten. Bei jeweils beiden Systemen war entweder Zugriff auf den Quellcode oder Zugriff auf den Datensatz vorhanden.

3.4.3.1 Samply.MDR

Als erstes MDR wurde das Samply.MDR [Kadioglu et al., 2016] für eine Integation untersucht und für eine Proof-of-Concept-Implementierung von QL⁴MDR ausgewählt. Das Samply.MDR wurde 2012 im Rahmen des deutsches Konsortiums für Translationale Krebsforschung (DKTK) entwickelt und der medizininformatischen Community als Open Source-Projekt zur Verfügung gestellt. Mittlerweile wird das System in vielen verschiedenen Projekten auf nationaler, aber auch europäischer Ebene genutzt [Kadioglu et al., 2018]. Gerade die Verwendung im MIRACUM-Konsortium im Rahmen der Medizinformatik-Initiative [Semler et al., 2018] macht das Samply.MDR durch die Nähe zum Datenintegrationszentrum zu einen wertvollen Kandidaten im Sinne der föderierten Metadaten. Zudem existiert eine Adapterentwicklung für die Abfrage durch den Common Data Element Browser, welches die Wiederverwendbarkeit von Metadaten steigern kann. Das System ist partiell kompatibel zum ISO 11179 Standard, es implementiert zwar die Repräsentationsebene, aber nicht die semantisch wertvolle konzeptionelle Ebene. Zudem wurden im Zuge der Benutzerfreundlichkeit und Nutzerakzeptanz Datenmodellerweiterungen vorgenommen, beispielsweise die nicht standardkonforme *DataElementGroup*.

Das System ist in Java entwickelt und folgt der Servlet-Spezifikation der Java Platform Enterprise Edition. Durch die Servlet-Struktur ist es modular aufgebaut und bietet neben der Hauptapplikation ein RESTful Interface als zusätzliches Servlet an. Mit dem Open Source-Programmcode bietet sich eine Integration im Sinne der Quellcodeanpassung an, doch wurde es als eigenständiges Modul entwickelt und daher am eigentlichen Quellcode keine Änderungen vorgenommen werden mussten. Bei der Implementierung konnte auf vorhandene Komponenten zurückgegriffen werden, wie beispielsweise Programmbibliotheken für die Datenzugriffsschicht. Zudem konnte das QL^{4}MDR-Modul angelehnt an das RESTful Interface in das bestehende Authentifikations- und Autorisierungssystem eingebunden werden. Eine Schwierigkeit während der Implementierung waren die Datenmodellerweiterungen, gerade die DataElementGroup, da sie sehr stark in die standard-konforme Datenrepräsentation eingreift. Da diese Entität nicht im Standard enthalten ist, sollte sie nicht über QL^{4}MDR abgefragt werden. Doch das Samply-Schema konnte adaptiert werden, sodass die zusätzliche DataElementGroups ebenfalls abgebildet wurde. Allerdings wurde eine standardkonforme Adaption vorgenommen. Die DataElementGroups konnten als komplexe Datenelemente behandelt werden, die aus mehreren Datenelementen, einer Bezeichnung und einer Definition bestehen. Somit sind sie als Datenelemente abfragbar inklusive der darunter gruppierten Datenelemente.

QL^{4}MDR konnte erfolgreich als Proof-of-Concept in das bestehende System implementiert werden und wurde auch in das Stammrepository[1] der Samply Entwickler-Community aufgenommen.

3.4.3.2 MDM Portal Pragmatic MDR

Das MDM Portal der Universität Münster, wie in Kapitel 2.3 vorgestellt, bietet die größte, öffentlich zugängliche Sammlung klinischer Formulare weltweit [Dugas et al., 2016]. Mit ca. 470.000 Datenelementen ist die Integration des MDM-Datenbestandes ein gewinnbringender Schritt für die föderierte Abfrage von Metadaten. Die Daten des MDM werden im sogenannten *Pragmatic MDR* bereitgestellt [Hegselmann et al., 2021], welches hier das Integrationsziel ist. Das Pragmatic MDR synchronisiert sich mit dem datenhaltenden MDM und ist in Java implementiert, zudem als Open Source verfügbar und basiert auf Spring Boot. Für eine schnellere Suche in dem großen Datenbestand wird Apache Solr mit angepassten Suchindexen genutzt. Zudem stellt es eine öffentlich

[1] https://bitbucket.org/medicalinformatics/samply.mdr.ql4mdr

zugängliche ressourcenorientierte REST-API zur Abfrage und Wiederverwendung der Datenelemente zur Verfügung.

Für das MDM, bzw. das Pragmatic MDR sind mehrere Integrationsszenarien denkbar: Datentransformation und Quellcodeanpassung. Durch die Verfügbarkeit des Datensatz (siehe Kapitel 2.3) ist eine Datentransformation möglich. Hierbei müssen die Datenelemente aus dem CDISC ODM Format in ein ISO-kompatibles Schema transformiert werden; dazu sind in der Literatur validierte Mappings und Transformationspipelines zu finden [Ngouongo et al., 2013, Kock-Schoppenhauer et al., 2018c]. Durch die Verfügbarkeit des Quellcodes ist eine Codeanpassung ebenso möglich. Hierbei kann die Spring Boot-Applikation um eine weitere Komponente erweitert werden, welche sich dann in die bestehende Applikationsumgebung integriert. Ähnlich zur reinen Datensatztransformation muss auch bei der Integration einer QL⁴MDR-Schnittstelle der Datenbestand intern konvertiert werden, um als ISO-kompatibles Format ausgeliefert zu werden. Für die technische Integration kann auf eine gut gepflegte, speziell für Spring Boot implementierte, GraphQL-Library zurückgegriffen werden.

Die GraphQL-Library war auch ausschlaggebend für die Entscheidung, welcher Integrationsweg am besten geeignet ist. Die GraphQL-Library erwartet die Verwendung von Spring Boot in der Version 2.1.X oder neuer. Der Codestamm vom Pragmatic MDR hingegen war nur in Version 1.5.21 verfügbar und somit nicht kompatibel. Die Aktualisierung auf eine neuere Version war laut den Entwicklern angedacht, doch wurde kein zeitlicher Rahmen definiert. Somit war die Integration per Datentransformation das geeignete Szenario. Dafür waren zwei Schritte notwendig: die Transformation in einer Ausleitungsdatenbank und die Bereitstellung per QL⁴MDR. Das Mapping war jedoch durch die Vorarbeiten gegeben, sodass die Daten (siehe Kapitel 2.3) per Skript valide transformiert werden konnten. Bei der Ausleitungsdatenbank konnte zwischen einer Neo4j mit integrierter GraphlQL-Schnittstelle und dem QL⁴MDR-Pod gewählt werden. Nach eingängiger Untersuchung musste der Weg über Neo4j leider ausgeschlossen werden, da in der aktuellen Version 4.0 die native GraphQL-Unterstützung entfernt wurde und somit nur die Verwendung des Pods möglich war.

Der QL⁴MDR-Pod fungiert als simples MDR, welches QL⁴MDR als Kommunikationsschnittstelle vollständig implementiert. Der Pod kann neben dem produktiven MDR Daten für die Föderation separat bereitstellen. Damit kann ein MDR-Betreiber Daten, welche für eine weitere Nutzung im föderierten Kontext in Betracht kommen, versioniert in den Pod überspielen und somit *sensible* bzw. urheberrechtlich geschützte Me-

tadaten im produktiven MDR schützen. Diese Datentrennung ist vergleichbar mit dem Brückenkopfkonzept von Lablans et al. [Lablans et al., 2015], doch muss der Pod nicht gesondert geschützt werden und darf per Netzwerkverbindung erreichbar sein. Das bei Lablans et al. beschriebene Polling Konzept für Instanzdaten muss im Kontext von föderierten Metadaten keine Anwendung finden. Der Pod ist mit Spring Boot in Java entwickelt und nutzt die zuvor erwähnte GraphQL-Bibliothek. Für die Datenhaltung wird für Entwicklungszwecke eine H2 Database genutzt und für den produktiven Einsatz eine PostgreSQL-Datenbank. Die Daten sind als CSV-Datei aus dem MDM gegeben und werden per Python-Skript geladen, transformiert und über die Schnittstelle in den Pod geladen. Das System ist als Ausleitungsdienst konzipiert, sodass neue Daten einfach in Pods gespielt werden. Änderungen an den bestehenden Daten werden überschrieben. Im Sinne des Brückenkopf-Konzepts werden hier keine Daten in das Produktivsystem zurückgespielt.

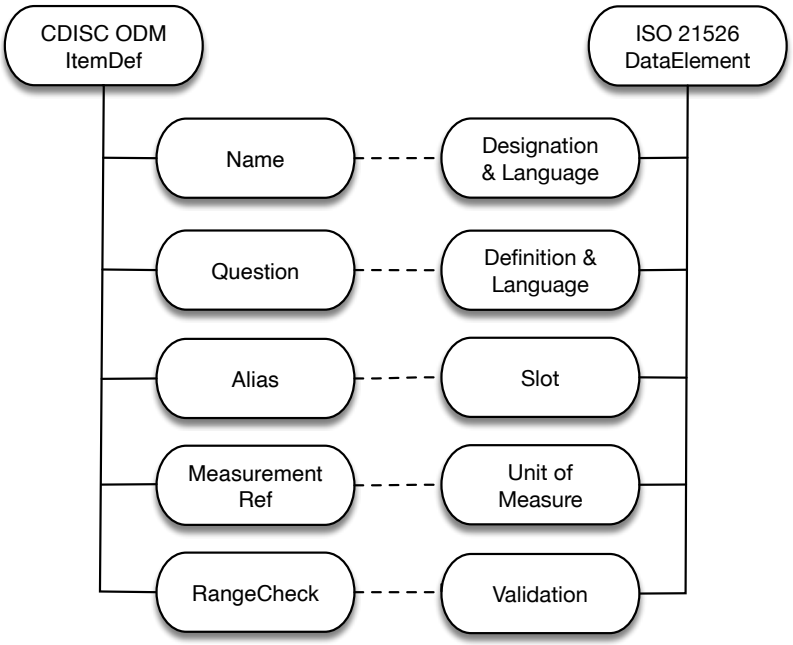


Abbildung 3.8: Die Grafik zeigt das Mapping einer CDISC ODM ItemDef in Version 1.3.1 in ein ISO 21526 DataElement. Alle notwendigen Attribute aus CDISC ODM können überführt werden. Das vollständige Mapping ist zu finden bei Kock-Schoppenhauer et al. [Kock-Schoppenhauer et al., 2018c].

Kapitel 4

Mehrwertdienste auf Basis der Metadaten

Im vierten Kapitel werden basierend auf der zuvor beschriebenen Metadatenschnittstelle Mehrwertdienste eingeführt. Die Dienste fügen sich zu einer umfassenden Pipeline zusammen, welche eine Metadaten-getriebene Datenintegration im klinischen Umfeld erreichen soll. Die Dienste adressieren dabei die einzelnen Schritte der Integration: Matching ([Deppenwiese et al., 2019]), Mapping ([Kern et al., 2018, Ulrich et al., 2020b]) und die Transformation. Sie sind ein Beispiel für mögliche Anwendungen basierend auf den erschlossenen Metadaten und dazugehörigen Daten.

Die Erschließung der projektspezifischen Metadaten war kein Selbstzweck, sondern soll die Verarbeitung der beschriebenen Daten ermöglichen und vereinfachen. Die zuvor eingeführte Schnittstelle soll die Kommunikation von Metadaten standardisieren und ein Metadaten-fokussiertes Ökosystem schaffen, welche die Datenintegration fördert. Dazu sollen atomare Mehrwertdienste konzipiert und implementiert werden, die es den Data Stewards ermöglichen, die projektspezifischen Metadaten und dazugehörigen Daten effektiv in klinische Datenintegrationszentren zu integrieren. Die Dienste sollen dabei möglichst eigenständig mit wenig Abhängigkeiten zu anderen externen Diensten funktionieren. Diese lose Kopplung soll durch den Einsatz von QL^{4}MDR als standardisierte Schnittstelle gewährleistet werden. Dadurch soll ein schneller und unkomplizierter Einsatz an verschiedenen Standorten ermöglicht werden. Die Dienste sollen als Container bereitgestellt werden, um eine horizontale Skalierung in den Datenintegrationszentren zu ermöglichen.

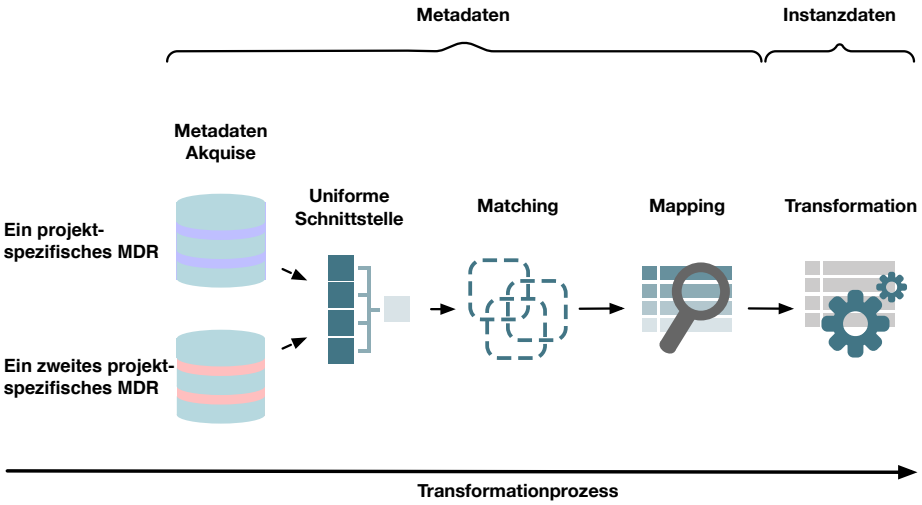


Abbildung 4.1: Durch die Anbindung heterogener Quellssysteme über eine uniforme Schnittstelle wird die einheitliche Durchsuchbarkeit von verteilten Metadaten ermöglicht. Über die Schnittstelle können die Daten angefragt und dann weiterverarbeitet werden. Dieses Grafik zeigt die einzelnen Verarbeitungsschritte der Metadaten für die Datenintegration.

4.1 Metadaten Matching

Wie in der Literaturübersichtsarbeit in Kapitel 2.2.2 gezeigt wurde, ist das Matching ein wichtiger Schritt für die weitere Verarbeitung von Metadaten für die Datenintegration. Das Matching ist nötig, um heterogene Metadaten zu integrieren und damit die klinischen Instanzdaten für die Sekundärnutzung verfügbar zu machen. Eine schnelle und flexible Integration erfordert aber auch ein flexibles Tooling, welches alle wichtigen Werkzeuge für das Matching vereint: MDRCupid [Deppenwiese et al., 2019]. MDRCupid unterstützt diesen Schritt, indem es eine konfigurierbare Matching-Toolbox bereitstellt, die lexikalische und statistische Matching-Ansätze beinhaltet. Die Matching-Konfiguration kann durch die manuelle Auswahl von Algorithmen und deren Gewichtungen oder durch die Verwendung des Optimierungsmoduls mit entsprechenden Trainingsdaten an unterschiedliche Zwecke angepasst werden.

