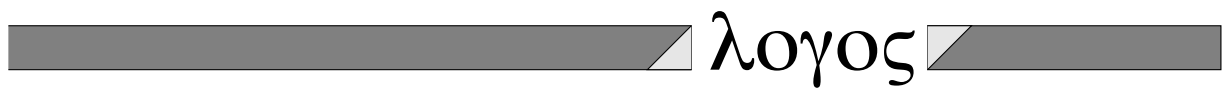


Frank Friedhelm Kramer

Ein allgemeiner Ansatz zur Metadaten-Verwaltung



Dissertation zur Erlangung des akademischen Grades
Doktor der Ingenieurwissenschaften (Dr.-Ing.)
der Technischen Fakultät der Christian-Albrechts-Universität zu Kiel,
eingereicht im Jahr 2018

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der
Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind
im Internet über <http://dnb.d-nb.de> abrufbar.

© Copyright Logos Verlag Berlin GmbH 2019
Alle Rechte vorbehalten.

ISBN 978-3-8325-4788-2

Logos Verlag Berlin GmbH
Comeniushof, Gubener Str. 47,
10243 Berlin
Tel.: +49 (0)30 42 85 10 90
Fax: +49 (0)30 42 85 10 92
INTERNET: <http://www.logos-verlag.de>

1 Einleitung

Ein in den letzten Jahren eher vernachlässigtes Gebiet des Datenmanagements ist die Verwaltung von Metadaten. Dabei war das Thema in den Anfängen der Datenbankmanagementsysteme (DBMS) durchaus relevant. Um einem DBMS zu ermöglichen, angelegte Strukturen zu erfassen und zu verwalten, wurde in den 1970er Jahren ein großer Aufwand in die Entwicklung des Data Dictionarys gesteckt. Dieses sollte die Metadaten für die angelegten Strukturen in einem Datenbankschema, wie die vorhandenen Objekte, Relationen und Schlüssel, erfassen und diese dem DBMS für seine Arbeit zur Verfügung stellen. Daraus resultiert bis heute ein Nutzen von Metadaten. Durch die Bekanntgabe der Strukturen, die in einem DBMS vorliegen, können die relevanten Funktionen zum Umgang mit den Daten, wie Einfüge- und Löschoperationen, generiert werden. Dies ist aber nur möglich, wenn die rigide Form des Data Dictionarys eingehalten wird. Das Ergebnis dieser Entwicklung ist heute in jedem gängigen DBMS zu finden und auch nicht mehr wegzudenken. Das Data Dictionary hat aber auch einen klaren Fokus. Es ist für das DBMS und dessen Operatoren entwickelt worden. Andere Nutzer profitieren nicht von dieser Form der Metadaten-Verwaltung. Dabei ergibt die Verwendung von Metadaten gerade bei unbedarften Nutzern durchaus Sinn. Durch die immer größer werdende Menge und Größe an Daten wird es heute für heterogene Nutzergruppen immer schwieriger, Informationen über die vorliegenden Daten nur aus diesen selbst zu erhalten und vernünftig zu bewerten. Das Problem wird auch noch dadurch erschwert, dass jede Nutzergruppe Daten unter unterschiedlichen Gesichtspunkten betrachtet, diese aber aufgrund der flach geklopften Strukturen in modernen, relationalen Datenbankmanagementsystemen nicht richtig erkennen kann.

Diese Probleme sind in der Praxis immer häufiger zu finden. So beschäftigen sich verschiedene Forschungsgruppen im Sonderforschungsbereich 1266 an der CAU [8] mit dem Transformationsprozess prähistorischer und archaischer Epochen. Schon innerhalb der Forschungsgruppe der Archäologen hat sich gezeigt, dass die beteiligten Nutzergruppen unterschiedliche Sichtweisen auf die für alle vorhandenen Daten haben. Ein einfaches Beispiel liegt in der Unterscheidung zwischen Dorf und Siedlung. Während Forscher, die sich mit dem Neolithikum befassen, über Dörfer reden, reden Forscher, die sich mit der Bronzezeit beschäftigen, über Siedlungen. Eigentlich meinen aber beide die gleichen Objekte und Daten. Somit wäre hier eine Integration der verschiedenen Gesichtspunkte für die gemeinsame Nutzung der vorhandenen Daten nötig. Dies würde den Forschern eine Chance geben, die gesammelten Daten zu verstehen und sich interdisziplinär darüber auszutauschen. Ein weiterer Fall aus der Praxis ist das Projekt „Future Ocean“ [4], das ebenfalls an der CAU durchgeführt wird. So wurde in Gesprächen mit den Mitarbeitern 2013 deutlich, dass die einzelnen Bereiche des „Future Ocean“ unterschiedliche Sichtweisen auf die Daten und somit auch auf die benötigten Metadaten haben. Als Beispiel wurde eine Forschungsfahrt aufgeführt, deren Verlauf über den Ozean für den Forscher an Bord eines Schiffes Metadaten für die Messproben sei, für den Kapitän des Schiffes hingegen Daten, die zur Navigation des Schiffes dienen. Hinzu kommt, dass die erhobenen Daten in unterschiedliche, heterogene Datencluster gespielt werden müssen, deren Format und Aufbau sich im Allgemeinen unterscheiden. Die momentane Lösung sind dabei Policies, die den

einzelnen Nutzergruppen mitgegeben werden, wobei es hier nötig wäre, zu erklären, warum eine Policy auf welche Weise notwendig ist. Ein letzter Praxisfall ist das Programmieren von großen Systemen oder Applikationen, an denen Teams über die gesamte Welt verteilt arbeiten. Die Teams sind meist räumlich getrennt und verständigen sich über das Internet miteinander. Dies stellt die nächsthöhere Ebene zum Programming-in-the-large¹ dar und soll im Folgenden als Programming-in-the-web bezeichnet werden. Das große Problem liegt darin, dass die einzelnen Nutzer innerhalb eines Projektes praktisch ‚on the fly‘ verstehen müssen, was die anderen Nutzer an dem System gemacht haben. Dies resultiert in einer Art von Kommunikationsproblem, da die verschiedenen heterogenen Nutzer sich austauschen müssen, um die Arbeiten und Veränderungen so zu kommunizieren, dass diese schnell und einfach durch die anderen Nutzer verstanden werden.

