Aus dem Institut für Medizinische Biometrie und Informatik der Universität Heidelberg
Direktor: Prof. Dr. sc. hum. Meinhard Kieser

Datenintegration in biomedizinischen Forschungsverbünden auf Basis von serviceorientierten Architekturen

Inauguraldissertation
zur Erlangung des Doctor scientiarum humanarum
an der
Medizinischen Fakultät Heidelberg
der
Ruprecht-Karls-Universität

vorgelegt von
MATTHIAS GANZINGER
aus
Schweinfurt
2014

Dekan: Prof. Dr. med. Claus R. Bartram
Doktormutter: Frau Prof. Dr. sc. hum. Petra Knaup-Gregori

INHALTSVERZEICHNIS

ABBILDUNGSVERZEICHNIS

TABELLENVERZEICHNIS

LISTINGS

ABKÜRZUNGEN

aCGH	Array comparative genomic hybridization
ANOVA	Analysis of Variance
ANSI	American National Standards Institute
API	Application Programming Interface
ATCC	American Type Culture Collection
BAO	BioAssay Ontology
BFO	Basic Formal Ontology
BMBF	Bundesministerium für Bildung und Forschung
BPEL	Business Process Execution Language
caBIG	Cancer Biomedical Informatics Grid
caCORE	Cancer Common Ontologic Representation Environment
caDSR	Cancer Data Standards Registry and Repository
caGRID	Cancer Grid
CBO	Clinical BioInformatics Ontology
CCONT	Cell Culture Ontology
cDNA	Complementary DNA
ChEBI	Chemical Entities of Biological Interest
CLIMA	Cell Line Integrated Molecular Authentication
CLO	Cell Line Ontology
CNV	Copy-Number Variations
CORBA	Common Object Request Broker Architecture

CQL	caGrid Query Language
CSM	Common Security Module
CSV	Character Separated Value
DDL	Data Definition Language
DFG	Deutsche Forschungsgemeinschaft
DIN	Deutsches Institut für Normung
DMZ	Demilitarized Zone
DNA	Deoxyribonucleic Acid
DNS	Desoxyribonukleinsäure
DO	Human Disease Ontology
DSMZ	Deutsche Sammlung von Mikroorganismen und Zellkulturen
ECACC	European Collection of Cell Cultures
EFO	Experimental Factors Ontology
ETL	Extract, Transform, and Load
EVS	Enterprise Vocabulary Service
EXACT	Experiment Actions
FACS	Fluorescence-activated Cell Sorting
FMA	Foundational Model of Anatomy
GAARDS	Grid Authentication and Authorization with Reliably Distributed Services
GO	Gene Ontology
HBV	Hepatitis B Virus
HCC	Hepatocellular Carcinoma
HCV	Hepatitis C Virus
HTML	Hypertext Markup Language
HTTPS	Hypertext Transfer Protocol Secure
HTTP	Hypertext Transfer Protocol
i2b2	Informatics for Integrating Biology and the Bedside
ICLC	Interlab Cell Line Collection
IEEE	Institute of Electrical and Electronics Engineers
IEV	INOH Event Ontology
INOH	Integrating Network Objects with Hierarchies
ISCN	International System for Human Cytogenetic Nomenclature
ISO	International Organization for Standardization
IT	Informationstechnologie
JSF	JavaServer Faces
JSR	Java Specification Request
KEGG	Kyoto Encyclopedia of Genes and Genomes
LDAP	Lightweight Directory Access Protocol
MCCL	Molecular Connections Cell Line Ontology
MeSH	Medical Subject Headings
mRNA	Messenger RNA
NCIm	NCI Metathesaurus