Es gibt mehrere Tools, die eine automatische oder halbautomatische Abgleich- oder Zuordnungsfunktionalität bieten. CUPID ist ein etablierter Schema-Matching-Algorithmus [Madhavan et al., 2001], der das Label, den Datentyp, die zulässigen Werte und die Gesamtstruktur berücksichtigt, aber nicht alle in ISO 11179-Datenelementen enthaltenen

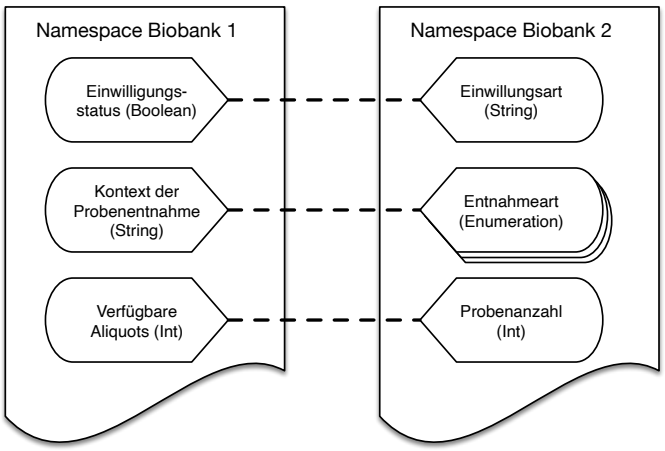


Abbildung 4.2: Die Grafik zeigt drei Matchings aus einem Biobankverbundsprojekt, indem Namespaces aufeinander abgebildet werden sollten. Dabei zeigen sich hier beispielhaft Probleme, welche während des Matchings auftreten: unterschiedliche Formate, Strukturen oder fachsprachliche Unterschiede.

Informationen unterstützt. Ein weiteres Beispiel ist die COMA++ Software [Aumueller et al., 2005], die die Verwendung und Parametrisierung mehrerer String-Matching-Algorithmen ermöglicht und SQL-, XML- und OWL-Eingabedateien akzeptiert. Wie anfangs eingeführt, sind die zu prozessierenden Studiendaten meist in EAV-Modellen gespeichert und es bedarf einer Strukturierung über Metadaten. Diese sind meist in ISO 11179-kompatiblen MDR-Systemen gespeichert und daher ist eine spezifische Unterstützung für ISO 11179-Datenelemente notwendig, welches die beiden Algorithmen nicht bieten. Andere Ansätze wie das MDM-Portal setzen darauf, Metadatenelemente mit semantischen Codes zu annotieren und diese Informationen zur Erkennung von Korrespondenzen zu verwenden. Das Ziel und das Alleinstellungsmerkmal von MDRCupid ist es, ein hochgradig konfigurierbares Matching-Tool für konforme Metadatenelemente nach ISO 11179 bereitzustellen. Das System besteht aus vier verschiedenen Komponenten: Toolbox, Datenbankdienst, Matching Service und Optimization Service.

Toolbox

Die Toolbox implementiert alle wichtigen Funktionen für das Matching und die Konfigurationsmöglichkeiten. Die Toolbox ist so konzipiert, dass sie von einem anderen Dienst verwendet werden kann und für den jeweiligen Usecase alle benötigten Algorithmen und Werkzeuge implementiert und bereitstellt. Um die besten Matches für ein gegebenes

Datenelement aus einer Menge von Kandidatenelementen zu finden, wird ein Ähnlichkeitswert für jede mögliche Übereinstimmung berechnet. Dieser Score wird als Summe gewichteter Ähnlichkeits-Scores für einzelne Datenelement-Attribute berechnet. Diese Attribute können als Metadaten der Metadaten betrachtet werden und umfassen z. B. Name und Datentyp. Die Menge der möglichen Attribute ist durch die Konformität zu den ISO-Standards begrenzt. Die Attributswerte (z. B. Geburtstag oder Geburtsdatum für Name und String oder *Date* für Datentyp) werden als einfache Inputstrings betrachtet, auf die Verarbeitungspipelines und ein lexikalischer Matching-Algorithmus angewendet werden können.

Für die Vorverarbeitung stehen Werkzeuge aus dem Apache OpenNLP-Projekt zur Verfügung: Tokenizer, POS-Tagger und Lemmatizer. Die vorverarbeiteten Datenelemente werden dann mit Hilfe eines Matching-Algorithmus verglichen und deren Ähnlichkeit über einen Score ausgedrückt. Die Toolbox bietet eine Vielzahl an gängigen lexikalischen Matching-Algorithmen:

- Kosinus-Ähnlichkeit
- Jaccard-Ähnlichkeit
- Ähnlichkeit anhand der Longest Common Subsequence
- Ähnlichkeit anhand der NGramme

Andere String-Matching-Algorithmen können leicht hinzugefügt werden, sofern sie zwei Bedingungen erfüllen. Der Algorithmus muss normalisiert sein, das heißt die Ergebnisse müssen unabhängig von der Länge der verglichenen Zeichenketten im Bereich zwischen 0 und 1 sein. Zudem sollen höhere Punktzahlen bessere Übereinstimmungen anzeigen. Für Algorithmen, die Abstände berechnen, kann dies leicht durch den Kehrwert des berechneten Abstands erreicht werden. Die Ähnlichkeitsbewertung für zwei Stringwerte aus zwei Datenelementen für dasselbe Attribut wird dann als Summe der gewichteten Ergebnisse für alle angewandten Verarbeitungspipelines errechnet. Die Gewichte für die Verarbeitungspipelines können konfiguriert werden, ebenso wie die zuvor erwähnten Attributgewichte. Neben dem Vergleich von einfachen String-Attributen (z. B. Namenswerte) werden auch Scores für andere Datenelemente Eigenschaften berechnet, die zur Gesamtbewertung des Datenelements beitragen. Da ein Datenelement eine

Liste von zulässigen Werten (Antwortoptionen) enthalten kann, wird ein Ähnlichkeitsmaß für zwei solche Listen benötigt, um sie in die Gesamtbewertung einzubeziehen. Um dieses Maß zu berechnen, wird die Ähnlichkeit zwischen jeder einzelnen Option in Datenelement A und jeder Option in Datenelement B mit einer konfigurierbaren Kombination der obigen lexikalischen Matching-Algorithmen berechnet. Diese Werte werden in einer Ähnlichkeitsmatrix gesammelt. Unter Verwendung des Kuhn-Munkres-Algorithmus [Munkres, 1957] wird dann höchstens eine Zahl aus jeder Spalte und Zeile so gewählt, dass die Summe dieser Werte maximiert wird. Jeder gewählte Wert repräsentiert eine Übereinstimmung zwischen zwei Antwortoptionen, und die maximierte Kombination stellt die insgesamt beste Übereinstimmung dar. Zudem werden relevante Strukturinformationen in den Matching-Prozess miteinbezogen. Bei Formularen aus klinischen Studien werden Fragen und somit die korrespondierenden Metadaten gruppiert und dadurch in einen Kontext gesetzt, z. B. Einschluss- oder Ausschlusskriterien. Dieser Kontext kann entscheidend für das Matching sein und so kann der Ähnlichkeitswert zwischen den korrespondierenden Gruppen den Gesamtscore beeinflussen. Die Relevanz von den Struktur- und Kontextinformationen konnte in einer vorherigen Arbeit erfolgreich gezeigt werden [Ulrich et al., 2018b]. Dort konnte durch die Modellierung der hochvernetzten Metadaten in Graphen durch die Verwendung der Strukturinformationen der Matchingaufwand reduziert und damit die Ergebnisse verbessert werden.

Datenbankdienst

Der Datenbankdienst stellt die zentrale Datenhaltung für MDRCupid dar. Dort werden alle relevanten Informationen für die beiden datenerzeugenden Dienste vorgehalten und persistiert. Der Dienst verwendet die NoSQL-Datenbank MongoDB und speichert fünf verschiedene Entitäten: Config, CupidMaps, Truth, FHIR ConceptMaps und FHIR Provenance. Die Entitäten werden per REST-API den anderen Diensten zur Verfügung gestellt. Ein *Config*-Objekt stellt eine vollständige Konfiguration für den Matcher dar. Die einzelnen Konfigurationen werden gespeichert, um das Matching später nachvollziehbar zu machen. Die *CupidMaps* sind die Ergebnisse des Matchings. Die Maps enthalten zu einem Quelldatenelement eine Liste mit möglichen Zieldatenelementen und dem entsprechenden Score. Zudem werden noch administrative Daten wie ein Zeitstempel, ein Querverweis zur verwendeten Konfiguration, sowie der Nutzer, der das Matching veranlasst hat. Die *Truths* sind validierte Matchings, welchen vom Optimierer verwendet

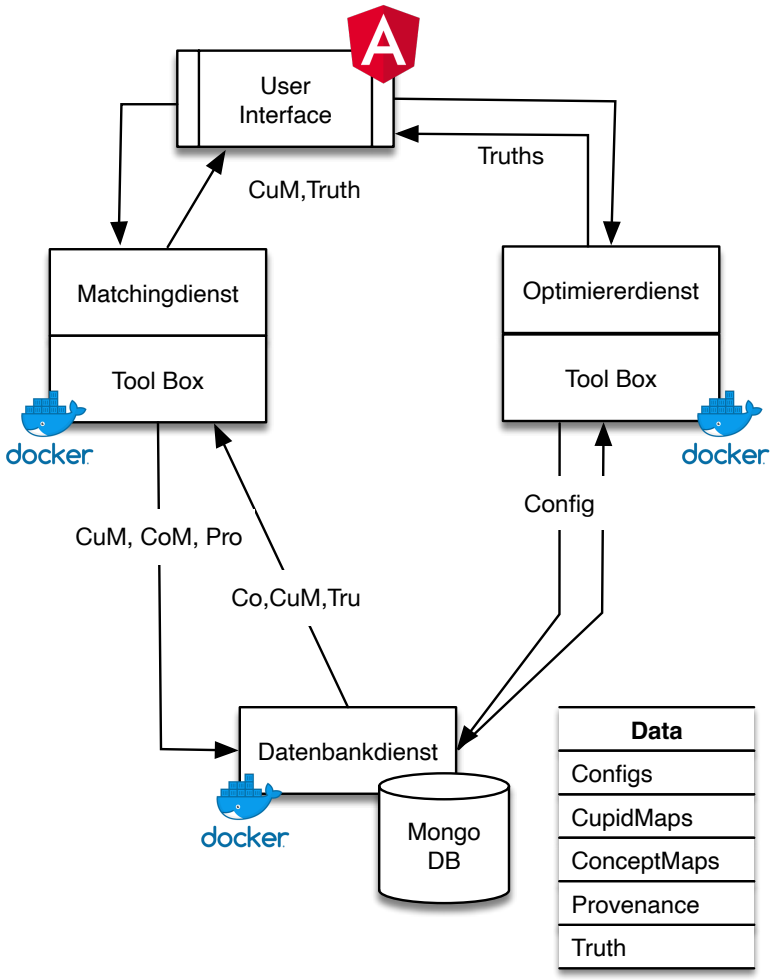


Abbildung 4.3: Die Grafik zeigt die einzelnen Komponenten von MDRCupid und deren Zusammenspiel. Das System besteht aus vier Einzeldiensten: UserUser Interface, Matchingdienst, Datenbankdienst und Optimiererdienst.

werden, um eine optimale Konfiguration zu berechnen. Die Truths beinhalten die Quell- und Zielschemata, die validierten Korrespondenzen und den erstellenden Nutzer. Die Datenbank speichert zudem auch die exportierten FHIR R4 *ConceptMaps* und die dazugehörigen *Provenance*-Informationen.

Matchingdienst

Der Matchingdienst verwendet die zuvor beschriebene Toolbox, um ein Matching zwischen Schemata zu erzeugen. Der Dienst stellt neben den Toolbox-Features auch die Daten für das User Interface bereit. Über eine modulare Interface-Architektur können veschiedene Ein- und Ausgabeformate unterstützt werden. Dafür müssen die Funktionen des jeweiligen Interfaces implementiert werden. MDRCupid unterstützt schon mehrere Eingabequellen: die zuvor beschriebene uniforme Schnittstelle QL^{4}MDR und einen Konnektor für das Samply.MDR. Für die Ausgabe sind zwei Konverter implementiert, Neo4j Cyphers und FHIR ConceptMaps. Der Ablauf, wie in Abbildung 4.4 beschrieben, startet durch einen MatchingRequest, der entweder vom User über das User Interface oder per API-Aufruf gestellt werden kann. Für einen MatchingRequest sind drei Parameter nötig: Quell- und Zielschemata und ggf. eine Konfiguration - referenziert über die korrespondierende URI. Der Dienst verbindet sich dann über die angegebenen Konnektoren mit den MDRs und lädt die Schemata. Wenn eine Konfiguration angehängt wurde, wird sie vom Datenbankdienst abgefragt; falls keine Config angegeben wurde, wird eine Standardkonfiguration geladen. Da der Matching-Prozess eine gewisse Zeit in Anspruch nimmt, wird dem anfragenden Client eine UUID als Token zurückgegeben. Über die UUID und über die zusätzliche Angabe des Ausgabeformats kann der Client dann eine ExportRequest stellen. Dadurch kann der Dienst um weitere Ausgabeformate erweitert werden ohne dass die APIs angepasst werden müssen. Zudem kann ein Ergebnis später in verschiedenen Formaten ausgeliefert werden.

Optimiererdienst

Da die Toolbox die Konfiguration von vielen Parametern erlaubt, enthält sie auch ein Optimierungsmodul, um den Prozess der Suche nach der besten Konfiguration für ein bestimmtes Matching-Problem zu ermöglichen. Die gefundene Konfiguration kann dann für ein anderes Matching-Problem genutzt werden. Durch die Verwendung von vorher validierten Matchings und dem darin enthaltenen impliziten Wissen sollen die Matchings fundierter und robuster werden. Da die optimale Konfiguration der Attributgewichte von den Algorithmen abhängt, die zur Berechnung der Attributähnlichkeiten verwendet werden, werden die Parameter in zwei Schritten berechnet. Zuerst werden die besten Verarbeitungspipes und deren Gewichte für jedes Attribut separat bestimmt. Im zweiten Schritt werden die Attributgewichte mit Hilfe der Algorithmuskonfiguration aus Schritt