1.1 Der Mehrwert durch explizite Metadaten

Eine wesentliche Lösung für dieses Bündel von Problemfällen ist die Generierung einer vernünftigen Übersicht über die Datenbestände, die in der entsprechenden Anwendung vorhanden sind. Diese Übersicht muss zusätzlich um Informationen angereichert werden, die beschreiben, was sich die Nutzer bei den einzelnen Datenbeständen gedacht haben, warum also einzelne Datenbestände in ihrer momentanen Form in dem System vorhanden sind. Die Verwendung von Metadaten ist eine denkbare Lösung. Metadaten besitzen das Potential durch ihre Aussagekraft viel zur Interpretation von Daten beitragen zu können. Eine Lösung über Metadaten sollte dabei hinreichend universell sein, so dass sie verschiedene Situationen und Problemfälle abbilden kann. Das Data-Dictionary-Management aus den DBMS kann dafür eine Ausrichtung sein, reicht aber als alleinige, universelle Lösung aufgrund der diskutierten Schwächen nicht aus. So ist die Bedeutung, welche Daten überhaupt als Metadaten aufgefasst werden können, durch mehrere Faktoren, wie das zu Grunde liegende Szenario, dem Kontext und den Nutzergruppen bestimmt. Daher sollte die Lösung ebenfalls eine hinreichend universelle Möglichkeit zur Beschreibung der Bedeutung von Metadaten bereitstellen, welche die wesentlichen Faktoren miteinbezieht und für eine Anwendung eine für alle Nutzergruppen verständliche, gemeinsame Beschreibung generiert. Mit der Bedeutung von Metadaten alleine kann aber nicht gearbeitet werden. Daher wird eine Struktur benötigt, welche mit der Beschreibung der Metadaten als Grundlage, eine Verwaltung von Metadaten erlaubt. Da die Grundlage dieser Struktur eine universelle Beschreibung ist, sollte die Struktur selbst ebenfalls hinreichend universell sein. Zu diesem Zweck könnten ein Stereotyp bzw. mehrere Stereotype zum Einsatz kommen, welche zwar den generellen Kern einer jeden Metadaten-Verwaltung bilden, aber für jede Anwendung spezialisiert werden, um den speziellen Anforderungen jeder Anwendung gerecht zu werden. Dies würde zu einer einfacheren und einheitlicheren Entwicklung von Strukturen für die Verwaltung von Metadaten führen. Hinzu kommt, dass anders als in einem DBMS, die Metadaten nicht komplett von einem System selbst generiert und in die Strukturen eingepflegt werden. Sie stammen aus

¹ Für die Beschreibung des Begriffs Programming-in-the-large siehe De Remer und Kron in [39].

verschiedenen, heterogenen Quellen und werden durch die verschiedenen Nutzergruppen in die Anwendung importiert, eingetragen oder generiert. Daher sollte eine mögliche Lösung zur Metadaten-Verwaltung Methoden bieten, in die spezialisierte Stereotype Metadaten einzupflegen, die an die Sichtweisen der jeweiligen Nutzer angepasst sind. Sind Metadaten in einem System zu finden, so müssen diese für die Nutzergruppen aufbereitet und so zur Verfügung gestellt werden, dass jede Gruppe ihre Metadaten passend zu ihrem Szenario finden kann. Hier könnte auf eine neue Art von Data Dictionary zurückgegangen werden, das als Repository für Metadaten zu nutzen sein könnte. Die einzelnen Tabellen im Dictionary müssten dabei unter Berücksichtigung der Nutzerinteressen und Szenarien über die im Schema zu findenden Metadaten generiert und aufbereitet werden. Zuletzt werden die Metadaten nicht allein von Nutzern verarbeitet. Die Metadaten werden auch zusammengestellt und an andere Quellen exportiert, damit mit diesen gearbeitet werden kann. So sollte eine Lösung zur Metadaten-Verwaltung, ähnlich zum Import, auch Methoden für den Export von Metadaten bereitstellen. Diese müsste so universell sein, dass die gleichen Metadaten ohne Aufwand an verschiedene, heterogene Zielsysteme geliefert werden könnten.