NCIP	National Cancer Informatics Program
NCIt	NCI Thesaurus
NCI	National Cancer Institute
ncRNA	Non-coding RNA
NGS	Next Generation Sequencing
NIH	National Institutes of Health
NISO	National Information Standards Organization
NLM	National Library of Medicine
OBI	Ontology for Biomedical Investigations
OBO	Open Biomedical Ontologies
OLS	Ontology Lookup Service
OWL	Web Ontology Language
PAP	Policy Administration Point
PDP	Policy Decision Point
PELICAN	Platform Enhancing Liver Cancer Networked Research
PEP	Policy Enforcement Point
QDA	Qualitative Datenanalyse
qPCR	Quantitative Polymerase Chain Reaction
RDF	Resource Description Framework
REST	Representational State Transfer
RNA	Ribonucleic Acid
SAML	Security Assertion Markup Language
SAN	Storage Area Network
SDK	Software Development Kit
SFB	Sonderforschungsbereich
SKOS	Simple Knowledge Organisation System
SNOMED	Systematized Nomenclature of Medicine
SOAP	Simple Object Access Protocol
SOA	Serviceorientierte Architektur
STR	Short Tandem Repeat
TMA	Tissue Microarray
TRR	Transregio
UMLS	Unified Medical Language System
UML	Unified Modeling Language
URI	Uniform Resource Identifier
UTS	UMLS-Terminology Services
VOCALIGN	Vocabulary Alignment
WSDL	Web Services Description Language
W3C	World Wide Web Consortium
XHTML	Extensible HyperText Markup Language
XMI	XML Metadata Interchange
XML	Extensible Markup Language

1 EINLEITUNG

GEGENSTAND UND BEDEUTUNG

Mit der Zusammenfassung von biomedizinischen Forschungsprojekten zu Forschungsverbünden erhofft man sich durch Synergie weitergehende Erkenntnisse zu gewinnen, als dies bei separaten Projekten der Fall wäre. Hierzu werden gemeinsame Ressourcen für die Erarbeitung der Ergebnisse verwendet.

Da die Teilprojekte eines Forschungsverbundes jedoch verschiedene Aspekte des übergeordneten Forschungsgebiets bearbeiten, fallen als Ergebnis Daten unterschiedlicher Art an. Um den größtmöglichen Nutzen aus diesen heterogenen Datenbeständen zu ziehen, ist es erforderlich, die Daten zusammenzuführen und die Auswertung der Daten über die Teilprojekte hinweg zu ermöglichen.

Von der Deutschen Forschungsgemeinschaft (DFG) werden zum Stichtag 28. November 2013 insgesamt 61 Forschungsverbünde als Sonderforschungsbereich (SFB) der Variante Transregio (TRR) gefördert [Deutsche Forschungsgemeinschaft 2013]. Bei diesen ist als Förderziel explizit festgeschrieben, dass die Beiträge der Verbundpartner „essentiell, komplementär und synergetisch" [Deutsche Forschungsgemeinschaft o. J.] sein müssen.

MOTIVATION

Obwohl in der biomedizinischen Forschung der Einsatz von Werkzeugen der Informationstechnologie (IT) zur Bearbeitung der Ergebnisse allgegenwärtig ist, kann die übergreifende Auswertung der Daten vielfach nicht erfolgen, da die verwendeten Lösungen nicht miteinander kombinierbar sind. Um aber eine IT-gestützte Auswertbarkeit der gewonnenen Erkenntnisse zu erreichen, ist es erforderlich, die Daten zusammenzuführen, sei es in einer gemeinsamen Datenbank oder durch die Vernetzung der Datenbestände.

Da die Daten heterogen sind (zum Beispiel Bilddaten und genomische Daten) ist die Vorgabe eines einheitlichen Datenformates für alle Teilprojekte nicht praktikabel. Vielmehr ist es erforderlich, die einzelnen Datenarten durch Metadaten zu beschreiben und so eine Auswertung über verschiedene Datenarten hinweg zu ermöglichen. Weiterhin bedarf es einer geeigneten Benutzerschnittstelle, die es den Forschern auch ohne tiefgreifende IT-Kenntnisse ermöglicht, ihre Fragestellung zu formulieren und mit Hilfe des Datenbestandes zu beantworten.