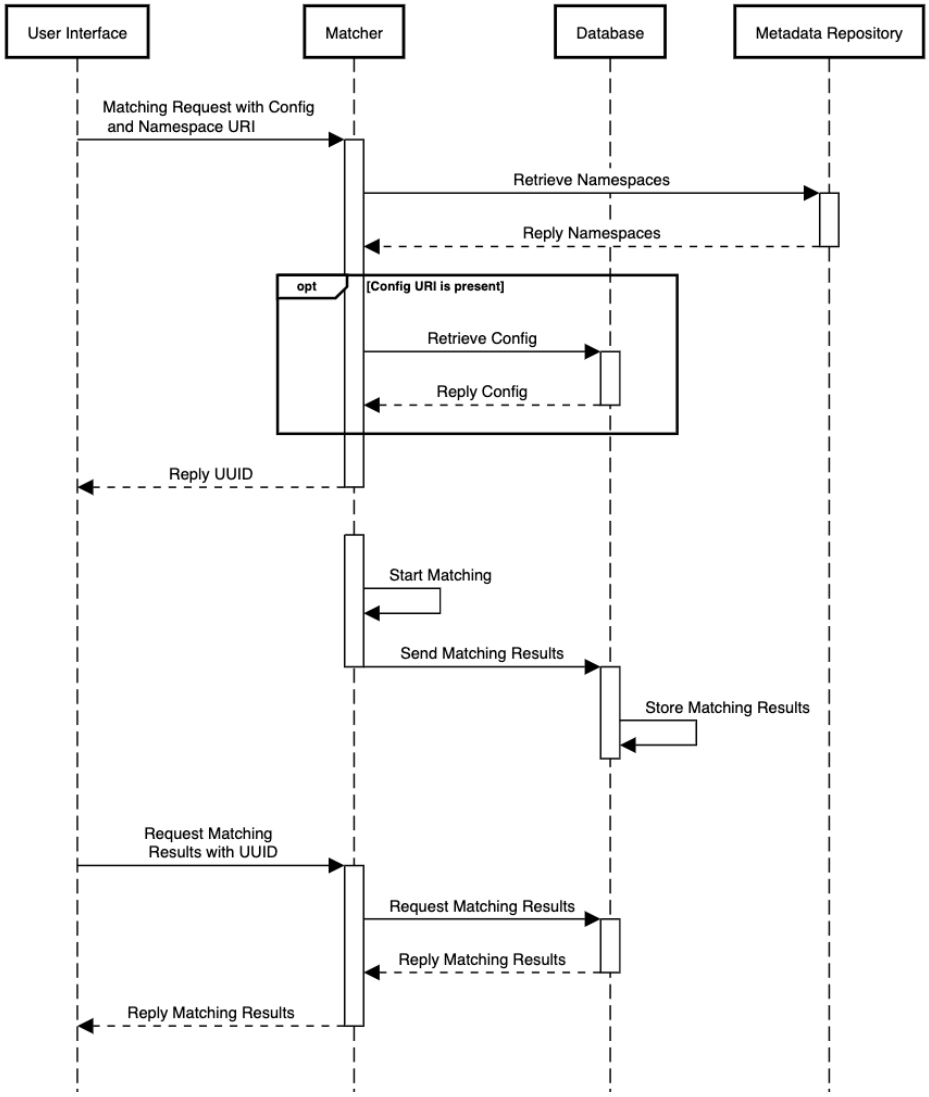


Abbildung 4.4: Das Sequenzdiagramm zeigt den Ablauf eines Matchingsprozesses. Ein User löst den Prozess über einen *Request* aus. Der Matcher quittiert die Anfrage mit einer UUID und fragt die benötigten Daten aus einem MDR ab. Ist der Vorgang abgeschlossen, kann der User mit der UUID die Ergebnisse abrufen.

eins berechnet. Dazu verwendet der Dienst eine Truth, also ein validiertes Matching zweier Schemata. Für die Algorithmusoptimierung werden alle möglichen Kombinationen von aktuell implementierten Präprozessoren und Matching-Algorithmen als Verarbeitungspipelines verwendet. Um die besten Kombinationen für ein bestimmtes Attribut zu finden, werden für jede Kombination von zwei Namespaces die Datenelement-Ähnlichkeitsscores für jedes Attribut für jede Pipe separat berechnet. Dann werden die besten drei Pipes

pro Datenelement ausgewählt und absteigend mit drei, zwei bzw. einem Punkt bewertet. Diese Bewertungspunkte werden für jede Pipe über alle Trainingsdaten summiert, normalisiert und dann als Gewichte dieser Pipes verwendet. Um zu entscheiden, welche Pipes die besten Ergebnisse liefern, stehen zwei verschiedenen Maße zur Verfügung. Für beide Bewertungen wird jeweils eine Liste von möglichen Matches benötigt, welche geordnet in absteigender Reihenfolge nach ihrer Ähnlichkeitsbewertung zum Zieldatenelement sortiert ist. Um zu berechnen, wie die Attribute gewichtet werden sollen, werden mit den im ersten Optimierungsschritt ermittelten Pipes und Gewichten die Scores und Rankings für zwei Schemata für jedes Attribut separat berechnet. Anschließend werden die Attributgewichte analog zu den Verarbeitungspipes während der Algorithmusoptimierung berechnet.

1. Für den Abstandsscore (4.1) wird der Score $s(\cdot)$ des ersten validen Elements m_j vom Score des höchstbewertete Elements m_1 in der Liste abgezogen.

$$\frac{1}{N}\sum_{i=1}^{N}\Big(s(m_j = n_i) - s(m_1)\Big) \tag{4.1}$$

2. Der Positionsscore (4.2) wird berechnet als der Kehrwert der Position $pos(\cdot)$ des ersten übereinstimmenden Elements m_j. Da es keine Metrik ist, bedeuten größere Werte hier also eine bessere Platzierung.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1}{pos(m_j = n_i)} \tag{4.2}$$

Evaluation

MDRCupid nahm an einer wissenschaftlichen Challenge, dem GMDS Mappathon 2018 teil. Der Mappathon ist eine Metadaten-Mapping-Challenge, die nach Methoden sucht, um korrespondierende Datenelemente innerhalb ähnlicher Datensätze zu finden und um Datenelemente untereinander zu korrelieren. Dafür wurden verschiedene annotierte Datensätze ausgegeben, welche als Ground Truth fungierten. Um die Anforderungen des Wettbewerbs zu erfüllen, wurde die Toolbox so modifiziert, dass sie die externe CSV-Dateien als Ground Truth unterstützt und darauf die Matching Pipelines optimieren konnte. Auf der Konferenz wurden dann neue Datensätze verteilt, die verarbeitet und per Cypher an einen bereitgestellten Validator gesendet wurden muss-

te. Der Mappathon-Validator berechnete den Null-Eins-Klassifikationsverlust und den Mappathon-Score gemäß einer definierten Bewertungsmatrix. MDRCupid gewann die Mappathon-Challenge in der Kategorie *Bestes Mapping* und konnte sich gegen sechs registrierte Teams durchsetzen. Die Ergebnisse sind in der entsprechenden Veröffentlichung zu finden [Kock-Schoppenhauer et al., 2018a].

4.2 Mapping Editor

Um die Ergebnisse aus dem Matching-Prozess für die Datenintegration nutzbar zu machen, müssen die Matches in Mappings umgewandelt werden. Dazu müssen die entstandenen Maps durch Domänenexperten gesichtet und validiert werden, zudem müssen Transformationsregeln erstellt werden. Diese Regeln werden später die Instanzdaten basierend auf dem Quellschemata angewendet, um sie ins Zielschema zu überführen. Ohne Computerunterstützung ist die Erstellung von Transformationsregeln zeitaufwändig und fehleranfällig. Trotz der Vielzahl von Open-Source-Tools zur Unterstützung von ETL-Prozessen [Majchrzak et al., 2011], haben sie alle den Mangel an Austauschbarkeit gemein: Die Transformationen werden lokal in einer proprietären Darstellung kodiert. Anbieter von Gesundheitsdaten und Integrationszentren, die mit dem gleichen Integrationsproblem konfrontiert sind, könnten vom Austausch dieser Transformationsschemata profitieren [Atzeni et al., 2019]. Es wird also ein unabhängiger und standardisierter Weg für den Austausch von Transformationsregeln benötigt. Diese sollten in einem standardisierten Format, welches Verwendung im Gesundheitswesen findet, bereitgestellt werden. Um die Reproduzierbarkeit zu fördern, müssen weiterhin Provenance-Informationen zur Verfügung stehen, um die Zuverlässigkeit der Transformationsregeln nachzuvollziehen, einschließlich der verwendeten Software und des entsprechenden Data Stewards. Als mögliche Standards werden HL7 FHIR und der ISO 21526 betrachtet. FHIR als aufstrebender Kommunikationsstandard für das Gesundheitswesen definiert insbesondere drei Ressourcen, die zu den Anforderungen passen: ConceptMap, StructureMap und Provenance. ConceptMaps können die konzeptionelle Beziehung zwischen Datenelementen abbilden und sind daher ein passendes Eingabeformat für den Editor. Der zuvor beschriebene Matchingdienst unterstützt den Export als ConceptMap und bewies deren Eignung für die vollständige Abbildung von Matches. Diese ConceptMaps enthalten verschiedene Informations-Tripel: Quell- und Ziel-URI zur Identifizierung der Datenelemente und ih-

re Ähnlichkeitsbeziehung, um ihre Verwandtschaft auszudrücken. StructureMaps werden definiert, um Regeln zur Transformation von Elementen in andere Repräsentationen darzustellen. Der ISO-Standard bietet im Vergleich zu seinem Vorgänger ISO 11179 durch die Einführung des Mapping-Packages die Möglichkeit Verbindungen zwischen Datenelementen direkt im MDR zu speichern. Es gibt zwei Mapping-Entitäten, das übergeordnete MDRMapping und die eingeschlossenen Maps, welche die direkten Verbindungen zwischen Datenelementen darstellen. Eine Map enthält mindestens ein Quell- und ein Zieldatenelement, erlaubt aber auch eine Komposition von Datenelementen. Zusätzlich muss ein Mapping-Type aus einem festen Katalog ausgewählt werden. Der Editor soll also FHIR-Ressourcen als Eingabeformat nutzen und als Ausgabeformate FHIR und eine ISO-konforme Darstellung verwenden.

Die Entwicklung des Editors konzentrierte sich auf einen modularen Aufbau mit guter Unterstützung für den Data Steward und eine Reproduzierbarkeit der erstellten Mappings. Der Anwender kann eine ConceptMap entweder manuell hochladen oder auf einen Ressourcenstandort auf einem verfügbaren FHIR-Server verweisen. Im letzteren Fall fragt der Editor nach einer mit der ConceptMap verknüpften Provenance-Ressource, um die wertvollen Informationen zu erhalten. Der Inhalt der ConceptMap wird mit zusätzlichen Details zum Quell- und Zieldatenelement über den MDR.Injector [Kern et al., 2018] visualisiert. Diese Bibliothek löst den URI in menschenlesbare Repräsentationen auf, um weitere nützliche Informationen über die entsprechenden Datenelemente interaktiv bereitzustellen. Der Editor basiert auf der JavaScript-Bibliothek Ace, die Syntax-Highlighting für verschiedene Skriptsprachen und Live-Checking zur Fehlerminimierung bietet. Der Data Steward kann die Transformationsregeln für jedes Tripel definieren, Beispieldaten zur Validierung des Mappings in Echtzeit verwenden und sich vom Editor einen Vorschlag unterbreiten lassen. Die präferierte Skriptsprache für die Transformationsregeln ist JavaScript, da es neben der Ausdrucksstärke für eine Verwendung im Browser-gestützten Editor geeignet ist. Zudem kann JavaScript im nächsten Verarbeitungsschritt (siehe Kapitel 4.3) effektiv genutzt werden. Die Speicherung der Transformationsregeln erfolgt in einem Datenbankdienst, der schon für den Matchingdienst verwendet wurde, siehe Kapitel 4.1. Der Editor nutzt OpenID Connect zur Wiederverwendung bestehender Authentifizierungssysteme, um die für die Provenance-Ressource benötigten Benutzerinformationen zu erhalten. Die Zuständigkeit des Data Stewards für die Transformation wird so einheitlich dokumentiert und, falls vorhanden, mit der bisherigen Provenance-Ressource verknüpft.

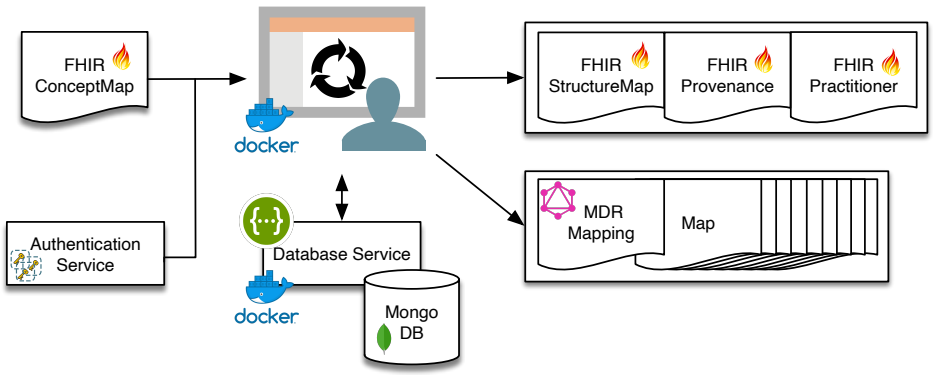


Abbildung 4.5: Der Editor verwendet standardisierte FHIR ConceptMaps als Eingabe. Die verpflichtende Anmeldung über OpenID Connect ist für Identifikation der Data Stewards notwendig und auch zur Abfrage von geschützten Metadaten nötig. Die erstellten Mappings inklusive der Transformationsregeln werden als FHIR R4-Ressourcen oder ISO 21526 MDRMappings exportiert.

Zudem wurde eine Vorschlagsfunktion implementiert, welche den Data Stewards vorhandene Transformationsregeln vorschlagen kann. Dazu wird einerseits in der Datenbank anhand der UUID gesucht, ob das Ziel- oder Quellelement schon einmal verarbeitet wurden. Anderseits können Transformationen ähnlicher Datenelemente gesucht werden. Hierzu wird für die Definition des Datenelements ein lokal sensitiver Hashwert berechnet. Lokal sensitive Hash-Funktionen sind so konzipiert, dass Hash-Kollisionen für zwei *nahe* Zeichenketten wahrscheinlicher sind als für entfernte Zeichenketten. Diese Algorithmen werden oft im Information Retrieval für die Dublettenerkennung genutzt [Joshi et al., 2014], aber auch in der Analyse von Genomdaten, speziell beim Sequenzalignment [Berlin et al., 2015]. Das Ergebnis wird dann mit den zuvor berechneten Werten aus der Datenbank verglichen. Die implementierte Vorschlagsfunktion basiert auf der lokal sensitiven Hash-Funktion MinHash[Broder, 1997]. Dafür werden beide Zeichenketten in *Schindeln (Shingles)* der Größe 3 aufgeteilt, gehasht und der Abstand über den Jaccard-Koeffizenten quantifiziert. Somit können Transformationsregeln ähnlicher Datenelemente gefunden werden und dem Data Steward zur Nutzung oder ggf. Anpassung vorgeschlagen werden. Der Editor exportiert das resultierende Mapping entweder als FHIR-Bundle mit einer StructureMap, einer Provenance und dem entsprechenden Steward in eine Practitioner-Ressource oder als QL^{4}MDR-kompatibles JSON, welches dann ins MDR zurückgespielt werden kann. Bei der Speicherung im MDR können die Provenance-Informationen leider nicht gespeichert werden. In einem vorangegangen Pro-

jekt [Ulrich et al., 2016] konnte gezeigt werden, dass Metadaten nach der ISO 11179 mit FHIR geeignet referenziert und kombiniert gespeichert werden konnten. Ein ähnliches Vorgehen wäre möglich, doch werden seitens FHIR die dafür benötigten Ressourcen nicht mehr unterstützt.

4.3 Transformation unter Verwendung von Mirth Connect

Die zuvor erstellten Mappings müssen im letzten Verarbeitungsschritt genutzt werden, um die erstellten Konvertierungsregeln auf Instanzdaten anzuwenden. Dazu wird eine Konvertierungsapplikation benötigt, welche die Regeln einliest und dann auf die richtigen Instanzdatenfelder anwendet. Als Konvertierungapplikation wurde Mirth Connect [Vireq, 2021] gewählt, eine Schnittstellen-Applikation, welche im Gesundheitswesen für die Verwaltung und Verarbeitung Nachrichten-gestützte Kommunikation genutzt wird. In der Literatur finden sich verschiedene Berichte, wie Mirth Connect erfolgreich für die Datenintegration im klinischen Umfeld genutzt wurde. So wird die Verwendung von Mirth Connect in verschiedenen US-amerikanischen Krankenhäusern[Bortis, 2008], ebenso wie Camacho Rodriguez et al. [Camacho Rodriguez et al., 2016]. In deren Studie wird Mirth Connect für die Aufbereitung von HL7 v2 Labornachrichten für die direkte Verwendung in Open Clinica beschrieben. Lin et al. [Lin et al., 2019] verwenden in ihrer Arbeit Mirth Connect für die Umstellung eines Krankenhausinformationssystems inklusive der Altdatenübernahme.