Die Existenz einer solchen Lösung würde zu einer klaren Verbesserung in den Problemfällen führen. Die Probleme, die durch die Sichtweise auf Daten und Metadaten entstehen, werden durch die Beschreibung von Metadaten unter Einbeziehung relevanter Faktoren, wie dem Kontext der Anwendung, den Szenarien und den relevanten Nutzergruppen, beseitigt. Hinzu kommt, dass eine gemeinsame Grundlage für das Verständnis der Metadaten gelegt wird, welche sich wiederum auf das Verständnis der eigentlichen Daten auswirkt, da bekannt ist, welche Metadaten für diese Daten existieren und verwendet werden können. Dieses Verständnis wirkt sich dann auch auf die Kommunikation untereinander aus, da alle beteiligten Nutzergruppen dieselbe Beschreibung vorliegen haben. Durch die Einführung eines Metadatenschemas als Stereotyp wird ermöglicht, für verschiedene, heterogene Gruppen und Szenarien ähnliche Schemata zu entwickeln, die einen gemeinsamen Kern besitzen. Dadurch wird es möglich, aus einem gemeinsamen Schema unterschiedliche Sichtweisen abzuleiten, die dann den Nutzergruppen, bedingt durch ihre eigene Sichtweise aufbereitet zur Verfügung gestellt werden. Da der Kern des Stereotyps dabei gleich ist, vereinfacht sich nicht nur der Datenaustausch zwischen den einzelnen, spezialisierten Schemata, sondern auch die Kommunikation über die Metadaten, die in diesen Schemata vorhanden sind. Durch das Bereitstellen von Methoden, die das Importieren und Exportieren von Metadaten in die Schemata erlauben, können die Probleme, die durch heterogene Quellen und Ziele entstehen, gelöst werden. Der Import und Export selbst kann durch angepasste Methoden für die jeweilige Quelle verwendet werden und gleichzeitig ein gemeinsames Schema nutzen. Dies erleichtert beim Import die Berücksichtigung der unterschiedlichen Sichtweisen, da jeder Sichtweise eine eigene Schnittstelle für den Import ihrer Metadaten zur Verfügung gestellt werden kann. Die Sichtweise kann dann durch einfache Extract-Transform-Load-Prozesse auf die Gesamtstruktur abgebildet werden. Für den Export ermöglicht die Lösung eine je nach Zielsystem angepasste Struktur, die beispielsweise das Problem der heterogenen Quellen beim Projekt „Future Ocean“ lösen würde. Für das Retrieval von Metadaten durch die Nutzer bildet das automatisch generierte Data Dictionary die Möglichkeit, aufbereitete Metadaten zu den passenden Daten im Sys-

tem zu erhalten und auch verteilten Nutzern über das Web zur Verfügung zu stellen. Da die einzelnen Tabellen mit Sicht auf die Nutzergruppen und Szenarien generiert werden, können Metadaten zielgerichtet und nach der gültigen Beschreibung bereitgestellt werden, was wiederum das Sichtweisenproblem der einzelnen Bereiche löst.

Eine universelle Metadaten-Verwaltung hat Prinzipien und Zielsetzungen zu erfüllen, die je nach Anwendung und Nutzergruppe abgeleitet werden müssen. Aus den Prinzipien sollte sich insgesamt ergeben, was es für die Nutzer bedeutet, eine vernünftige Metadaten-Verwaltung zu besitzen. Die Zielsetzung wiederum beschreibt, was für echte Ziele mit der Metadaten-Verwaltung in einer Anwendung für die Nutzergruppen erfüllt werden sollen. Eine Möglichkeit, diese Kriterien zu bestimmen, liegt in der Nutzung von Qualitätskriterien, wie sie von Thalheim und Nissen in [132] für Modelle definiert wurden. Dort wird die Qualität eines Modells über explizite Qualitätskriterien, wie Nachhaltigkeit, Dauerhaftigkeit, sowie einer abstrakten Liste von Parametern, die sich in interne, externe und Qualität-in-Benutzung aufgliedert, gemessen. Damit ist es für ein Modell auch möglich zu sagen, wofür es in welcher Qualität zu nutzen ist und wie gut sich Ziele des Modells erreichen lassen. Wird dieses Vorgehen auf die Prinzipien und die Zielsetzung angewendet, so kann es auch mittels Qualitätskriterien ausgedrückt werden. Die relevanten Qualitätskriterien werden dazu den entsprechenden Bereichen zugeordnet. Der aktuelle Zustand der Metadaten-Verwaltung und ihr Nutzen für die Nutzergruppen kann dann über die Berechnung und Betrachtung der einzelnen Qualitätskriterien festgelegt werden. Außerdem kann die Metadaten-Verwaltung so entwickelt werden, dass die Qualitätskriterien bestmöglich erfüllt würden. So könnte beispielsweise ein Teil der Zielsetzung einer Metadaten-Verwaltung für das Programmieren-im-Web eine übersichtliche und verständliche Verwaltung sein, welche die Prinzipien der Klarheit, Vollständigkeit und Übersichtlichkeit erfüllen sollte. Hinzu kommt, dass für die einzelnen Nutzergruppen, welche die Verwaltung nutzen, Kriterien bestimmt werden können, welche für sie im Zentrum stehen und somit eine Art ‚Hauptkriterium‘ bilden. Bei der Nutzung der Metadaten-Verwaltung können dann diese Kriterien für die einzelnen Nutzergruppen im Fokus stehen und speziell behandelt werden.

1.2 Ausgangsposition der Forschung

Bevor sich an der Konstruktion einer solchen Lösung versucht wird, soll an dieser Stelle zuerst ein kurzer Überblick gegeben werden, welche Konzepte es bereits für die Verwaltung von Metadaten in der Vergangenheit gab, wie der aktuelle Stand im Bereich der Metadaten-Verwaltung ist und ob nicht vielleicht schon ein Ansatz für eine solche Lösung existiert. Für die gesamte Problematik der allgemeinen Definition von Metadaten sei auf Abschnitt 2.1 verwiesen, der eine ganze Fülle von Definitionen für Metadaten umfasst und analysiert. An dieser Stelle sei nur so viel vorweggenommen, dass die Definition der Metadaten als ‚Daten über Daten‘, wie sie beispielsweise in [133], [25], [27], [11], [135] beschrieben wird, bei weitem nicht ausreicht.