Auch wenn die gemeinsame Auswertung der Daten vorrangiges Ziel ist, muss dennoch sichergestellt sein, dass die Daten der einzelnen Projekte hinreichend geschützt werden. Daher ist es erforderlich, den Zugriff auf die Daten zu kontrollieren, sodass eine unberechtigte Verwendung der Daten wirksam verhindert wird.

In dieser Arbeit wird anhand eines Forschungsverbundes zum Hepatozellulären Karzinom (engl. Hepatocellular Carcinoma, HCC) untersucht, mit welchen Methoden und Werkzeugen eine IT-Plattform gestaltet werden kann, welche die integrierte Auswertung der Daten des Verbundes ermöglicht.

PROBLEMATIK

Diese Arbeit entstand im Rahmen des Teilprojektes *Z2: Integrierte Informationsplattform und projektübergreifende biostatische Auswertungen* aus dem SFB/TRR77 *Leberkrebs – von der molekularen Pathogenese zur zielgerichteten Therapie*.

In diesem Verbund werden beispielsweise Microarraydaten, histologische Schnittbilder und klinische Daten erzeugt oder verarbeitet. Bisher werden die Daten häufig in Tabellen mit Hilfe der Software Microsoft Excel erfasst und bearbeitet. Dies führt dazu, dass die in den einzelnen Arbeitsgruppen erstellten Datensätze in unterschiedlichen Formaten vorliegen und nicht ohne manuelle Eingriffe zusammengeführt werden können. Weiterhin kann nicht ohne Weiteres nachvollzogen werden, wie die Daten bearbeitet wurden, sodass eine mangelhafte Datenqualität unter Umständen nicht erkannt wird.

1.1 PROBLEMSTELLUNG

Es existiert keine Plattform, auf der die in den Projekten eines Forschungsverbundes gewonnenen heterogenen Daten systematisch integriert, ausgewertet und für andere Projektgruppen bereitgestellt werden können. Insbesondere fehlen Konzepte, um die Daten durch Metadaten so zu beschreiben, dass ein automatisches Auswerten auch heterogener Datenquellen ermöglicht wird.

In biomedizinischen Forschungsverbünden kommen heterogene Datendefinitionen und -strukturen zum Einsatz. Weiterhin sind die Beziehungen zwischen den verschiedenen Datenarten unklar. Zusammen führt dies dazu, dass wissenschaftliche Daten nicht ohne Weiteres projektübergreifend auswertbar sind.

Weiterhin haben die Eigentümer wissenschaftlicher Daten unterschiedliche und teilweise hohe Ansprüche an die Vertraulichkeit der in den Verbund eingebrachten Daten. Werden diese Ansprüche von einem Datenintegra-

tionswerkzeug nicht erfüllt, so führt dies zu mangelnder Akzeptanz bei potenziellen Nutzern.

1.2 ZIELSETZUNG

Aus der Problemstellung werden folgende Ziele abgeleitet, die im Rahmen dieser Arbeit erreicht werden sollen:

ZIEL 1: Erfassung der Anforderungen von Forschungsverbünden an eine IT-Plattform zur Datenintegration und Erstellung eines Referenzmodells für Anforderungen.

ZIEL 2: Entwurf einer serviceorientierten Architektur (SOA) für eine IT-Plattform zur Integration und projektübergreifenden Analyse der wissenschaftlichen Daten in einem biomedizinischen Forschungsverbund.

ZIEL 3: Spezifikation von Metadaten zur Integration der heterogenen wissenschaftlichen Daten im biomedizinischen Forschungsverbund für projektübergreifende Auswertungen.

1.3 VORGEHENSWEISE

Zum Erreichen der formulierten Ziele werden die in den folgenden Abschnitten aufgezeigten Schritte durchgeführt.