Die Kommunikation in Mirth Connect erfolgt über so genannte Kanäle (*channels*). Kanäle bestehen aus einem Filter, der nachgelagerten Datentransformation und Datenausleitung. Der Filter überprüft eingehende Nachrichten, ob sie für die Transformation dieses Kanals geeignet sind. Die korrekten Daten werden dann anhand von Transformationsregeln umgeformt; hierfür verwendet Mirth Connect JavaScript. Mirth Connect schränkt hierbei die Funktionalitäten von JavaScript nicht ein und stellt zusätzlich für eine einfache Benutzung *Code Templates* bereit. Die in Templates enthaltenen vordefinierten JavaScript-Funktionen erhöhen die Wiederverwendbarkeit von Codesegmenten. Desweiteren können in den Code Templates über Mirth Extensions native Java Packages zur Laufzeit nachgeladen werden und somit auch externe Dienste per HTTP-Anfragen eingebunden werden. Die transformierten Nachrichten können dann über verschiedene

Protokolle aus- und weitergeleitet werden. Mirth unterstützt hierbei eine Vielzahl an Standardprotokollen wie Transmission Control Protocol (TCP), Hypertext Transfer Protocol (HTTP), File Transfer Protocol (FTP), aber auch medizininformatische Protokolle: HL7 Minimum Lower Layer Protocol (MLLP) und Digital Imaging and Communications in Medicine (DICOM). Mirth erlaubt auch ein direktes Schreiben ins Dateisystem. Im Zusammenhang mit den speziellen Anwendungsprotokollen unterstützt Mirth auch weitere medizininformatische Formate wie HL7 v2 und HL7 FHIR. Beispielsweise kann eine eingehende Nachricht aus einem Laborinformationssystem für die Verwendung in einem Forschungs-Data Warehouse pseudonymisiert werden, in eine FHIR Observation umgewandelt und direkt übertragen werden. Zudem können administrative Teilinformationen, wie ein Datum und eine Pseudonymisierungsart in eine weitere Datenbank für Dokumentationszwecke geschrieben werden.

Mirth Channel Factory

Ein Mirth Channel mit allen wichtigen Definitionen, wie Filtern und Transformationsregeln, wird in einer XML-Datei serialisiert. Somit kann die Erstellung von Channels außerhalb von Mirth Connect erfolgen und die Transformationslogik in die Channel eingefügt werden. Dafür wurde die *Mirth Channel Factory* implementiert. Die *Factory* braucht für die Erstellung einer validen Channel-Definition drei Eingabeparameter: die Quell- und Zielstruktur und die dazugehörigen Transformationsregeln. Alle Informationen sind in den vorherigen Schritten (siehe Kapitel 4.1 und 4.2) erstellt oder sind im MDR über QL^{4}MDR verfügbar. Die Quellstruktur wird einerseits für den Filter als Referenzstruktur genutzt und andererseits um die Eingabefelder bei der Transformation richtig zu benennen. Die werden dann über die Transformationsregeln auf die Zielstruktur übertragen. Die Ziel- und Quellstruktur müssen Base64 kodiert in der Channel Definition gespeichert werden, die Transformationsregeln können als JavaScript-Quellcode direkt in das korrespondierende Segment integriert werden.

Nachdem alle Transformationsregeln verarbeitet wurden, gibt die Mirth Channel Factory einen vollständigen Mirth Connect-Kanal aus.

Es sind keine weiteren manuellen Schritte notwendig. Des Weiteren sind die generierten Kanäle nicht projektspezifisch, d.h. die generierten Channels können in jeder Mirth Connect-Anwendung importiert und genutzt werden. Mirth Connect stellt ein offizielles Docker Image zur Verfügung und kann somit über Container schnell bereitgestellt werden. Über die eingebauten Skalierungssysteme wie Docker Swarm kann eine gut

skalierbare Datenintegrationsmaschinerie erstellt werden, welche Metadaten-getriebene Nachrichten aus ihrem Quellformat in das gewünschte Zielformat überführt. Durch die Vielzahl an unterstützten Formaten und produktive Verwendung vom Mirth Connect im Gesundheitswesen kann die Mirth Channel Factory auf Basis der vorangegangen Metadatenverarbeitung die Datenintegration von lokalen Projekten beschleunigen und durch die einfache Adaption und Serialisierung der Channels diese über Institutionsgrenzen ermöglichen.

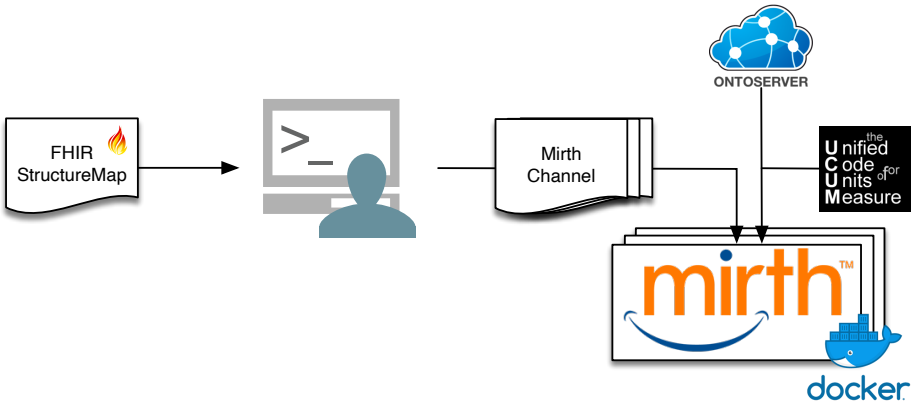


Abbildung 4.6: Die Grafik zeigt die Verarbeitung der FHIR StructureMaps zu Mirth Channels. Die Transformationsregeln aus den StructureMaps werden von Mirth Connect genutzt, um eingehende Nachrichten zu konvertieren. Dabei können auch externe Dienste für die Konvertierung eingebunden werden, beispielsweise ein Terminologieserver für die semantische Anreicherung der Daten oder weitere spezialisierte Konvertierungsdienste.

Kapitel 5

Diskussion und Zusammenfassung

Im Verlauf der Arbeit wurden Metadaten als Integrationswerkzeug ausführlich untersucht und ihre Verwendungen evaluiert, zudem wurde eine Integrationsstrategie für Metadaten vorgestellt. Im folgenden Kapitel werden die eingangs aufgeworfenen Forschungsfragen diskutiert:

1. Sind Metadaten für die Datenintegration geeignet?
2. Welche Vorteile ergeben sich, wenn (Meta-)daten in einen föderalen Verbund gegeben werden?
3. Kann die Verwendung von Metadaten die Sekundärnutzung von klinischen Daten fördern?
4. Werden klassische MDR-Systeme durch die voranschreitende Verbreitung von FHIR abgelöst?

1. Sind Metadaten für die Datenintegration geeignet?
Die eingehende Untersuchung von Metadaten im Rahmen der systematischen Literaturanalyse, sowie relevanten Vorarbeiten [Kock-Schoppenhauer et al., 2018b, Kock-Schoppenhauer et al., 2019] zeigten, dass Metadaten eine detailreiche Beschreibung von Forschungsdaten ermöglichen. Durch die Abstraktion und Atomisierung der Daten lässt sich die Ausdrucksvielfalt von klinischen Studien abbilden und vor allem angemessen strukturieren. Zudem können anhand der strukturellen Abstraktion große Datenmengen auf Konsistenz und Qualität überprüft werden [Kahn et al., 2016, Schmidt et al., 2021]. Folglich lassen sich nachfolgende Probleme in der Verwendung frühzeitig erkennen und an die Datenquellen kommunizieren. Doch für die gewinnbringende Anwendung müssen Metadaten gut definiert und gepflegt werden, um möglichst vielfältige Verwendungsmöglichkeiten zu gewährleisten. Allein die Erstellung von Metadaten liefert keinen

Mehrwert. Es bedarf weiterer Faktoren und Komponenten für einen erfolgreichen und langfristigen Einsatz in der Datenintegration. Diese Faktoren wurden in der Analyse herausgearbeitet und flossen in eine neue Komponentenklassifikation von Metadaten (siehe Abbildung 2.11). Der Klassifikation folgend, sollten Metadaten durch eine *Schema Definition* anhand von Standards, einem Metadatenschema mit strukturellen Annotationen und weiteren semantischen Deskriptoren für Inhalt und Kontext beschrieben sein. Die entwickelte Struktur erhöht die Vergleichbarkeit und unterstützt damit den Austausch der (Meta-)Daten. Zudem fördert die Verwendung von Standards und maschinenlesbaren Strukturinformationen die Verarbeitungsmöglichkeiten wie das automatisierte Matching. Neben dem strukturellen und möglichst standardisierten Fundament ist eine semantische Beschreibung der Metadaten ausschlaggebend für eine langfristige Verwendung. Durch die Annotationen können Inhalt und Kontext maschinenauswertbar beschrieben werden. Da gerade im Bereich der klinischen Studien viel über vordefinierte Auswahllisten ausgedrückt wird, ist es hier entscheidend, dass bis auf die unterste Code-Value-Ebene annotiert wird. Dadurch sind nicht nur die Fragen maschinell auswertbar, sondern auch alle möglichen Antwortausprägungen. Neben der inhaltlichen Beschreibung, sollte eine strukturelle Annotation der klinischen Studien über die LOINC Document Ontology erfolgen. Dies kann ein Matching und Mapping für die weitere Verwendung vereinfachen. Wie die Definition der Metadaten wird auch die Annotierung meist händisch durchgeführt und stellt damit einen ressourcen- und zeitintensiven Prozess dar. Daher müssen für eine langfristige Verwendung der Annotationen grundsätzliche organisatorische Angaben vorhanden sein. Wie die vorherige Analyse (siehe Kapitel 2.3) von Annotationen zeigte, ist neben der Auswahl des richtigen Vokabulars auch eine systematische Angabe des verwendeten Systems inklusive der verwendeten Version zwingend erforderlich. Sind alle benannten Voraussetzung erfüllt, ist die erste Forschungsfrage positiv zu beantworten ist: Metadaten sind durch ihre Flexibilität ein geeignetes Werkzeug für die Datenintegration von klinischen Projektdaten. Durch die Metadaten kann einerseits der Kontext und anderseits die Qualität nachvollziehbar geprüft werden. Dennoch müssen auch die Metadaten sorgfältig erstellt, annotiert bzw. gepflegt werden und es fehlt an einem flächendeckenden und einheitlichen Zugang zu den Daten. Hier kann das beschriebene Verbundskonzept eine Verbesserung erzielen.

2. Warum sollten (Meta-)daten in einen föderalen Verbund gegeben werden?

Die eigenen Metadaten in einem Verbund zu veröffentlichen, soll die Sichtbarkeit und den gemeinsamen Nutzen stärken. Jedoch stehen diesem Vorgehen auch Bedenken gegenüber. Die Herausgabe der eigenen Daten an Dritte erweckt stets Vorbehalte. Dies gilt ebenso für Metadaten. Ungleich zu klinischen Patientendaten, wo die Herausgabe datenschutzrechtliche Hindernisse birgt, sind Metadaten in dieser Hinsicht unproblematisch. Allerdings kommt es dabei maßgeblich auf den Zeitpunkt der Veröffentlichung an. Denn durch die Metadaten lassen sich Rückschlüsse auf das eigene Forschungsvorhaben und die Methodik schließen. Dies birgt im Sinne der incentive-getriebenen Wissenschaft ein nicht zu leugnendes Problem. Dieses Problem besteht gleichwohl meist temporär. Die nachträgliche Veröffentlichung der Metadaten birgt überdies den Vorteil, dass die verwendeten Daten besser nachvollziehbar sind und gegebenenfalls der Wissenschaft erneut zur Verfügung stehen können [Dugas et al., 2015]. Im Weiteren erzeugt die Teilnahme an einem Verbund mehr Aufwand durch die Bereitstellung und Pflege der Software, durch eine evtl. redundante Lagerung der Daten entsteht Synchronisationsaufwand, wie in Kapitel 3.4.3 beschrieben. Gegenüber dem Mehraufwand eines Verbundsbeitritts stehen die dadurch gewonnenen Möglichkeiten. Metadaten dienen in erster Linie der Beschreibung der eigentlichen Forschungsdaten. Durch die Veröffentlichung der Metadaten wird die Sichtbarkeit erhöht und die generelle Datenverfügbarkeit gestärkt. Allein durch die Kenntnis darüber, dass ein Datensatz zu einer bestimmten Thematik existiert, birgt bereits einen Mehrwert. Dadurch kann beispielsweise die kostenintensive und langwierige Datenerhebung erspart werden. Erst die höhere Sichtbarkeit der Forschungsdaten ermöglicht eine Wiederverwendung durch andere Forschungsgruppen. Dies führt zu Kooperationen und Veröffentlichungen und steigert die wissenschaftliche Verwertung der Daten. Zudem kann eine Kombination verschiedener Datensätze neue Ergebnisse produzieren, die nur mit einem Datensatz nicht möglich wären. Zudem kann die langfristige Archivierung der Metadaten eine dedizierte Aufgabe des Verbunds sein – das Vorhalten der Metadaten wird mitunter in den FAIR-Prinzipien [Wilkinson et al., 2016] explizit gefordert. Ein solcher Verbund lebt von der produktiven Teilnahme der MDR-Betreiber und der Bereitschaft, Daten zur Verfügung zu stellen. Die Erfahrung aus vergangenen Projekten [Lablans et al., 2015] zeigt: Je mehr Teilnehmende aktiv partizipieren, desto höher ist die wissenschaftliche Ausbeute für alle Beteiligten des Verbunds.

3. Kann die Verwendung von Metadaten die Sekundärnutzung von klinischen Daten fördern?

Eine Sekundärnutzung klinischer Daten findet gegenwärtig wenig Anwendung, u.a. da wichtige Informationen bezüglich der Daten nicht ausreichend dokumentiert sind. In einer Diabetesstudie kann es beispielsweise entscheidend sein, ob eine Blutzuckermessung vor oder nach dem Essen durchgeführt wurde. Für eine gesicherte Sekundärnutzung spielen daher verschiedene Faktoren eine Rolle: Es braucht eine standardisierte Strukturbeschreibung, um die Daten syntaktisch zu verstehen. Zudem werden maschinenlesbare Kontextinformationen benötigt, die den Erhebungskontext beschreiben, um zu entscheiden, welche Daten genutzt werden können. Abschließend muss die Qualität der Daten überprüfbar sein, sodass die Ergebnisse auch verwertbar sind. Diese drei Faktoren lassen sich über Metadaten erfüllen. Durch die Verwendung von Metadatenstandards und eine Annotation auf Attributsebene sind Informationen maschinenlesbar vorhanden und lassen sich dadurch leichter und nachhaltiger verarbeiten. Zudem können durch das Matching und Mapping der Schemata mehr Daten erschlossen werden, welche dann einer Sekundärnutzung zur Verfügung stehen. Die Bereitstellung der Metadaten und damit das maschinenlesbare Verständnis der dahinterliegenden Datensätze ist der Grundstein für einen allgemeinen Wissensgewinn und die Sekundärnutzung. Für einen flächendeckenden Einsatz werden darüber hinaus besser Routinedaten benötigt, die kontextsensitiv beschrieben werden. Sind alle diese strukturellen und qualitativen Hindernisse überwunden, bedarf es immer menschlicher Interaktion. Diese ist zeitintensiv aber notwendig zur Sicherung der Qualität und Konsistenz. Denn vergleichbar zur Sekundärnutzung sind die projektspezifischen Metadaten in erster Linie für einen anderen Zweck erhoben worden und dies erschwert eine Verarbeitung: sinnbildlich dafür steht das Datenelement mit dem Namen *Unequivocalprogressivediseaseinnontargetlesionsisbasedon: (pleasedecribe)* [Ulrich et al., 2017].