Wird die Verwaltung von Metadaten betrachtet, kann zuerst geschaut werden, was in der Vergangenheit für Konzepte entwickelt wurden, um Metadaten sinnvoll zu verwalten. Ein guter Überblick darüber kann bei Marco in [88] gefunden werden. Neben dem schon

erwähnten Data Dictionary für relationale Datenbanken wurden in den siebziger Jahren auch die sogenannten *computer aided software engineering tools* (CASE-Tools) entwickelt. Ursprünglich entwickelt, um den Entwicklungsprozess für Datenbanken und Softwareanwendungen zu unterstützen, wurden sie auch als Metadaten-Verwaltung genutzt, da sie auch Daten über die Daten speicherten, die sie ursprünglich verwalteten. So wurde auf Forderung vieler Nutzer von CASE-Tools eine Schnittstelle geschaffen, um die Metadaten aus verschiedenen Tools zu verbinden und auszulesen. Aus der Komplexität der Integration verschiedener CASE-Tools in eine Schnittstelle fing 1987 die Entwicklung des *CASE data interchange format* (CDIF) durch die *Electronic Industries Alliance* (EIA) an. Das CDIF ermöglichte es, ein Metamodell zu entwickeln, auf dessen Grundlage die CASE-Tool-Integration durchgeführt werden konnte. Das CDIF und die CASE-Tools setzten sich aber letztendlich nicht durch. Ein Grund dafür ist, dass in den achtziger Jahren verschiedene große Unternehmen wie IBM ankündigten, dass sie eine neue Art von Metadaten-Verwaltung auf den Markt bringen würden. Dabei sollte es sich um die ersten, sogenannten Metadaten-Repositories handeln, welche noch auf Großrechnern zum Einsatz kamen. Aus dieser Ankündigung entstanden die ersten Metadaten-Initiativen, die sich aber erst einmal nur auf technische Metadaten fokussierten. Bei diesen Metadaten handelte es sich um aufpolierte Daten, wie sie auch schon in den Data Dictionaries der relationalen Datenbanken zu finden waren. Erst in den neunziger Jahren konnte ein Wert in Metadaten-Repositories erkannt werden. Es wurden Client-Server-Strukturen genutzt, um es den Nutzern oder anderen Systemen wie Entscheidungssystemen zu ermöglichen, auf die Metadaten zuzugreifen und diese zu verarbeiten. Diese Art von Metadaten-Repository-System wird auch heute noch in Büchern, wie bei Tannenbaum in [121] oder Marco in [88], als Systemarchitektur für eine Metadaten-Verwaltung vorgeschlagen, wobei sich natürlich einzelne Faktoren über die Zeit weiter angepasst haben.

Wird der Forschungsbereich der Metadaten und deren Verwaltung in der jüngeren Vergangenheit und heute untersucht, so werden unterschiedliche Themenbereiche gefunden, auf die sich die Forschung konzentriert. Die drei Hauptthemen, zu denen sich die meisten Artikel finden lassen, sind: Entwicklungen von Frameworks für Metadaten und deren Verwaltung, Modelle für Metadaten und Anwendungen, in denen Metadaten zur Lösung von Problemen zum Einsatz kommen. Im Folgenden soll ein kurzer Überblick über einige Arbeiten in den oben aufgeführten Themenbereichen gegeben werden. Es handelt sich dabei aber nur um einen Ausschnitt potentieller Arbeiten, da eine vollständige Betrachtung aller Arbeiten den Umfang dieses Abschnitts bei weitem überschreiten würde. Den Anfang sollen die Frameworks machen. Diese sind meist für einen bestimmten Anwendungsbereich gedacht oder umfassen eine bestimmte Menge von thematisch zusammengehörenden Metadaten. Das Framework repräsentiert hierbei nicht nur ein Modell, sondern auch die technische Umsetzung sowie die benötigten Strukturen und Umgebungen. Bei Tsou in [135] wird ein Framework für Geo-Metadaten beschrieben, das für das Suchen, Indexieren und Auslesen von entsprechenden Metadaten verwendet werden kann. Farinha und Triqueriros beschreiben in [46] ein erweiterbares Framework für Qualitätsmetadaten, welches zur strukturellen Validierung von Daten verwendet werden kann. Ein Framework zur Suche von Lernobjekten, die mit Metadaten aus dem entsprechenden Themengebiet ausgestattet sind, wird von Song in [117] beschrieben. Im Bereich der Landwirtschaftsforschung wird von Yunpeng et al. in

[34] ein Framework entwickelt, welches für die Fusion von Informationen und das Nutzen von Informationsservices genutzt werden kann. Mit dem „MetaFa“, das von Ikebe et al. in [64] erläutert wird, existiert ein Framework, mit dessen Hilfe auf Basis von Metadaten in datenintensiven Applikationen Daten leicht gefunden und ausgetauscht werden können. Anitha und Mukaherjee beschreiben in [11] unter dem Begriff „Metadata as a Service“ (MaaS) ein Framework, welches das Retrieval von Daten aus einer Cloud ermöglicht. Die Metadaten bilden dazu den entsprechenden Service. Eine etwas andere Art von Framework kann bei Handschuh et al. in [54] gefunden werden. Das dort präsentierte Framework wird zur semiautomatischen Generierung von Metadaten genutzt, nachdem es für den entsprechenden Anwendungsfall trainiert wurde. Zuletzt beschreiben Reis et al. in [105] ein Framework, das für den Austausch von Metadaten mit XML als Grundlage verwendet werden kann.