Zu Ziel 1:

Die Erstellung des Referenzmodells für Ziele und Anforderungen von Forschungsverbünden an eine IT-Plattform zum Austausch ihrer Daten erfolgt auf Basis der im SFB/TRR77 gewonnenen Erkenntnisse. Hierzu werden verschiedene Quellen aus dem Verbund identifiziert, welche aus verschiedenen Perspektiven zur Erfassung der Ziele und Anforderungen von Forschungsverbünden beitragen können. Die Ziele und Anforderungen werden abstrahiert, geordnet und miteinander in Beziehung gesetzt. Schließlich entsteht ein Anforderungsmodell, welches so allgemein gehalten ist, dass es für andere Forschungsverbünde als den SFB/TRR77 als Referenz dienen kann.

Zu Ziel 2:

Grundlage für die Architektur der Datenaustauschplattform sind die Ziele und Anforderungen der intendierten Nutzer. Somit wird zunächst eine SFB/TRR77-spezifische Instanz des Referenzmodells für Anforderungen aus Ziel 1 erstellt und dokumentiert. Von besonderer Relevanz ist dabei die

Frage, welche projektübergreifenden Auswertungen durch die serviceorientierte IT-Plattform für den SFB/TRR77 bereitgestellt werden sollen. Auch die Sicherheitsanforderungen des SFB/TRR77 werden bei der Entwicklung der Architektur berücksichtigt.

Auf Basis der Anforderungen wird untersucht, inwiefern Frameworks für die medizinische Forschung wie das Cancer Biomedical Informatics Grid (caBIG) oder Informatics for Integrating Biology and the Bedside (i2b2) eine geeignete Grundlage zur Umsetzung der SFB/TRR77-SOA darstellen, beziehungsweise welche Anpassungen für den Einsatz im Forschungsverbund erforderlich sind. Um die Integration der Daten zu ermöglichen, müssen die im Rahmen von Ziel 3 definierten Metadaten in einem Metadatenverzeichnis bereitgestellt und gepflegt werden.

Zur Bereitstellung der Projektdaten werden entsprechende SOA-Schnittstellen spezifiziert und implementiert. Schließlich wird die entwickelte SOA im Rahmen des SFB/TRR77 unter dem Namen Platform Enhancing Liver Cancer Networked Research (PELICAN) prototypisch umgesetzt.

Zu Ziel 3:

In einem ersten Schritt sind die in den biomedizinischen Projekten verwendeten Daten sowie die an ihrer Erzeugung beteiligten Prozesse zu analysieren. Von besonderer Bedeutung sind dabei

- die im Projekt verwendete Terminologie,
- die Datenstrukturen sowie
- die Art der Datenhaltung.

Aus diesen Analyseergebnissen wird zunächst eine Referenzterminologie für einen biomedizinischen Forschungsverbund entworfen. Diese Referenzterminologie wird anschließend mit eingeführten, standardisierten Terminologien aus der Medizin abgeglichen.

Weiterhin wird untersucht, welche im Verbund verwendeten Datenarten einer besonderen Behandlung bezüglich ihrer Metadaten bedürfen. Für diese Datenarten wird eine Metadatendefinition in Form einer Ontologie entwickelt.

Sowohl die Referenzterminologie als auch die Ontologie sind öffentlich verfügbar, sodass sie auch außerhalb des SFB/TRR77 angewandt werden können.

2 GRUNDLAGEN

2.1 DAS HEPATOZELLULÄRE KARZINOM

Der SFB/TRR77 hat das Ziel, verschiedene Aspekte des HCC zu erforschen. Durch dieses Ziel werden die im Verbund verwendeten Daten und Analysemethoden bestimmt. Daher wird in den folgenden Abschnitten ein Überblick über die (bio)medizinischen Aspekte des HCC gegeben.

Krebs

Umgangssprachlich spricht man bei malignen Neoplasien von *Krebs* [Pschyrembel, Bach 2010]. Dabei handelt es sich um die Neubildung von Gewebe auf Grund einer gestörten Wachtumsregulation. Man unterscheidet zwischen *benignen* und *malignen* Neoplasien: Im benignen Fall ist die Neubildung lokal zum umgebenden Gewebe begrenzt, im malignen Fall kann die Neubildung in das umliegende Gewebe eindringen und es zerstören [Robbins et al. 2010].