4. Werden klassische MDR-Systeme durch die voranschreitende Verbreitung von FHIR abgelöst?

Die separate Kuratierung von Metadaten erzeugt einen erheblichen Mehraufwand aufgrund der erforderlichen Bereitstellung eines MDR-Systems und der konsequenten Pflege und Aktualisierung der Daten. In der systematischen Analyse wurden aber auch verschiedene andere Metadatenansätze herausgearbeitet, siehe Kapitel 2.2.2. Ein Ansatz ist dabei von besonderem Interesse: Grewe et al. [Grewe et al., 2011] beschrieb ein generisches

Modell, welches fünf wichtige Eigenschaften aufweist: Erweiterbarkeit, Modularität, Verfeinerungen, Mehrsprachigkeit, maschinelle Verarbeitbarkeit. Diese fünf Merkmale erfüllt der Standard FHIR. In seiner Simplizität liegt der Grund seiner großen Verbreitung. Wie in Kapitel 2.1.3 beschrieben, zeichnet sich FHIR durch eine simple Modularisierung in Ressourcen und einen maschinenverarbeitbaren Erweiterungsmechanismus aus, der sich auch für eine projektspezifische Profilierung eignet. Vor diesem Hintergrund ist in den Blick zu nehmen, ob es sich überhaupt lohnt, klassische Metadaten zu pflegen, oder ob FHIR ausreichend ist. FHIR stellt mit den modularen Ressourcen Schemata, also strukturelle Metadaten, öffentlich und kostenlos bereit. Die Schemata sind dementsprechend bekannt und projektspezifische Erweiterungen sind zudem meist auf Simplifier [1] öffentlich verfügbar. Somit haben wir ein MDR-System gefüllt mit unzähligen Projekten, die ihre Daten über FHIR harmonisieren, wie der deutsche GECCO-Datensatz für die COVID-19 Forschung [Sass et al., 2020]. Die Instanzdaten sind entsprechend konform zu den Schemata vorhanden und können über die FHIR *Validator* gegen die Schemata geprüft und standardisiert übertragen werden. Die Validator können sogar tiefgehende Prüfungen vornehmen, beispielsweise, ob eine genutzte Annotation aus einer spezifischen Version des vorgegebenen Vokabulars entstammt. FHIR bietet zudem auch Ressourcen für die Beschreibung und den Austausch von Strukturinformationen, die auch in dieser Arbeit Verwendung fanden: ConceptMap, StructureMap und Provenance. Folglich bietet FHIR viele vielversprechende Eigenschaften für eine Ablösung klassischer Ansätze mit MDR-Systemen nach ISO 11179/21526. Doch bei genauer Betrachtung unterscheiden sich die beiden Ansätze grundlegend: FHIR folgt dem Top-Down-Prinzip, wohingegen projektspezifische MDRs nach dem Bottom-Up-Prinzip ihre Daten erheben. FHIRs vordefinierte Schemata widersprechen dem Bottom-Up-Prinzip, auch wenn durch die Erweiterbarkeit eine sinnvolle, aber gefährliche Freiheit entsteht. Im Rahmen einer Lehrveranstaltung mussten beispielsweise Studierende verschiedene bekannte medizinische *Entitäten* wie den Impfausweis mittels FHIR modellieren. Jede Gruppe modellierte den bekannten Pass jedoch im Detail unterschiedlich. Hierbei trat ein Problem in der Arbeit mit FHIR deutlich hervor: FHIR ist strikt genug, um die Erweiterungen zu benötigen, aber ausreichend flexibel, um Modellvarianz zuzulassen. Die Daten aus klinischen Projekten sind heterogen und brauchen ein präzises Werkzeug zur Beschreibung. Ein kuratiertes, ISO-kompatibles MDR-System ist flexibel und durch Annotationen ausdrucksstark

[1] https://simplifier.net/

genug, um die Varianz von klinischen Projektdaten abzubilden. Aufgrund der Atomisierung und ohne vordefinierte Kompositionen sind Metadaten flexibler verwendbar und dadurch gegenüber FHIR für die Sekundärnutzung besser geeignet. Betrachtet man zudem die eingeführte Komponentenklassifikation, liegen die ISO-Standards und FHIR auf verschiedenen Ebenen. FHIR als Sammlung von Ressourcen ist eine Sammlung von Metadatenschemata, wohingegen die ISO-Standards eine Stufe darüber als Schema Definitions die Schemata prägen. Diese Trennung konnte in einem vorangegangen Projekt erfolgreich gezeigt werden [Ulrich et al., 2016]: Metadaten dienen der Repräsentation der strukturellen Inhalte und mit FHIR lassen sich die in beispielsweise Fragebögen zusammengestellten Informationen besser austauschen. Für einen Austausch der atomisierten Datenelemente ist FHIR jedoch nicht geeignet, da es keine geeigneten Ressourcen dafür gibt. Diese Lücke hingegen soll die neue Schnittstelle QL^{4}MDR schließen. Die Frage, ob FHIR klassische MDR-Systeme vollständig ablösen kann, ist somit zu verneinen. Im Gegenteil ist eher davon auszugehen, dass MDR-Systeme einen höheren Stellenwert bekommen werden, weil der Wunsch nach Datenverfügbarkeit und Bearbeitung der Fülle an Daten im Vordergrund stehen wird. So haben die klassischen Systeme auch in FHIR-Zeiten ihre Daseinsberechtigung. Sie werden benötigt, um die Bottom-Up-Projekte in all ihren Facetten maschinenlesbar abzubilden.

Zusammenfassung

Metadaten sind ein leistungsfähiges Werkzeug der Datenintegration und können zur Identifizierung, Beschreibung und Verarbeitung von Informationen hilfreich sein, auch wenn ihre nachhaltige Definition eine Herausforderung darstellt. Im Rahmen dieser Arbeit wurden daher die Erstellung von Metadaten als wichtiger Grundstein für die Verarbeitung und die weitere Verwendung von Metadaten im klinischen Kontext untersucht. Dabei lieferte eine systematische Literaturanalyse, neben harmonisierten Definitionen der Terme Matching, Mapping und Transformation, einen Überblick über die Verwendung und die damit verbundenen Probleme und korrespondieren Lösungen. Zudem wurde eine neue Komponentenklassifikation vorgestellt, welche den strukturellen Aufbau von Metadatenschemata beleuchtet und durch den ganzheitlichen Überblick inhaltliche Missverständnisse zwischen Domänenexperten verhindern soll. Aus den Erkenntnissen der Analyse und vorangegangen Projekten wurde der Bedarf nach einem föderalen Metadatenkonzept erkannt. Dazu wurde eine Anforderungsanalyse aufgestellt, um wichtige Voraussetzungen und Faktoren für einen erfolgreichen Metadatenverbund zu erörtern.

Da der potenzielle Erfolg des Verbundes von der aktiven Partizipation der standortspezifischen MDR-Betreiber abhängig ist, wurde eine strukturierte Umfrage durchgeführt, um den Bedarf und mögliche Hindernisse zu ermitteln. Bestärkt durch die positive Rückmeldung wurde eine Bestandsaufnahme potenzieller MDR-Systeme mit medizininformatischem Bezug aufgestellt und eingehend untersucht. Dabei stellte sich heraus, dass die gepflegten Metadaten nur schwer aus den einzelnen Systemen abfragbar waren. Um diese Siloisierung und Kommunikationsprobleme zu überwinden, wurde eine uniforme Schnittstelle, QL^{4}MDR, speziell für die Verfügbarkeit und den Austausch von Metadaten konzipiert und erfolgreich in verschiedene MDR-Systeme integriert. Die Schnittstelle kombiniert das ISO-konforme 11179-3 Datenmodell mit modernen Webtechnologien und soll durch die Erweiterung auf den ISO 21526 auch zukünftige MDR-Systeme unterstützen und die standardübergreifende Metadatenkommunikation sicherstellen.

Basierend auf dem uniformen Schnittstellenkonzept wurden verschiedene Mehrwertdienste implementiert. Durch die technische Hilfe bei dem Matching, Mapping und Transformieren sollen klinische Metadaten einfacher verarbeitet werden. Dies stellt einen ersten Schritt für die technische Integration der klinischen Projektdaten für eine mögliche Sekundärnutzung dar. Metadaten als zentrales Werkzeug erwiesen sich als geeignet, sogleich die reine Erstellung von Metadaten kein Selbstzweck ist: Nur wohl definierte und kuratierte Metadaten sind langfristig ein effektives Werkzeug für die Integration klinischer Projektdaten.

Abkürzungsverzeichnis

KIS Krankenhausinformationssystem . . . 2

RDF Resource Description Framework . . . 3

NISO National Information Standards Organization . . . 7

W3C World Wide Web Consortium . . . 7

HL7 Health Level 7 . . . 16

FHIR Fast Healthcare Interoperability Resources . . . 16

SKOS Simple Knowledge Organization System . . . 16

PRISMA Preferred Reporting Items for Systematic Reviews and Meta-Analyses . . . 20

CDISC Clinical Data Interchange Standards Consortium . . . 27

ODM Operational Data Model . . . 27

OMOP Observational Medical Outcomes Partnership . . . 27

IHE Integrating the Healthcare Enterprise . . . 27

DEX Data Element Exchange . . . 27

ASTM American Society for Testing and Materials . . . 27

CCR Continuity of Care Record . . . 27

XML Extensible Markup Language . . . 27

OWL Web Ontology Language . . . 27

JSON-LD JavaScript Object Notation for Linked Data . . . 27

ClaML Classification Markup Language . . . 27

UMLS Unified Medical Language System . . . 27

SNOMED CT Systematized Nomenclature of Medicine - Clinical Terms . . . 27

LOINC Logical Observation Identifiers Names and Codes . . . 27

MedDRA Medical Dictionary for Regulatory Activities . . . 27

5 Abkürzungsverzeichnis

EAD Encoded Archival Description . . . 27

GILS Global Information Locator Service . . . 27

VRA Core Visual Resources Association Core . . . 27

CIMI Consortium for Interchange of Museum Information . . . 27

CSDGM Content Standard For Digital Geospatial Metadata . . . 27

ONIX ONline Information eXchange . . . 27

MARC MAchine-Readable Cataloging . . . 27

TMA DES Tissue Microarray Data Exchange Specification . . . 27

EXIF Exchangeable Image File Format . . . 27

INSPIRE Infrastructure for Spatial Information in the European Community . . . 27

DCAT Data Catalog Vocabulary . . . 27

NLP Natural Language Processing . . . 30

CDE Common Data Element . . . 36

JSON JavaScript Object Notation . . . 36

NLM National Library of Medicine . . . 38

WHO Weltgesundheitsorganisation . . . 38

ICD-10 Klassifikation der Krankheiten und verwandter Gesundheitsprobleme . . . 38

ICD-10-GM Klassifikation der Krankheiten und verwandter Gesundheitsprobleme - Deutsche Modifikation . . . 38

FSN Fully Specified Name . . . 40

MDM Medical-Data-Models-Portal . . . 41

PCE Postkoordinierten Ausdrücke . . . 47

ECL Expression Constraint Language . . . 47

DICOM Digital Imaging and Communications in Medicine . . . 49

TD Top-Down . . . 50

BU Bottom-Up . . . 50

TMF Technologie- und Methodenplattform für die vernetzte medizinische Forschung e.V. . . . 52

DACH Deutschland, Österreich und die Schweiz 56

METeOR Metadata Online Registry 59

CaDSR Cancer Data Standards Registry and Repository 59

MII Medizininformatik Initiative 59

ELGA elektronische Gesundheitsakte 59

CKM Clinical Knowledge Manager 59

JWT JSON Web Token 61

HTTP Hypertext Transfer Protocol 61

TCP Transmission Control Protocol 94

FTP File Transfer Protocol 94

MLLP Minimum Lower Layer Protocol 94

Literaturverzeichnis

[Aschhoff et al., 2013] Aschhoff, M., Rimatzki, B., Breil, B., and Haas, P. (2013). Model-Mapping zwischen ISO 11179-basiertem Meta Data Repository und gängigen Standards für die Standardisierung klinischer Dokumente. page DocAbstr.154. German Medical Science GMS Publishing House.

[Ashish et al., 2016] Ashish, N., Dewan, P., and Toga, A. W. (2016). The GAAIN entity mapper: An active-learning system for medical data mapping. *Frontiers in Neuroinformatics*, 9(JAN2016):1–10.

[Atzeni et al., 2019] Atzeni, P., Bellomarini, L., Papotti, P., and Torlone, R. (2019). Meta-mappings for schema mapping reuse. *Proceedings of the VLDB Endowment*, 12(5):557–569.

[Aumueller et al., 2005] Aumueller, D., Do, H.-H., Massmann, S., and Rahm, E. (2005). Schema and Ontology Matching with COMA++. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, SIGMOD '05, pages 906–908, New York, NY, USA. ACM.

[Australian Institute of Health and Welfare, 2021] Australian Institute of Health and Welfare (2021). METeOR home. https://meteor.aihw.gov.au/content/index.phtml/itemId/181414, Zugriffsdatum: 26. Oktober 2021.

[Baek and Sugimoto, 2012] Baek, J. and Sugimoto, S. (2012). A task-centric model for archival metadata schema mapping based on the records lifecycle. *International Journal of Metadata, Semantics and Ontologies*, 7(4):269–282.

[Ballinger et al., 2004] Ballinger, K., Box, D., Curbera, F., Davanum, S., Ferguson, D., Graham, S., Liu, C. K., Leymann, F., Lovering, B., and Nadalin, A. (2004). Web services metadata exchange (WS-MetadataExchange). *OASIS draft*.

[Batini et al., 1986] Batini, C., Lenzerini, M., and Navathe, S. B. (1986). A comparative analysis of methodologies for database schema integration. *ACM computing surveys (CSUR)*, 18(4):323–364.

[Bellahsene et al., 2011] Bellahsene, Z., Bonifati, A., and Rahm, E., editors (2011). *Schema Matching and Mapping*. Data-Centric Systems and Applications. Springer-Verlag, Berlin Heidelberg.

[Bender and Sartipi, 2013] Bender, D. and Sartipi, K. (2013). HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. In *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems*, pages 326–331. IEEE.

[Benson and Grieve, 2016] Benson, T. and Grieve, G. (2016). *Principles of Health Interoperability*. Health Information Technology Standards. Springer International Publishing.

[Berlin et al., 2015] Berlin, K., Koren, S., Chin, C.-S., Drake, J. P., Landolin, J. M., and Phillippy, A. M. (2015). Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nature Biotechnology*, 33(6):623–630.

[Bernstein et al., 2017] Bernstein, M. N., Doan, A., and Dewey, C. N. (2017). MetaSRA: Normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, 33(18):2914–2923.

[Berry and Edgar, 2019] Berry, S. D. and Edgar, H. J. (2019). Standardizing data from the dead. *Studies in Health Technology and Informatics*, 264:1427–1428.

[Bortis, 2008] Bortis, G. (2008). Experiences with Mirth: An open source health care integration engine. In *Proceedings of the 30th International Conference on Software Engineering*, pages 649–652.

[Braunstein and Detmer, 2016] Braunstein, M. L. and Detmer, D. (2016). Interoperable informatics for health enterprise transformation. *Journal of Enterprise Transformation*, 6(3-4):110–119.