Neben den Frameworks werden auch häufig Modelle für Metadaten entwickelt. Diese sind ähnlich zu den Frameworks üblicherweise im Kontext einer konkreten Anwendung angesiedelt. Im Fokus dieses Themenbereichs steht die Konstruktion eines konkreten Modells sowie die Begründung, warum dieses Modell in dem betrachteten Anwendungsbereich verwendbar ist. Allgemeinere Modelle werden auch teilweise beschrieben und es wird auf deren Spezialisierbarkeit auf konkrete Anwendungsfälle verwiesen. Die Modelle bilden aber nur eine Grundlage für eine komplette Metadaten-Verwaltung im entsprechenden Anwendungsbereich. In der Arbeit von Chen et al. [28] wird ein adaptives Metadatenmodell vorgestellt, das für die Beschreibung von Interfaces von domain-spezifischen Services verwendet werden kann. Paganelli et al. entwickeln im Bereich des Dokumentenmanagements in [98] ein Modell, welches es erlaubt, unstrukturierte Dokumente in Anwendungen wiederzuverwenden, zu integrieren und sie interoperabel zu nutzen. Das Modell von Ehrig et al. in [44] hingegen kommt in semantischen Peer-to-Peer-Systemen (P2P-Systemen) mit dem Zweck zum Einsatz, dort die semantischen Informationen zu kodieren, um verschiedene heterogene, sich ausschließende Sichten auf relevante Bereiche des P2P-Systems für den Informationsaustausch zu ermöglichen. Ein weiterer Einsatz eines Metadatenmodells kann in der Arbeit von Wu und Jin in [148] gefunden werden. Dort wird für den Bereich des Semantic Web, und der P2P-Systeme ein Modell entwickelt, welches verschiedene Features aus dem Bereich der Ontologien mit dem Ziel kombiniert, zu kodieren, Referenzen mit semantischen Informationen anzupassen und flexibel Wissen innerhalb eines Peers zu beschreiben. Orgel et al. haben in [96] ein Metadatenmodell entwickelt, das im Fachgebiet des heterogenen Kulturerbes genutzt wird und dort für das Mapping von Datenmodellen für Datenquellen auf die übliche Repräsentation in die relevanten Bereiche zum Einsatz kommt. Für die sogenannten RAGE-Werte entwickeln Georgiev et al. in [49] ein Modell für angewandte Spiele, welches einerseits die RAGE-Werte selbst und andererseits auch deren potentielle Nutzung in den drei Dimensionen „Technologie“, „Gaming“ und „Pädagogik“ beschreibt. Ein etwas allgemeineres Metadatenmodell wird von Papageorgiou et al. in [99] entwickelt, welches, geleitet durch Metadaten, eine statistische Verarbeitung ermöglicht, indem Metadaten in einem statistischen Informationssystem (SIS) generiert werden. Über SIS können dann die Daten und deren Metadaten manipuliert werden. Bental et al. konzentrieren sich in [18] mit ihrem Modell für Metadaten auf den Nutzer, indem sie ihm mit dem Modell ermöglichen, Ressourcen, welche im Web zu finden sind, zu selektieren. Für Toolkits im Bereich von grid-

basierten Problemlösungsumgebungen (PSE) entwickelten Comito, Mastroianni und Talia in [32] ein Metadatenmodell, welches das Design und die Implementierung einzelner Umgebungen ermöglicht. Ji, Nammee und Kang haben in [70] ein Metadatenmodell entwickelt, das eine transparente und vergrößerte Nutzung von Services im Bereich des Broadcasting von Medieninhalten unterstützt. Eine speziellere Art eines durchaus allgemeineren Modells liefern Duong et al., welche Metadaten als Kontext auffassen und so ein Modell schaffen, das Daten, Metadaten und Kontext vereint, um die Bedeutung und Information über die Umgebung von detaillierten Daten zu erhalten. Zuletzt gibt es auch Ansätze für formale Modelle. Eines wäre das von Cardinaels et al. in [25] entwickelte, formale Metadatenmodell für Lernmetadaten, welches ein besseres, formales und rigoroses Schlussfolgern über die Lernmetadaten des Learning Object Metadata Standard (LOM) erlaubt.

Der letzte große Themenbereich beschäftigt sich mit der Anwendung von Metadaten zur Lösung eines konkreten Problems in einem Anwendungsbereich. Dabei kann es sich um die Beschreibung einer Menge von Metadaten handeln, die eine entwickelte Lösung unterstützt, um Schemata, die eine Menge von Metadaten beschreiben, die in der Lösung verwendet werden oder um eine komplette Metadaten-Verwaltung, die eine Gesamtlösung darstellt. Eine Gemeinsamkeit dieser Arbeiten liegt in dem gleichen Anwendungsbereich. Die Wiederverwendbarkeit der Lösung in anderen heterogenen Problembereichen wird dabei meist nicht betrachtet. So beschreiben Apps und Macintyre in [12], wie mit Hilfe der Metadaten, die im Dublin Core Standard vorhanden sind, bibliographische Informationen für Journale erhalten werden können. Dafür wird auch ein erweitertes Metadatenschema angegeben. Für den Bereich des Data Governance beschreiben Helvoirt und Weigand in [136] ein Metadaten-Management, das es erlaubt, Transparenz zu schaffen, indem mit Metadaten nicht nur dokumentiert wird, sondern auch Policies erzwungen werden. Daraus können als Ergebnis Geschäfts- und Datenregeln gefolgert werden. Charalabidis et al. entwickeln in [26] ein Metadatenschema basierend auf Ontologien, das dafür genutzt wird, E-Governance-Ressourcen, wie Services, XML-Schemata oder öffentliche Stellen zu beschreiben. Ziel ist es, einen formalisierten Austausch zwischen Portalen und Registern zu ermöglichen. Für die Kontrolle und Nutzung von föderierten Datenbanksystemen im medizinischen Bereich entwickeln Rantree et al. in [106] eine Lösung, die zu diesem Zweck Metadaten verwendet. Zhang et al. beschreiben in [150] ein Metadaten-Management, das es erlaubt, in kabellosen Sensor-Netzwerken adäquate Informationen für das Ausführen von Anfragen an das Netzwerk bereitzustellen. Metadaten können auch verwendet werden, um in der digitalen Forensik bestehende Verbindungen zwischen Dateien, wie Bildern, über metadaten-basierte Verbindungen aufzuzeigen und zu suchen, wie Raghavan und Raghavan in [104] zeigen. Juszczyszyn zeigt in [71] die Modellierung von evolutionären Metadaten, um mit diesen Metadaten Aktivitäten eines simulierten, sich verändernden Multi-Agenten-Systems auszudrücken. In [14] wird von Baldonado et al. die komplette Metadatenarchitektur des digitalen Bibliotheksmanagements der Stanford Universität beschrieben. Dies bietet eine Infrastruktur, um die Interoperabilität zwischen heterogenen und autonomen Services der digitalen Bibliothek zu ermöglichen. Zuletzt beschreiben Pardillo et al. in [101] für ein Data Warehouse, wie der OLAP-Würfel mit Metadaten angereichert werden kann, um Anfragen an den Würfel über Tools von Nutzern zu ermöglichen.