Gewebe von Erwachsenen wird aus stark differenzierten Zellen gebildet, die ihre Spezialisierung über Zellgenerationen hinweg beibehalten. Im Gegensatz zu frei lebenden Zellen, wie Bakterien, sind die Zellen mehrzelliger Lebewesen darauf ausgelegt, miteinander zu kooperieren. Dazu existieren vielfältige Mechanismen der interzellulären Kommunikation, die sicherstellen, dass die einzelne Zelle zum Wohle des gesamten Organismus ruht, wächst, sich teilt, sich differenziert oder stirbt. Jedoch tragen die meisten differenzierten Zellen das vollständige Genom ihres Lebewesens in sich und somit mehr Information als zur Ausführung ihrer Aufgabe erforderlich ist. Weiterhin behalten viele Zellen die Fähigkeit bei, zu wachsen und sich zu teilen, sodass Regenerationsprozesse wie bei der Wundheilung möglich sind. Diese Vielseitigkeit und Autonomie der Zellen birgt jedoch die Gefahr, dass diese durch Störung ihrer Mechanismen zur Genexpression Fähigkeiten ausprägen, die ihrem Zelltyp im Normalfall nicht zur Verfügung stehen. Dies kann zu abnormalen Zell-Phänotypen führen, die ihre vorgesehene Differenzierung ganz oder teilweise verloren haben und deren Verhalten unvereinbar mit ihrer vorgesehenen Rolle im Organismus ist. Im ungünstigsten Fall verliert die Zelle die Fähigkeit, sich den Regulationsmechanismen zu unterwerfen und beginnt ein unkontrolliertes Wachstum auf Kosten des umgebenden Gewebes und letztlich des gesamten Organismus [Alberts 2008; Weinberg 2007].

Die Entstehung von Krebs wird durch verschiedene Arten von *Karzinogenen* gefördert. Diese verändern das Genom physikalisch (zum Beispiel Röntgenstrahlung) oder chemisch (zum Beispiel Hydrocarbonate), wobei diese Modifikationen zufällig neoplastische Eigenschaften der Zelle hervorrufen können. Ebenso können einige Viren Krebs auslösen. Die Information, die für die Ausprägung maligner Eigenschaften erforderlich ist, liegt dabei schon im Genom der Zelle vor. Im Normalfall werden sie jedoch nicht oder nur in geringem Umfang exprimiert, man spricht von einem *Proto-Onkogen*. Im Zuge der Krebsentwicklung wird aus dem Proto-Onkogen das eigentliche *Onkogen*, welches schließlich der Krebszelle maligne Fähigkeiten verleiht. Umgekehrt existieren im Genom *Tumor-Suppressor-Gene*, die im Normalfall eine Wirkung entfalten, die der Krebsentstehung entgegen wirkt. Wird die Expression dieser Gene gestört, so erhöht sich die Wahrscheinlichkeit der Tumorentwicklung [Robbins et al. 2010; Weinberg 2007].

Breiten sich Tumoren so weit aus, dass sie das Lymphsystem oder Blutgefäße erreichen, so können sie *Metastasen* bilden und sich in andere Teile des Körpers ausbreiten. Da der Organismus so an verschiedenen Stellen geschädigt wird, ist das Auftreten von Metastasen einer der gefährlichsten Aspekte der Krebserkrankung [Robbins et al. 2010].

Leberkrebs

Unter der Bezeichnung Leberkrebs werden die bösartigen Neubildungen der Leber verstanden. Genauer betrachtet unterscheidet man zwischen *primären* und *sekundären* Lebertumoren. Sekundäre Karzinome sind Lebermetastasen, die von Karzinomen anderer Organe in die Leber gestreut wurden [Robbins et al. 2010]. Zu den primären Malignomen der Leber zählt neben dem Gallengangskarzinom das Hepatozelluläre Karzinom (HCC), welches weltweit 70% bis 85% aller Leberkarzinome ausmacht [Ahmed et al. 2008].