[Breil et al., 2012] Breil, B., Kenneweg, J., Fritz, F., Bruland, P., Doods, D., Trinczek, B., and Dugas, M. (2012). Multilingual medical data models in ODM format: A

novel form-based approach to semantic interoperability between routine healthcare and clinical research. *Applied Clinical Informatics*, 3(3):276–289.

[Broder, 1997] Broder, A. Z. (1997). On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

[Bruland et al., 2017] Bruland, P., Doods, J., Storck, M., and Dugas, M. (2017). What information does Your EHR contain? Automatic generation of a clinical metadata warehouse (CMDW) to support identification and data access within distributed clinical research networks. In Gundlapalli, AV and Jaulent, MC and Zhao, D, editor, *Studies in Health Technology and Informatics*, volume 245 of *Studies in Health Technology and Informatics*, pages 313–317. Chinese Med Informat Assoc.

[Buna, 2016] Buna, S. (2016). *Learning GraphQL and Relay*. Packt Publishing Ltd.

[Camacho Rodriguez et al., 2016] Camacho Rodriguez, J. C., Stäubert, S., and Löbe, M. (2016). Automated import of clinical data from HL7 messages into OpenClinica and tranSMART using Mirth Connect. In *Exploring Complexity in Health: An Interdisciplinary Systems Approach*, pages 317–321. IOS Press.

[Canakoglu et al., 2019] Canakoglu, A., Bernasconi, A., Colombo, A., Masseroli, M., and Ceri, S. (2019). GenoSurf: Metadata driven semantic search system for integrated genomic datasets. *Database : the journal of biological databases and curation*, 2019:baz132.

[Charles et al., 2013] Charles, V., Isaac, A., Fernie, K., Dallas, C., Gavrilis, D., and Angelis, S. (2013). Achieving interoperability between the CARARE schema for monuments and sites and the Europeana Data Model. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 115–125.

[Chu et al., 2018] Chu, Y. C., Kuo, W. T., Cheng, Y. R., Lee, C. Y., Shiau, C. Y., Tarng, D. C., and Lai, F. (2018). A Survival Metadata Analysis Responsive Tool (SMART) for web-based analysis of patient survival and risk. *Scientific Reports*, 8(1).

[Cimino, 1998] Cimino, J. (1998). Desiderata for Controlled Medical Vocabularies in the Twenty-First Century. *Methods of information in medicine*, 37(4-5):394–403.

[Corradi et al., 2012] Corradi, L., Porro, I., Schenone, A., Momeni, P., Ferrari, R., Nobili, F., Ferrara, M., Arnulfo, G., and Fato, M. M. (2012). A repository based on a dynamically extensible data model supporting multidisciplinary research in neuroscience. *BMC medical informatics and decision making*, 12(1):115.

[Covitz et al., 2003] Covitz, P. A., Hartel, F., Schaefer, C., De Coronado, S., Fragoso, G., Sahni, H., Gustafson, S., and Buetow, K. H. (2003). caCORE: A common infrastructure for cancer informatics. *Bioinformatics*, 19(18):2404–2412.

[Cunningham et al., 2016] Cunningham, S. G., Carinci, F., Brillante, M., Leese, G. P., McAlpine, R. R., Azzopardi, J., Beck, P., Bratina, N., Bocquet, V., Doggen, K., Jarosz-Chobot, P. K., Jecht, M., Lindblad, U., Moulton, T., Metelko, Nagy, A., Olympios, G., Pruna, S., Skeie, S., Storms, F., di Iorio, C. T., and Massi Benedetti, M. (2016). Core standards of the EUBIROD project: Defining a European diabetes data dictionary for clinical audit and healthcare delivery. *Methods of Information in Medicine*, 55(2):166–176.

[Daniel et al., 2014] Daniel, C., Sinaci, A., Ouagne, D., Sadou, E., Declerck, G., Kalra, D., Charlet, J., Forsberg, K., Bain, L., Mead, C., Hussain, S., and Laleci Erturkmen, G. B. (2014). Standard-based EHR-enabled applications for clinical research and patient safety: CDISC - IHE QRPH - EHR4CR & SALUS collaboration. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2014:19–25.

[De Jong et al., 2019] De Jong, F., Maegaard, B., De Smedt, K., Fišer, D., and Van Uytvanck, D. (2019). Clarin: Towards fair and responsible data science using language resources. In *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, pages 3259–3264.

[Deppenwiese et al., 2019] Deppenwiese, N., Duhm-Harbeck, P., Ingenerf, J., and Ulrich, H. (2019). MDRCupid: A Configurable Metadata Matching Toolbox. *Studies in Health Technology and Informatics*, 264:88–92.

[Dinu and Nadkarni, 2007] Dinu, V. and Nadkarni, P. (2007). Guidelines for the effective use of entity-attribute-value modeling for biomedical databases. *International Journal of Medical Informatics*, 76(11-12):769–779.

[Dugas, 2018] Dugas, M. (2018). Neuroinflammatory Biobank - Portal für Medizinische Datenmodelle (MDM-Portal).

[Dugas et al., 2019] Dugas, M., Hegselmann, S., Riepenhausen, S., Neuhaus, P., Greulich, L., Meidt, A., and Varghese, J. (2019). Compatible data models at design stage of medical information systems: Leveraging related data elements from the MDM portal. *Studies in Health Technology and Informatics*, 264:113–117.

[Dugas et al., 2015] Dugas, M., Jöckel, K.-H., Friede, T., Gefeller, O., Kieser, M., Marschollek, M., Ammenwerth, E., Röhrig, R., Knaup-Gregori, P., and Prokosch, H.-U. (2015). Memorandum Open Metadata. Open Access to Documentation Forms and Item Catalogs in Healthcare. *Methods of Information in Medicine*, 54(4):376–378.

[Dugas et al., 2016] Dugas, M., Neuhaus, P., Meidt, A., Doods, J., Storck, M., Bruland, P., and Varghese, J. (2016). Portal of medical data models: Information infrastructure for medical research and healthcare. *Database*, 2016.

[Eichenlaub et al., 2014] Eichenlaub, N., Morgan, M., and Masak-Mida, I. (2014). Undressing fashion metadata: Ryerson University Fashion Research Collection. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 191–195.

[Fielding and Taylor, 2000] Fielding, R. T. and Taylor, R. N. (2000). *Architectural Styles and the Design of Network-Based Software Architectures*. PhD Thesis, University of California, Irvine.

[Francis et al., 2013] Francis, W., Atkinson, R., Box, P., Rankine, T., Woodman, S., and Kostanski, L. (2013). Model-driven data harvesting to publish provenance for geospatial references. In *Proceedings of the 7th International Conference on Knowledge Capture: Knowledge Capture in the Age of Massive Web Data, K-CAP 2013*, page 121, New York, New York, USA. ACM Press.

[Frisendal, 2018] Frisendal, T. (2018). Getting the Structure Right. In *Visual Design of GraphQL Data*, pages 45–60. Springer.

[Frosini et al., 2018] Frosini, L., Bardi, A., Manghi, P., and Pagano, P. (2018). An aggregation framework for digital humanities infrastructures: The parthenos experience. *Scires-It*, 8(1):33–50.

[Gonçalves and Musen, 2019] Gonçalves, R. S. and Musen, M. A. (2019). The variable quality of metadata about biological samples used in biomedical experiments. *Scientific data*, 6(1):1–15.

[Gonzalez-Beltran et al., 2018] Gonzalez-Beltran, A. N., Campbell, J., Dunn, P., Guijarro, D., Ionescu, S., Kim, H., Lyle, J., Wiser, J., Sansone, S. A., and Rocca-Serra, P. (2018). Data discovery with DATS: Exemplar adoptions and lessons learned. *Journal of the American Medical Informatics Association*, 25(1):13–16.

[GraphQL Foundation, 2021] GraphQL Foundation (2021). Who's Using — GraphQL. https://graphql.org/users/, Zugriffsdatum: 26. Oktober 2021.

[Grewe et al., 2011] Grewe, J., Wachtler, T., and Benda, J. (2011). A Bottom-up Approach to Data Annotation in Neurophysiology. *Frontiers in Neuroinformatics*, 5.

[Guerra and Fernandes, 2013] Guerra, E. and Fernandes, C. (2013). A Qualitative and quantitative analysis on metadata-based frameworks usage. In Murgante, B and Misra, S and Carlini, M and Torre, CM and Nguyen, HQ and Taniar, D and Apduhan, BO and Gervasi, O., editor, *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 7972 LNCS of *Lecture Notes in Computer Science*, pages 375–390. Ho Chi Minh City Int Univ; Univ Perugia; Monash Univ; Kyushu Sangyo Univ; Univ Basilicata; Off Naval Res.

[Hall and McMahon, 2016] Hall, S. R. and McMahon, B. (2016). The implementation and evolution of STAR/CIF ontologies: Interoperability and preservation of structured data. *Data Science Journal*, 15.

[Hammad et al., 2020] Hammad, R., Barhoush, M., and Abed-alguni, B. H. (2020). A Semantic-Based Approach for Managing Healthcare Big Data: A Survey. *Journal of Healthcare Engineering*, 2020:e8865808.

[Haslhofer and Klas, 2010] Haslhofer, B. and Klas, W. (2010). A survey of techniques for achieving metadata interoperability. *ACM Computing Surveys*, 42(2):1–37.

[Hegselmann et al., 2021] Hegselmann, S., Storck, M., Gessner, S., Neuhaus, P., Varghese, J., Bruland, P., Meidt, A., Mertens, C., Riepenhausen, S., and Baier, S. (2021).

Pragmatic MDR: A metadata repository with bottom-up standardization of medical metadata through reuse. *BMC medical informatics and decision making*, 21(1):1–14.

[Holz et al., 2019] Holz, C., Kessler, T., Dugas, M., and Varghese, J. (2019). Core data elements in acute myeloid leukemia: Results from a unified medical language system-based semantic analysis and experts' review. *Journal of Medical Internet Research*, 21(8):e13554.

[Howarth, 2003] Howarth, L. C. (2003). *Designing a Common Namespace for Searching Metadata-Enabled Knowledge Repositories: An International Perspective*, volume 37 of *Cataloging & Classification Quarterly*. Taylor & Francis.

[Huang et al., 2017] Huang, G., Yuan, M., Li, C., and Sun, Q. (2017). Research on ontology generation and evaluation method in oil field based on the MDR. *Journal of Computational Methods in Sciences and Engineering*, 17(4):665–676.

[Hübner et al., 2020] Hübner, U., Esdar, M., Hüsers, J., Liebe, J.-D., Naumann, L., Thye, J., and Weiß, J.-P. (2020). It-report gesundheitswesen— schwerpunkt–wie reif ist die gesundheits-it aus anwenderperspektive?,: Befragung der bundesdeutschen krankenhäuser.

[Hume et al., 2016] Hume, S., Aerts, J., Sarnikar, S., and Huser, V. (2016). Current applications and future directions for the CDISC Operational Data Model standard: A methodological review. *Journal of Biomedical Informatics*, 60:352–362.

[ISO, 2017] ISO (2017). ISO 15836-1: Information and documentation — the dublin core metadata element set — part 1: Core elements. https://www.iso.org/standard/71339.html.

[ISO/IEC, 2013] ISO/IEC (2013). ISO/IEC 11179-3:2013 - Information technology – Metadata registries (MDR) – Part 3: Registry metamodel and basic attributes. https://www.iso.org/standard/50340.html.

[ISO/IEC, 2019] ISO/IEC (2019). ISO/TS 21526:2019 - Health informatics — Metadata repository requirements (MetaRep). https://www.iso.org/standard/71041.html.

[Ivanschitz et al., 2018] Ivanschitz, B. P., Lampoltshammer, T. J., Mireles, V., Revenko, A., Schlarb, S., and Thurnay, L. (2018). A semantic catalogue for the data market Austria. In *CEUR Workshop Proceedings*, volume 2198.

[Jeong et al., 2014] Jeong, S., Kim, H. H., Park, Y. R., Kim, J. H., and Kim, J. H. (2014). Clinical data element Ontology for unified indexing and retrieval of data elements across multiple metadata registries. *Healthcare Informatics Research*, 20(4):295–303.

[Joshi et al., 2014] Joshi, B., Bista, U., and Ghimire, M. (2014). Intelligent Clustering Scheme for Log Data Streams. In Gelbukh, A., editor, *Computational Linguistics and Intelligent Text Processing*, Lecture Notes in Computer Science, pages 454–465, Berlin, Heidelberg. Springer.

[Kadioglu et al., 2018] Kadioglu, D., Breil, B., Knell, C., Lablans, M., Mate, S., Schlue, D., Serve, H., Storf, H., Ückert, F., Wagner, T., Weingardt, P., and Prokosch, H.-U. (2018). Samply.MDR - A Metadata Repository and Its Application in Various Research Networks. *Studies in Health Technology and Informatics*, 253:50–54.

[Kadioglu et al., 2016] Kadioglu, D., Weingardt, P., Ückert, F., et al. (2016). Samply. MDR-Ein Open-Source-Metadaten-Repository. *German Medical Science GMS Publishing House*.

[Kahn et al., 2016] Kahn, M. G., Callahan, T. J., Barnard, J., Bauck, A. E., Brown, J., Davidson, B. N., Estiri, H., Goerg, C., Holve, E., and Johnson, S. G. (2016). A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data. *Egems*, 4(1).

[Kapsner et al., 2021] Kapsner, L. A., Mang, J. M., Mate, S., Seuchter, S. A., Vengadeswaran, A., Bathelt, F., Deppenwiese, N., Kadioglu, D., Kraska, D., and Prokosch, H.-U. (2021). Linking a Consortium-Wide Data Quality Assessment Tool with the MIRACUM Metadata Repository. *Applied Clinical Informatics*, 12(04):826–835.

[Kern et al., 2018] Kern, J., Tas, D., Ulrich, H., Schmidt, E. E., Ingenerf, J., Ückert, F., and Lablans, M. (2018). A Method to use Metadata in legacy Web Applications: The Samply.MDR.Injector. *Stud Health Technol Inform*.

[Kim et al., 2019] Kim, H. H., Park, Y. R., Lee, K. H., Song, Y. S., and Kim, J. H. (2019). Clinical MetaData ontology: A simple classification scheme for data elements of clinical data based on semantics. *BMC Medical Informatics and Decision Making*, 19(1):166.

[Kock-Schoppenhauer et al., 2018a] Kock-Schoppenhauer, A.-K., Bruland, P., and Kadioglu, D. (2018a). Mappathon – A Metadata Mapping Challenge for Secondary Use. Technical report, Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS), 63. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (GMDS).

[Kock-Schoppenhauer et al., 2019] Kock-Schoppenhauer, A.-K., Bruland, P., Kadioglu, D., Brammen, D., Ulrich, H., Kulbe, K., Duhm-Harbeck, P., and Ingenerf, J. (2019). Scientific Challenge in eHealth: MAPPATHON, a Metadata Mapping Challenge. *Studies in Health Technology and Informatics*, 264:1516–1517.

[Kock-Schoppenhauer et al., 2018b] Kock-Schoppenhauer, A.-K., Hartung, L., Ulrich, H., Duhm-Harbeck, P., and Ingenerf, J. (2018b). Practical Extension of Provenance to Healthcare Data Based on the W3C PROV Standard. *Studies in Health Technology and Informatics*, 253:28–32.

[Kock-Schoppenhauer et al., 2018c] Kock-Schoppenhauer, A.-K., Ulrich, H., Wagenzink, S., Duhm-Harbeck, P., Ingenerf, J., Neuhaus, P., Dugas, M., and Bruland, P. (2018c). Compatibility Between Metadata Standards: Import Pipeline of CDISC ODM to the Samply.MDR. *Studies in health technology and informatics*, 247:221–225.