Neben diesen drei Hauptbereichen gibt es natürlich noch weitere spezielle Arbeiten, die sich beispielsweise mit der Generierung von Metadaten oder dem Import von Metadaten beschäftigen. Exemplarisch für einen solchen kleineren Themenbereich steht das Gebiet, das sich sehr allgemein mit dem Thema Metadaten-Management beschäftigt. So beschreiben Bücher wie das von Tannenbaum [121] und Marco [88] oder Artikel wie die von Owens [97], Neuroth [93] und Roskiewicz [107] allgemein bestimmte Bereiche der Metadaten-Verwaltung, wie die Erstellung eines Repositorys, den Umgang mit Metadatenstandards, die Entwicklung von Metadaten oder den sinnvollen Einsatz einer Metadaten-Verwaltung in einem Betrieb. Hier stehen meist keine konkreten Umsetzungen im Vordergrund, sondern das Aufzeigen von Strategien oder das Anleiten beim Umgang mit Metadaten und deren Verwaltung. Damit sollte es möglich sein, eine konkrete Metadaten-Verwaltung in einem passenden Anwendungsszenario zu konstruieren. Solche Arbeiten können daher eher als Leitfaden für eine gute Metadaten-Verwaltung angesehen werden.

1.3 Forschungsagenda

Die Herausforderung dieser Arbeit liegt in der Entwicklung eines konkreten, allgemeinen Ansatzes, welcher die in Abschnitt 1.1 beschriebene Lösung realisiert. Die Forschungsfragen (FF), die dafür im Mittelpunkt stehen, sind:

- FF1:** Existiert eine Definition für Metadaten, die als universelle Beschreibung genutzt werden kann?
- FF2:** Kann eine solche Definition geeignet formalisiert werden?
- FF3:** Kann ein Modell basierend auf einem Schema konstruiert werden, welches die relevanten Kategorien von Metadaten in Stereotypen abbildet?
- FF4:** Wie sind Sichtweisen auf ein allgemeines Modell zu stützen, abzuleiten und vernünftig zu verwalten?
- FF5:** Wie muss eine Lösung zur Realisierung der Metadaten-Verwaltung aufgebaut und prototypisch als Proof of Concept umgesetzt werden?
- FF6:** Kann die entwickelte Lösung in einem realen Anwendungsfall bestehen?

Die Forschungsfragen FF1 und FF2 bauen auf einander auf. So ist es zuerst einmal wichtig herauszufinden, ob es überhaupt möglich ist, eine Definition für Metadaten zu finden, die so allgemein ist, dass sie in verschiedenen heterogenen Szenarien Anwendung finden kann. Die Herausforderung liegt dabei in den verschiedenen Aspekten, wie der Sichtweise der Nutzer auf Daten und Metadaten, die auf die eigentliche Definition einwirken. Es muss also eine Möglichkeit gefunden werden, diese Aspekte in eine Definition mit aufzunehmen, sodass sie zwar der Definition zugeordnet werden können, aber nicht Teil der eigentlichen Beschreibung der Metadaten sind. Kann eine solche in FF1 gesuchte Definition angegeben werden, so wird mit FF2 eine formale Darstellung dieser Definition gesucht. Wie genau die passenden mathematischen Methoden, Theorien, Beweise und Konstrukte aussehen und

wie sie zu einer formalen Definition zusammengestellt werden müssen, bildet die große Herausforderung für diesen Bereich.

Mit der FF3 geht die Konstruktion einer Metadaten-Verwaltung von einer eher theoretischen in eine praktischere Betrachtungsweise über. Eine positive Beantwortung dieser Frage führt zu einem hinreichend universellen Schema, welches zum Verwalten von Metadaten verwendet werden kann. Hierbei finden sich eine Reihe von Herausforderungen, die für ein solches Schema gelöst werden müssen. Zuerst wird eine Kategorisierung von Metadaten benötigt, die eine bestmögliche Abdeckung von relevanten Metadaten liefert. Basierend auf diesen Kategorien müssen Schemata entwickelt werden, welche so allgemein sind, dass sie in verschiedenen Szenarien mit möglichst geringem Anpassungsaufwand verwendet werden können. Dies führt zu einer Reihe von Stereotypen. Für diese wird zuletzt ein Konzept benötigt, um die Schemata, die für die Kategorien entwickelt werden, als Stereotype nutzbar zu machen. Das in FF3 gesuchte Schema ist aber nur ein Teil der Lösung, das für eine allgemeine Metadaten-Verwaltung benötigt wird. Damit es einen Nutzen erbringt, muss es möglich sein, für verschiedene Nutzergruppen, welche die Stereotypen verwenden möchten, zu bestimmen, was diese Nutzer für Sichtweisen auf die Metadaten in ihrer Anwendung konkret benötigen. Aber nicht nur das Identifizieren, sondern auch das Zusammenstellen und das mögliche Umdefinieren dieser Sichtweisen sind die zentralen Herausforderungen an eine Metadaten-Verwaltung. Somit wird die FF4 zu einer der wichtigsten Forschungsfragen für eine allgemeine Metadaten-Verwaltung, da sie den Nutzen der Stereotype und somit des gesamten Schemas determiniert.

Das Ergebnis aus den FF3 und FF4 wird dann in den beiden Forschungsfragen FF5 und FF6 verwendet, um an eine konkrete Metadaten-Verwaltung zu gelangen und diese für einen Anwendungsfall zu testen. Das Ergebnis der FF5 ist nicht nur das Konzept eines Systems, welches die Verwaltung von Metadaten auf Basis der Ergebnisse aus FF1 und FF3 ermöglicht, sondern auch eine prototypische Implementierung, welche als Proof of Concept verwendet werden kann. Die Herausforderungen hier, liegen dabei vor allem in der Konzeptionierung und Implementierung einer Umsetzung, die es ermöglichen muss, Metadaten zu importieren, auszulesen, zu verarbeiten und zu exportieren. Dazu gehört auch die geeignete Wahl und Verwendung technischer Mittel zur Implementierung. Das so entwickelte Proof of Concept hat bis dahin aber das Problem, das noch nicht gezeigt ist, dass es in der Realität auch sinnvoll und gewinnbringend eingesetzt werden kann. Dies wird erst über die FF6 und deren Ergebnis geregelt, indem der Prototyp in einem konkreten Anwendungsfall zum Einsatz kommt. Die Integration und Verwendung des Prototyps zur Metadaten-Verwaltung für einen konkreten Anwendungsfall stellt hier die größte Herausforderung dar.

1.4 Aufbau der Arbeit

Diese Arbeit verfolgt das Ziel, einen konkreten, allgemeinen Ansatz für die Verwaltung von Metadaten zu entwickeln. Dazu wird in Kapitel 2 zuerst untersucht, ob eine allgemeine Definition von Metadaten möglich ist. Durch die Untersuchung zweier Szenarien,

der Kollaboration und der Integration, wird herausgestellt, dass Metadaten mit Hilfe eines Modells definiert werden können. Außerdem werden sieben Kategorien ermittelt, zu denen Metadaten zugeordnet werden können. Als Konsequenz daraus wird ein allgemeines Metadatenmodell angegeben, welches es ermöglicht, abhängig von den Szenarien, in denen Metadaten verwendet werden, eine Definition zu bestimmen. Für das Modell wird auch gezeigt, dass es in seiner Rolle als Metadatenmodell adäquat und verlässlich ist. Außerdem wird sowohl die Notwendigkeit von Stereotypen zur Verwaltung von Metadaten herausgearbeitet als auch die Funktionalität ermittelt, die für eine Verwaltung vorhanden sein muss.

Die Realisierung der benötigten Stereotypen wird in Kapitel 3 behandelt. Dazu werden als grundlegende Technik, die Konzepte der Datenbankkomponenten und der Kabelbäume eingeführt. Eine Datenbankkomponente benötigt ein Schema als Kern und besitzt eine Reihe von Interfaces, die für den Input und Output von Daten innerhalb der Komponente verwendet werden. Für die bereits ermittelten Kategorien von Metadaten werden dann die Schemata konstruiert, welche die Kerne für die Komponenten repräsentieren und den Stereotypen für das Metadatenmodell entsprechen. Als relevante Kategorien werden dazu die Bereiche „Administration“, „Provenance“, „Qualität“, „Struktur“, „Service“, „Raum“ und „Zeit“ betrachtet und entsprechende Schemata entwickelt. Die daraus resultierenden Komponenten können dann mit Hilfe der Kabelbäume verbunden werden und formen ein Metadaten-Data-Dictionary, das die Aufgabe eines internen Repository-Systems erfüllt.

Damit die Komponenten und Kabelbäume für die Verwaltung verwendet werden können, beschäftigt sich das Kapitel 4 mit der Realisierung der Funktionalität, die für die Metadaten-Verwaltung nötig ist. Hierzu werden zuerst die grundlegend benötigten Konzepte, wie erweiterte Datenbanksichten und Mediatoren, beschrieben. Daraus entsteht ein System zur Verwaltung der Metadaten, das als Mediator realisiert wird. Der Mediator besitzt vier Kernbereiche, und zwar „Create“, „Import“, „Retrieval“ und „Export“, in denen er die entsprechende Funktionalität bereitstellen muss. Für die einzelnen Bereiche wird jeweils die konkrete Realisierung beschrieben.

Um zu zeigen, dass das Konzept der Arbeit funktioniert, wird in Kapitel 5 eine reale Anwendung mit Hilfe der Komponenten und des Mediators entwickelt. Als Anwendung dient die Entwicklung einer Webseite für das Internationale Kontinentale Tiefbohrprogramm (ICDP) in Deutschland. Für die Anwendung wird eine Daten- und Metadaten-Verwaltung benötigt, sowie spezielle Funktionen, wie eine eigenschaftsbasierte Suche, die einen Einsatz des entwickelten Konzepts ermöglicht. Für den Nachweis des Konzepts werden in dem Kapitel die angepassten Datenbankkomponenten beschrieben, wie sie in der Anwendung verwendet werden. Außerdem wird auf die Funktionen eingegangen, die mit Hilfe des Mediators erstellt werden, um die Anforderungen an die Anwendung zu erfüllen.