[Kock-Schoppenhauer et al., 2019a] Kock-Schoppenhauer, A.-K., Kroll, B., Lambarki, M., Ulrich, H., Stahl-Toyota, S., Habermann, J., Duhm-Harbeck, P., Ingenerf, J., and Lablans, M. (2019a). One Step Away from Technology but One Step Towards Domain Experts—MDRBridge: A Template-Based ISO 11179-Compliant Metadata Processing Pipeline. *Methods of Information in Medicine*, 58(S 02):e72–e79.

[Kock-Schoppenhauer et al., 2019b] Kock-Schoppenhauer, A.-K., Kroll, B., Lambarki, M., Ulrich, H., Stahl-Toyota, S., Habermann, J. K., Duhm-Harbeck, P., Ingenerf, J., and Lablans, M. (2019b). One Step Away from Technology but One Step Towards Domain Experts-MDRBridge: A Template-Based ISO 11179-Compliant Metadata Processing Pipeline. *Methods of Information in Medicine*, 58(S 02):e72–e79.

[Kock-Schoppenhauer et al., 2021] Kock-Schoppenhauer, A.-K., Schreiweis, B., Ulrich, H., Reimer, N., Wiedekopf, J., Kinast, B., Busch, H., Bergh, B., and Ingenerf, J. (2021). Medical Data Engineering - Theory and Practice. In Leucker, M. and Lamo, Y., editors, *The International Health Data Workshop (HEDA 2021)*, Communications

in Computer and Information Science, page 16 pages (in press), Berlin, Heidelberg. Springer.

[Ku et al., 2014] Ku, H. S., Kim, S., Kim, H. H., Chung, H. J., Park, Y. R., and Kim, J. H. (2014). Dialysisnet: Application for integrating and management data sources of hemodialysis information by continuity of care record. *Healthcare Informatics Research*, 20(2):145–151.

[Kumari and Rath, 2015] Kumari, S. and Rath, S. K. (2015). Performance comparison of soap and rest based web services for enterprise application integration. In *Advances in Computing, Communications and Informatics (ICACCI), 2015 International Conference On*, pages 1656–1660. IEEE.

[Lablans et al., 2015] Lablans, M., Kadioglu, D., Muscholl, M., and Ückert, F. (2015). Exploiting Distributed, Heterogeneous and Sensitive Data Stocks while Maintaining the Owner's Data Sovereignty. *Methods of information in medicine*, 54(4):346–352.

[Li et al., 2013] Li, X., Yan, T., Gao, F., Zhou, L., Yu, J., and Guo, Z. (2013). Design of data management system for seafloor observatory network. In *Proceedings of International Conference on Service Science, ICSS*, pages 147–150. IEEE.

[Li et al., 2012] Li, Z., Wen, J., Zhang, X., Wu, C., Li, Z., and Liu, L. (2012). ClinData Express–a metadata driven clinical research data management system for secondary use of clinical data. *AMIA Annual Symposium Proceedings*, 2012:552–557.

[Liberati et al., 2009] Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., Clarke, M., Devereaux, P. J., Kleijnen, J., and Moher, D. (2009). The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Medicine*, 6(7):e1000100.

[Lightsey, 2001] Lightsey, B. (2001). Systems engineering fundamentals. Technical report, DEFENSE ACQUISITION UNIV FT BELVOIR VA.

[Lin et al., 2019] Lin, J., Ranslam, K., Shi, F., Figurski, M., and Liu, Z. (2019). Data Migration from Operating EMRs to OpenEMR with Mirth Connect. In *ITCH*, pages 288–292.

[Lopprich et al., 2014] Lopprich, M., Jones, J., Meinecke, M. C., Goldschmidt, H., and Knaup, P. (2014). A Reference Data Model of a Metadata Registry Preserving Semantics and Representations of Data Elements. *Studies in Health Technology and Informatics*, 205:368–372.

[Lunesu et al., 2011] Lunesu, M. I., Pani, F. E., and Concas, G. (2011). Using a standards-based approach for a multimedia knowledge-base. In *KMIS 2011 - Proceedings of the International Conference on Knowledge Management and Information Sharing*, pages 87–95. SciTePress - Science and Technology Publications.

[Lyttleton et al., 2011] Lyttleton, O., Treanor, D., Wright, A., and Lewis, P. (2011). Using XML to encode TMA DES metadata. *Journal of Pathology Informatics*, 2(1):40.

[Madhavan et al., 2001] Madhavan, J., Bernstein, P. A., and Rahm, E. (2001). Generic Schema Matching with Cupid. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 49–58, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Majchrzak et al., 2011] Majchrzak, T. A., Jansen, T., and Kuchen, H. (2011). Efficiency evaluation of open source ETL tools. In *Proceedings of the 2011 ACM Symposium on Applied Computing*, pages 287–294. ACM.

[Marques and Ferreira, 2020] Marques, I. C. and Ferreira, J. J. (2020). Digital transformation in the area of health: Systematic review of 45 years of evolution. *Health and Technology*, 10(3):575–586.

[Martínez-Romero et al., 2019] Martínez-Romero, M., O'Connor, M. J., Egyedi, A. L., Willrett, D., Hardi, J., Graybeal, J., and Musen, M. A. (2019). Using association rule mining and ontologies to generate metadata recommendations from multiple biomedical databases. *Database : the journal of biological databases and curation*, 2019.

[Mate et al., 2019a] Mate, S., Bürkle, T., Kapsner, L. A., Toddenroth, D., Kampf, M., Sedlmayr, M., Castellanos, I., Prokosch, H.-U., and Kraus, S. (2019a). A Method for the Graphical Modeling of Relative Temporal Constraints. *Journal of Biomedical Informatics*, page 103314.

[Mate et al., 2019b] Mate, S., Kampf, M., Rödle, W., Kraus, S., Proynova, R., Silander, K., Ebert, L., Lablans, M., Schüttler, C., and Knell, C. (2019b). Pan-European Data

Harmonization for Biobanks in ADOPT BBMRI-ERIC. *Applied clinical informatics*, 10(04):679–692.

[Mate et al., 2015] Mate, S., Köpcke, F., Toddenroth, D., Martin, M., Prokosch, H.-U., Bürkle, T., and Ganslandt, T. (2015). Ontology-Based Data Integration between Clinical and Research Systems. *PLOS ONE*, 10(1):e0116656.

[Maumet et al., 2016] Maumet, C., Auer, T., Bowring, A., Chen, G., Das, S., Flandin, G., Ghosh, S., Glatard, T., Gorgolewski, K. J., Helmer, K. G., Jenkinson, M., Keator, D. B., Nichols, B. N., Poline, J. B., Reynolds, R., Sochat, V., Turner, J., and Nichols, T. E. (2016). Sharing brain mapping statistical results with the neuroimaging data model. *Scientific Data*, 3:160102.

[Miles and Bechhofer, 2009] Miles, A. and Bechhofer, S. (2009). SKOS simple knowledge organization system reference. *W3C recommendation*, 18:W3C.

[Miller et al., 1988] Miller, S. P., Neuman, B. C., Schiller, J. I., and Saltzer, J. H. (1988). Kerberos authentication and authorization system. In *In Project Athena Technical Plan*. Citeseer.

[Milward, 2019] Milward, D. (2019). Model driven data management in healthcare. In *MODELSWARD 2019 - Proceedings of the 7th International Conference on Model-Driven Engineering and Software Development*, pages 107–118. SCITEPRESS - Science and Technology Publications.

[Munkres, 1957] Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38.

[Nadkarni, 2011] Nadkarni, P. M. (2011). *Metadata-Driven Software Systems in Biomedicine: Designing Systems That Can Adapt to Changing Knowledge*. Springer Science & Business Media.

[Nadkarni and Marenco, 2013] Nadkarni, P. M. and Marenco, L. N. (2013). *Data Integration: An Overview*. Elsevier.

[Ngouongo et al., 2013] Ngouongo, S. M., Löbe, M., and Stausberg, J. (2013). The ISO/IEC 11179 norm for metadata registries: Does it cover healthcare standards in empirical research? *Journal of Biomedical Informatics*, 46(2):318–327.

[Ngouongo and Stausberg, 2011] Ngouongo, S. M. and Stausberg, J. (2011). Integration of classifications and terminologies in metadata registries based on ISO/IEC 11179. *Studies in Health Technology and Informatics*, 169:744–748.

[NISO, 2017] NISO (2017). Understanding Metadata: What is Metadata, and What is it For?: A Primer — NISO website. https://www.niso.org/publications/understanding-metadata-2017.

[O'hara, 2019] O'hara, K. (2019). Data Trusts: Ethics, Architecture and Governance for Trustworthy Data Stewardship. https://eprints.soton.ac.uk/428276/.

[Ott et al., 2019] Ott, S., Rinner, C., and Duftschmid, G. (2019). Expressing Patient Selection Criteria Based on HL7 V3 Templates Within the Open-Source Tool ART-DECOR. *Studies in Health Technology and Informatics*, 260:226–233.

[Papež and Mouček, 2017] Papež, V. and Mouček, R. (2017). Applying an archetype-based approach to electroencephalography/event-related potential experiments in the EEGBase resource. *Frontiers in Neuroinformatics*, 11.

[Park and Tosaka, 2010] Park, J. R. and Tosaka, Y. (2010). Metadata creation practices in digital repositories and collections: Schemata, selection criteria, and interoperability. *Information Technology and Libraries*, 29(3):104–116.

[Park and Kim, 2010] Park, Y. R. and Kim, J. H. (2010). Achieving interoperability for metadata registries using comparative object modeling. In *Studies in Health Technology and Informatics*, volume 160, pages 1136–1139.

[Park et al., 2013] Park, Y. R., Yoon, Y. J., Kim, H. H., and Kim, J. H. (2013). Establishing semantic interoperability of biomedical metadata registries using extended semantic relationships. In Lehmann, CU and Ammenwerth, E and Nohr, C, editor, *Studies in Health Technology and Informatics*, volume 192 of *Studies in Health Technology and Informatics*, pages 618–621.

[Pathak et al., 2011] Pathak, J., Wang, J., Kashyap, S., Basford, M., Li, R., Masys, D. R., and Chute, C. G. (2011). Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: The eMERGE Network experience. *Journal of the American Medical Informatics Association*, 18(4):376–386.

[Qamar et al., 2007] Qamar, R., Kola, J. S., and Rector, A. L. (2007). Unambiguous data modeling to ensure higher accuracy term binding to clinical terminologies. In *AMIA Annual Symposium Proceedings*, volume 2007, page 608. American Medical Informatics Association.

[Rebaï et al., 2015] Rebaï, R. Z., Mnif, F., Zayani, C. A., and Amous, I. (2015). Adaptive global schema generation from heterogeneous metadata schemas. In *Procedia Computer Science*, volume 60, pages 197–205.

[Riepenhausen et al., 2019] Riepenhausen, S., Varghese, J., Neuhaus, P., Storck, M., Meidt, A., Hegselmann, S., and Dugas, M. (2019). Portal of Medical Data Models: Status 2018. In *EFMI-STC*, pages 239–240.

[Rodrigues et al., 2019] Rodrigues, J., Castro, J. A., da Silva, J. R., and Ribeiro, C. (2019). Hands-On Data Publishing with Researchers: Five Experiments with Metadata in Multiple Domains. *Communications in Computer and Information Science*, 988:274–288.

[Sass et al., 2020] Sass, J., Bartschke, A., Lehne, M., Essenwanger, A., Rinaldi, E., Rudolph, S., Heitmann, K. U., Vehreschild, J. J., von Kalle, C., and Thun, S. (2020). The German Corona Consensus Dataset (GECCO): A standardized dataset for COVID-19 research in university medicine and beyond. *BMC Medical Informatics and Decision Making*, 20(1):341.

[Schmidt et al., 2021] Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., Stausberg, J., Damerow, S., Huebner, M., Schmidt, B., Sauerbrei, W., and Richter, A. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, 21:63.

[Semler et al., 2018] Semler, S. C., Wissing, F., and Heyder, R. (2018). German Medical Informatics Initiative. *Methods of Information in Medicine*, 57(S 1):e50–e56.

[Smits et al., 2015] Smits, M., Kramer, E., Harthoorn, M., and Cornet, R. (2015). A comparison of two Detailed Clinical Model representations: FHIR and CDA. *European Journal of Biomedical Informatics*, 11(2).

[Solbrig, 2000] Solbrig, H. R. (2000). Metadata and the reintegration of clinical information: ISO 11179. *MD computing: computers in medical practice*, 17(3):25.

[Song et al., 2014] Song, T. M., Park, H. A., and Jin, D. L. (2014). Development of health information search engine based on metadata and ontology. *Healthcare Informatics Research*, 20(2):88–98.

[Späth and Grimson, 2011] Späth, M. B. and Grimson, J. (2011). Applying the archetype approach to the database of a biobank information management system. *International Journal of Medical Informatics*, 80(3):205–226.

[Specka et al., 2019] Specka, X., Gärtner, P., Hoffmann, C., Svoboda, N., Stecker, M., Einspanier, U., Senkler, K., Zoarder, M. A., and Heinrich, U. (2019). The BonaRes metadata schema for geospatial soil-agricultural research data – Merging INSPIRE and DataCite metadata schemes. *Computers and Geosciences*, 132:33–41.

[Stausberg and Harkener, 2019] Stausberg, J. and Harkener, S. (2019). Bridging Documentation and Metadata Standards: Experiences from a Funding Initiative for Registries. *Studies in health technology and informatics*, 264:1046–1050.

[Stöhr et al., 2021] Stöhr, M. R., Günther, A., and Majeed, R. W. (2021). ISO 21526 Conform Metadata Editor for FAIR Unicode SKOS Thesauri. In *German Medical Data Sciences: Bringing Data to Life*, pages 94–100. IOS Press.

[Trani et al., 2018] Trani, L., Atkinson, M., Bailo, D., Paciello, R., and Filgueira, R. (2018). Establishing Core Concepts for Information-Powered Collaborations. *Future Generation Computer Systems*, 89:421–437.

[Ulrich et al., 2020a] Ulrich, H., Germer, S., Kock-Schoppenhauer, A.-K., Kern, J., Lablans, M., and Ingenerf, J. (2020a). A Smart Mapping Editor for Standardised Data Transformation. *Studies in Health Technology and Informatics*, 270:1185–1186.

[Ulrich et al., 2019a] Ulrich, H., Kern, J., Kock-Schoppenhauer, A.-K., Lablans, M., and Ingenerf, J. (2019a). Towards a Federation of Metadata Repositories: Addressing Technical Interoperability. *Studies in Health Technology and Informatics*, 253:55–59.

[Ulrich et al., 2019b] Ulrich, H., Kern, J., Tas, D., Kock-Schoppenhauer, A.-K., Ückert, F., Ingenerf, J., and Lablans, M. (2019b). QL4MDR: A GraphQL query language

for ISO 11179-based metadata repositories. *BMC Medical Informatics and Decision Making*, 19(1):45. https://doi.org/10.1186/s12911-019-0794-z.

[Ulrich et al., 2016] Ulrich, H., Kock, A. K., Duhm-Harbeck, P., Habermann, J. K., and Ingenerf, J. (2016). Metadata repository for improved data sharing and reuse based on HL7 FHIR. *Studies in Health Technology and Informatics*, 228:162–166.

[Ulrich et al., 2017] Ulrich, H., Kock-Schoppenhauer, A.-K., Andersen, B., and Ingenerf, J. (2017). Analysis of Annotated Data Models for Improving Data Quality. *Studies in Health Technology and Informatics*, 243:190–194.