1.5 Abgrenzung der Arbeit

Diese Arbeit beschäftigt sich mit Metadaten und deren Verwaltung. Hierbei umfasst der Begriff „Metadaten“ nur die Elemente, die sich mit der allgemeinen Definition der Daten über Daten erfassen lassen. Zur besseren Einordnung der Arbeit kann der sogenannte Meta

Object Facility (MOF)-Standard der Object Management Group (OMG) [6] verwendet werden. Die Tabelle 1.1 zeigt die erste grobe Einordnung in diesen Standard. Dabei liegt die mit M_3 bezeichnete Ebene der Frameworks explizit außerhalb des Rahmens dieser Arbeit. Auf der untersten Ebene liegen die Exemplare oder auch Instanzen, wobei es sich im Kontext der Arbeit um die Metadaten handelt, die den Objekten in Form von Daten zugeordnet sind. Auf der M_1 liegt das in dieser Arbeit vorgestellte Metadatenmodell. Dabei handelt es sich um ein informatives Modell, welches orthogonal zum eigentlichen Datenbankmodell liegt. Zuletzt liegen auf der Ebene M_2 die generischen Metadaten-Schemata dieser Arbeit, welche als konfigurierbares Modell über der M_1 verstanden werden kann, das je nach Szenario und Metadatenmodell angepasst wird.

Tabelle 1.1: Einordnung in den Meta Object Facility (MOF)-Standard [6]

Abstraktionsebene	Modell-Tableau	Metadaten-Verwaltung
M_3	Framework	außerhalb des Rahmens der Arbeit
M_2	Rahmen	generische Schemata als konfigurierbares Modell für M_1
M_1	Modell	informatives, allgemein nutzbares Modell orthogonal zum DB-Modell
M_0	Exemplare	Daten als Objekte

Das Problem an der Einordnung in den MOF-Standard liegt darin, dass die generischen Schemata, die in dieser Arbeit entwickelt werden, eigentlich kein allgemeines Metamodell darstellen, welches aber auf der M_2 gefordert wird. Wie die Abbildung 1.1 zeigt, können diese Schemata auf eine Zwischenebene M_{12} eingeordnet werden, die eigentlich mit Hilfe eines Metamodells instantiiert werden muss. Je nach Szenario und Anwendung werden dann entsprechende Schemata für die Metadaten-Verwaltung in die M_1 spezialisiert. Der Fokus dieser Arbeit liegt dabei nicht auf der Entwicklung eines allgemeinen Metamodells, welches auf der M_2 liegt. Die Arbeit soll lediglich die generischen Schemata konstruieren, aus denen die nutzbaren Modelle für die M_1 spezialisiert werden können.

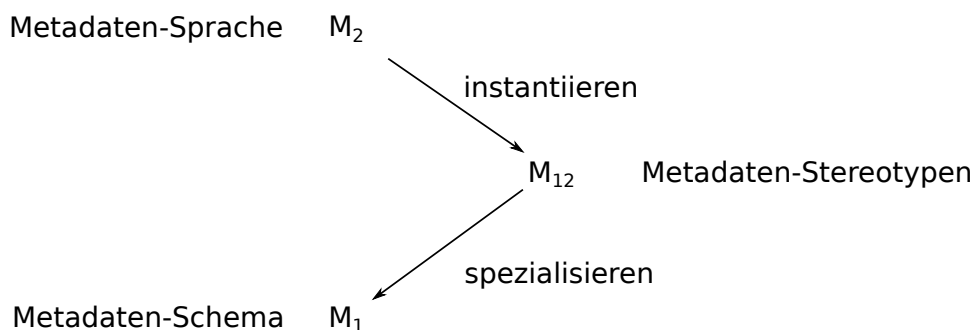


Abbildung 1.1: Anpassung der Einordnung in den MOF-Standard

Auch wenn in der Arbeit ein erster Ansatz für eine allgemeine Metadaten-Verwaltung präsentiert wird, so wird dieser nicht als strikte Handlungsvorschrift angesehen, wie mit Problemen im Bereich der Metadaten umzugehen ist. Daraus ergibt sich auch, dass diese Arbeit keine Belehrung für die Community im Bereich Metadaten, wie die OMG oder die Community des Dublin Cores, darstellt, sondern nur einen Ansatz zur stereotypischen Anwendung von Metadaten als ein adäquates Mittel zur Verwaltung vorstellt. Die Lösung selbst kann zwar in einer Reihe von Szenarien zum Einsatz kommen, was den hier entwickelten Ansatz hinreichend allgemein macht. Es wird aber keine allumfassende Lösung jedes erdenklichen Problems im Bereich der Metadaten-Verwaltung entwickelt, da dies aufgrund der Vielfalt und Komplexität des Problems nicht im Rahmen einer einzigen Arbeit möglich ist.

Trotz der Konzentration auf die deutsche Sprache werden einzelne Begriffe in dieser Arbeit in englischer Sprache verwendet. Die Verwendung der englischen Begriffe kann unterschiedliche Gründe haben. So kann es sich um Fachbegriffe handeln, die im Kontext der Informatik bereits auf Englisch verwendet werden. Auch werden manche Begriffe bereits in der deutschen Sprache in ihrem englischen Original verwendet und werden nicht übersetzt. Zuletzt existieren für manche englische Fachbegriffe keine adäquate Übersetzung in die deutsche Sprache.