[Ulrich et al., 2022] Ulrich, H., Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., and Ingenerf, J. (2022). Understanding the Nature of Metadata – A Systematic Review. *Journal of Medical Internet Research*. http://dx.doi.org/10.2196/25440.

[Ulrich et al., 2020b] Ulrich, H., Kock-Schoppenhauer, A.-K., Drenkhahn, C., Löbe, M., and Ingenerf, J. (2020b). Analysis of ISO/TS 21526 Towards the Extension of a Standardized Query API. *Studies in Health Technology and Informatics*, 275:202–206.

[Ulrich et al., 2018a] Ulrich, H., Kock-Schoppenhauer, A.-K., Duhm-Harbeck, P., and Ingenerf, J. (2018a). Modelling a Metadata Repository in a Graph Database using Neo4j.

[Ulrich et al., 2018b] Ulrich, H., Kock-Schoppenhauer, A.-K., Duhm-Harbeck, P., and Ingenerf, J. (2018b). Using Graph Tools on Metadata Repositories. *Studies in Health Technology and Informatics*, 253:55–59.

[Urban, 2014] Urban, R. J. (2014). The 1:1 Principle in the age of Linked Data. In *Proceedings of the International Conference on Dublin Core and Metadata Applications*, pages 119–128.

[Van de Sompel et al., 2004] Van de Sompel, H., Nelson, M. L., Lagoze, C., and Warner, S. (2004). Resource harvesting within the OAI-PMH framework. *D-lib magazine*, 10(12).

[Varghese et al., 2018a] Varghese, J., Fujarski, M., Hegselmann, S., Neuhaus, P., and Dugas, M. (2018a). Cdegenerator: An online platform to learn from existing data models to build model registries. *Clinical Epidemiology*, 10:961–970.

[Varghese et al., 2018b] Varghese, J., Sandmann, S., and Dugas, M. (2018b). Web-based information infrastructure increases the interrater reliability of medical coders: Quasi-experimental study. *Journal of medical Internet research*, 20(10):e274.

[Vireq, 2021] Vireq (2021). Mirth Connect » We healthcare. https://www.vireq.com/produkte/mirth-connect/, Zugriffsdatum: 26. Oktober 2021.

[Vos et al., 2012] Vos, R. A., Balhoff, J. P., Caravas, J. A., Holder, M. T., Lapp, H., Maddison, W. P., Midford, P. E., Priyam, A., Sukumaran, J., Xia, X., and Stoltzfus, A. (2012). NeXML: Rich, extensible, and verifiable representation of comparative data and metadata. *Systematic Biology*, 61(4):675–689.

[W3C, 2001] W3C (2001). Metadata at W3C. https://www.w3.org/Metadata/.

[Warzel et al., 2003] Warzel, D. B., Andonyadis, C., McCurry, B., Chilukuri, R., Ishmukhamedov, S., and Covitz, P. (2003). Common Data Element (CDE) Management and Deployment in Clinical Trials. *AMIA Annual Symposium Proceedings*, 2003:1048.

[Webber and Robinson, 2018] Webber, J. and Robinson, I. (2018). *A Programmatic Introduction to Neo4j*. Addison-Wesley Professional.

[Wilkinson et al., 2016] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3:160018.

[Woodley, 2008] Woodley, M. S. (2008). Crosswalks, metadata harvesting, federated searching, metasearching: Using metadata to connect users and information. In *Introduction to Metadata*. Getty Research Institute.

[Wulff et al., 2018] Wulff, A., Haarbrandt, B., and Marschollek, M. (2018). Clinical Knowledge Governance Framework for Nationwide Data Infrastructure Projects. In *eHealth*, pages 196–203.

[Yuliant and Karna, 2017] Yuliant, R. and Karna, N. (2017). Knowledge sharing filtering on OAI-PMH. In *2016 International Conference on Information Technology Systems and Innovation, ICITSI 2016 - Proceedings*.

[Zozus and Bonner, 2017] Zozus, M. N. and Bonner, J. (2017). Towards data value-level metadata for clinical studies. In Lau, F and BartleClar, J and Bliss, G and Borycki, E and Courtney, K and Kuo, A, editor, *Studies in Health Technology and Informatics*, volume 234 of *Studies in Health Technology and Informatics*, pages 418–423.

Eigene Publikationen

Erst und Letztautorenschaften

[1] **Ulrich, H.**, Kock-Schoppenhauer, A.-K., Deppenwiese, N., Gött, R., Kern, J., Lablans, M., Majeed, R. W., Stöhr, M. R., Stausberg, J., Varghese, J., Dugas, M., and Ingenerf, J. (Accepted). Understanding the Nature of Metadata – A Systematic Review, Journal of Medical Internet Research, 25440. https://doi.org/10.2196/25440

[2] **Ulrich, H.**, Behrend, P., Wiedekopf, J., Drenkhahn, C., Kock-Schoppenhauer, A.-K., and Ingenerf, J. (2021). Hands on the Medical Informatics Initiative Core Data Set—Lessons Learned from Converting the MIMIC-IV. Studies in Health Technology and Informatics, 283, 119–126. https://doi.org/10.3233/SHTI210549

[3] **Ulrich, H.**, Germer, S., Kock-Schoppenhauer, A.-K., Kern, J., Lablans, M., and Ingenerf, J. (2020). A Smart Mapping Editor for Standardised Data Transformation. Studies in Health Technology and Informatics, 270, 1185–1186. https://doi.org/10.3233/SHTI200354

[4] **Ulrich, H.**, Kock-Schoppenhauer, A.-K., Drenkhahn, C., Löbe, M., and Ingenerf, J. (2020). Analysis of ISO/TS 21526 Towards the Extension of a Standardized Query API. Studies in Health Technology and Informatics, 275, 202–206. https://doi.org/10.3233/SHTI200723

[5] **Ulrich, H.**, Kern, J., Tas, D., Kock-Schoppenhauer, A.-K., Ückert, F., Ingenerf, J., and Lablans, M. (2019). QL4MDR: A GraphQL query language for ISO 11179-based metadata repositories. BMC Medical Informatics and Decision Making, 19(1), 45. https://doi.org/10.1186/s12911-019-0794-z

[6] **Ulrich, H.**, Kern, J., Kock-Schoppenhauer, A.-K., Lablans, M., and Ingenerf, J. (2019). Towards a Federation of Metadata Repositories: Addressing Technical Inte-

roperability. Studies in Health Technology and Informatics, 253, 55–59. https://doi.org/10.3233/SHTI190808

[7] Deppenwiese, N., Duhm-Harbeck, P., Ingenerf, J., and **Ulrich, H**. (2019). MDRCupid: A Configurable Metadata Matching Toolbox. Studies in Health Technology and Informatics, 264, 88–92. https://doi.org/10.3233/SHTI190189

[8] **Ulrich, H.**, Kock-Schoppenhauer, A.-K., Duhm-Harbeck, P., and Ingenerf, J. (2018). Using Graph Tools on Metadata Repositories. Studies in Health Technology and Informatics, 253, 55–59. Ausgezeichnet mit dem **GMDS Best Paper Award 2018** https://doi.org/10.3233/978-1-61499-896-9-55

[9] **Ulrich, H.**, Kock-Schoppenhauer, A.-K., Andersen, B., and Ingenerf, J. (2017). Analysis of Annotated Data Models for Improving Data Quality. Studies in Health Technology and Informatics, 243, 190–194. https://doi.org/10.3233/978-1-61499-808-2-190

[10] **Ulrich, H.**, Kock, A. K., Duhm-Harbeck, P., Habermann, J. K., and Ingenerf, J. (2016). Metadata repository for improved data sharing and reuse based on HL7 FHIR. Studies in Health Technology and Informatics, 228, 162–166. https://doi.org/10.3233/978-1-61499-678-1-162

Koautorenschaften

[1] Kock-Schoppenhauer, A.-K., Schreiweis, B., **Ulrich, H.**, Reimer, N., Wiedekopf, J., Kinast, B., Busch, H., Bergh, B., and Ingenerf, J. (2021). Medical Data Engineering—Theory and Practice. In M. Leucker and Y. Lamo (Editor.), The International Health Data Workshop (HEDA 2021), pp. 269-284. Springer. https://doi.org/10.1007/978-3-030-87657-9_21

[2] Wiedekopf, J., Drenkhahn, C., **Ulrich, H.**, Kock-Schoppenhauer, A.-K., and Ingenerf, J. (2021). Providing ART-DECOR ValueSets via FHIR Terminology Servers – A Technical Report. Studies in Health Technology and Informatics. 283, 127 - 135. https://doi.org/10.3233/SHTI210550

[3] Wiedekopf, J., **Ulrich, H.**, Essenwanger, A., Kiel, A., Kock-Schoppenhauer, A.-K., and Ingenerf, J. (2021). Desiderata for a Synthetic Clinical Data Generator. Studies

in Health Technology and Informatics, 281, 68–72.
https://doi.org/10.3233/SHTI210122

[4] Banach, **A., Ulrich**, H., Kroll, B., Kiel, A., Ingenerf, J., and Kock-Schoppenhauer, A.-K. (2021). APERITIF - Automatic Patient Recruiting for Clinical Trials Based on HL7 FHIR. Studies in Health Technology and Informatics, 281, 58–62.
https://doi.org/10.3233/SHTI210120

[5] Reimer, N., **Ulrich, H.**, Busch, H., Kock-Schoppenhauer, A.-K., and Ingenerf, J. (2021). OpenEHR Mapper – A Tool to Fuse Clinical and Genomic Data Using the openEHR Standard. Studies in Health Technology and Informatics. 278, 86 - 93.
https://doi.org/10.3233/SHTI210055

[6] Riech, K. P., **Ulrich, H.**, Ingenerf, J., and Andersen, B. (2021). Mapping of medical device data from ISO/IEEE 11073-10207 to HL7 FHIR. GMS Medizinische Informatik, Biometrie und Epidemiologie, 17(2), Doc08. https://doi.org/10.3205/mibe000222

[7] Banach, A., **Ulrich, H.**, Kroll, B., Kiel, A., Ingenerf, J., and Kock-Schoppenhauer, A.-K. (2021). Benefits of MII Core Dataset and HL7 FHIR-Based Tooling for Automated Recruiting Purposes. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 125. https://doi.org/10.3205/21gmds027

[8] Drenkhahn, C., **Ulrich, H.**, and Ingenerf, J. (2021). An ontology-based tool to visually compare LOINC subsets. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 233. https://doi.org/10.3205/20gmds195

[9] Banach, A., Kock-Schoppenhauer, A.-K., **Ulrich, H.**, and Ingenerf, J. (2021). Evaluating Automated Methods for Metadata Quality in Healthcare. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 221.
https://doi.org/10.3205/20gmds192

[10] Drenkhahn, C., Burmester, S., Ballout, S., **Ulrich, H.**, Wiedekopf, J., and Ingenerf, J. (2020). Using FHIR terminology services-based tools for mapping of local microbiological pathogen terms to SNOMED CT at a German university hospital. 16. DVMD-Fachtagung, Leipzig, 25. bis 26. Februar 2021. https://dvmd.de/wp-content/uploads/2020/11/A-148.pdf

[11] Kock-Schoppenhauer, A.-K., Kroll, B., Lambarki, M., **Ulrich, H.**, Stahl-Toyota, S., Habermann, J. K., Duhm-Harbeck, P., Ingenerf, J., and Lablans, M. (2019). One Step Away from Technology but One Step Towards Domain Experts—MDRBridge: A Template-Based ISO 11179-Compliant Metadata Processing Pipeline. Methods of Information in Medicine, 58(S 02), e72–e79.
https://doi.org/10.1055/s-0039-3399579

[12] Kock-Schoppenhauer, A.-K., Bruland, P., Kadioglu, D., Brammen, D., **Ulrich, H.**, Kulbe, K., Duhm-Harbeck, P., and Ingenerf, J. (2019). Scientific Challenge in eHealth: MAPPATHON, a Metadata Mapping Challenge. Studies in Health Technology and Informatics, 264, 1516–1517. https://doi.org/10.3233/SHTI190512

[13] Deppenwiese, N., Kock-Schoppenhauer, A.-K., **Ulrich, H.**, Duhm-Harbeck, P., and Ingenerf, J. (2019). Automatic Evaluation of Metadata Quality in ISO 11179-3 Conformant Healthcare Metadata Repositories. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 86. https://doi.org/10.3205/19gmds036

[14] Kock-Schoppenhauer, A.-K., Wagenzink, S., Deppenwiese, N., **Ulrich, H.**, Simon, F., Meyer-Saraei, R., Heeger, C.-H., Graf, T., Tilz, R., Ingenerf, J., and Duhm-Harbeck, P. (2019). Potential Secondary Use of Medical Data in Monocentric In-House Clinical Trials. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 93. https://doi.org/10.3205/19gmds025

[15] Kern, J., Tas, D., **Ulrich, H.**, Schmidt, E. E., Ingenerf, J., Ückert, F., and Lablans, M. (2018). A Method to use Metadata in legacy Web Applications: The Samply.MDR.Injector. Studies in Health Technology and Informatics, 253, 45-49
https://doi.org/10.3233/978-1-61499-896-9-45

[16] Kock-Schoppenhauer, A.-K., Hartung, L., **Ulrich, H.**, Duhm-Harbeck, P., and Ingenerf, J. (2018). Practical Extension of Provenance to Healthcare Data Based on the W3C PROV Standard. Studies in Health Technology and Informatics, 253, 28–32.
https://doi.org/10.3233/978-1-61499-896-9-28

[17] Kock-Schoppenhauer, A. K., **Ulrich, H**., Wagen-Zink, S., Duhm-Harbeck, P., Ingenerf, J., Neuhaus, P., Dugas, M., and Bruland, P. (2018). Compatibility Between Metadata Standards: Import Pipeline of CDISC ODM to the Samply.MDR. Studies

in Health Technology and Informatics, 247, 221–225. https://doi.org/10.3233/978-1-61499-852-5-221

[18] Andersen, B., Kasparick, M., **Ulrich, H.**, Franke, S., Schlamelcher, J., Rockstroh, M., and Ingenerf, J. (2018). Connecting the clinical IT infrastructure to a service-oriented architecture of medical devices. Biomedizinische Technik, 63(1), 57–68. https://doi.org/10.1515/bmt-2017-0021

[19] Deppenwiese, N., **Ulrich, H.**, Wrage, J.-H., Kock, A.-K., and Ingenerf, J. (2017). Entwicklung eines Tools zum Konvertieren von HL7 FHIR Questionnaires in das MOLGENIS EMX Format. GMS Medizinische Informatik, Biometrie und Epidemiologie, DocAbstr. 207. https://doi.org/10.3205/17gmds148

[20] Kock-Schoppenhauer, A.-K., Kamann, C., **Ulrich, H.**, Duhm-Harbeck, P., and Ingenerf, J. (2017). Linked Data Applications Through Ontology Based Data Access in Clinical Research. Studies in Health Technology and Informatics, 235, 131–135. https://doi.org/10.3233/978-1-61499-753-5-131

[21] Andersen, B., **Ulrich, H.**, Schlichting, S., Golatowski, F., Timmermann, D., Ingenerf, J., and Kasparick, M. (2016). Point-of-Care Medical Devices and Systems Interoperability: A Mapping of ICE and FHIR. 2016 IEEE Conference on Standards for Communications and Networking (CSCN 2016), 1–5. https://doi.org/10.1109/CSCN.2016.7785165

Wettbewerbsteilnahme

Ulrich H., Deppenwiese N.: Erster Platz beim Mappathon - A Metadata Mapping Challenge im Rahmen der GMDS 2018,