Wie andere Krebsarten auch entsteht HCC durch die genetische Mutation von funktionalen Zellen des entsprechenden Gewebes, in diesem Fall der Leberzellen (Hepatozyten). Dabei gehen zelluläre Mechanismen, die das Wachstum der Zellen steuern, verloren und es kommt in der Folge zu stark erhöhter und unkontrollierter Teilung dieser Zellen (Proliferation) bis hin zum Tumor [Weinberg 2007].

Epidemiologie

Während die Bedeutung von HCC in Europa eher untergeordnet ist, ist diese Krebsart, wie in Tabelle 1 dargestellt, weltweit betrachtet an siebter Stelle der krebsbezogenen Neuerkrankungen bei der Frau und an fünfter Stelle

Tabelle 1: Weltweite Inzidenz und Mortalität von Krebs in Tausend für das Jahr 2008 (Auszug nach [Jemal et al. 2011])

INZIDENZ		MORTALITÄT	
FRAUEN	MÄNNER	FRAUEN	MÄNNER
Brust 1.384	Lunge 1.095	Brust 458	Lunge 951
Darm 570	Prostata 904	Lunge 427	***Leber 478***
Zervix 530	Darm 664	Darm 288	Magen 464
Lunge 514	Magen 641	Zervix 275	Darm 321
Bauch 349	***Leber 522***	Magen 274	Speiseröhre 276
Gebärmutter 287	Speiseröhre 237	***Leber 218***	Prostata 258
Leber 226	Blase 297	Ovarien 240	Leukämie 144

beim Mann. Am weitesten verbreitet ist HCC in Asien und Zentralafrika. Bezüglich des Alters ist das Risiko, an HCC zu erkranken in Regionen mit insgesamt niedrigem Erkrankungsrisiko um 75 Jahre am größten, während es in Hochrisikogebieten eher bei 60 bis 65 Jahren liegt [El-Serag, Rudolph 2007].

Für rund 80% aller HCC-Fälle spielt eine chronische virale Hepatitis eine Rolle bei der Entstehung [Caldwell, Park 2009]. Diese Fälle können wiederum etwa zu zwei Dritteln dem Hepatitis B Virus (HBV) und zu einem Drittel dem Hepatitis C Virus (HCV) zugeordnet werden [Bosch et al. 1999; Perz et al. 2006].

Pathogenese

Ein HCC kann sich auf der Basis einer Reihe von Vorerkrankungen entwickeln. Zu den bedeutendsten gehören nach [Gomaa et al. 2008]:

- Hepatitis durch Infektion mit HBV oder HCV
- Leberzirrhose beispielsweise durch Alkoholmissbrauch
- Aufnahme von Aflatoxin-B1 über die Nahrung

Diese Vorerkrankungen können beispielsweise zu einer chronischen Entzündung der Leber führen, was für die Leberzellen permanente Zyklen aus Nekrose und Regeneration bedeutet. Dabei werden schließlich karzinogene Pathways aktiviert, sodass ein HCC entstehen kann. Das HCC kann aber auch ohne vorherige Entzündung der Leber entstehen, beispielsweise bei Exposition gegenüber den mutagenen Eigenschaften von Aflatoxin-B1 [Farazi, DePinho 2006].

Man bringt bisher mehrere Ereignisse auf genetischer Ebene mit der Entwicklung des HCC in Verbindung. Dazu gehören die Inaktivierung des Tumorsuppressorgens p53, Mutationen bei β-Catenin sowie die Überexprimierung verschiedener Mitglieder der ErbB-Rezeptorfamilie und des MET-Rezeptors [Farazi, DePinho 2006].

Therapie

Sofern die Leberfunktion dies zulässt, ist die Leberresektion bei Patienten mit einem einzelnen Tumor die bevorzugte Behandlungsmethode [EASL-EORTEC 2012]. Sollte die Resektion nicht in Frage kommen, kann unter bestimmten Umständen die Lebertransplantation eine geeignete Alternative sein. Als weitere Alternative für nicht operable Patienten steht die Hochfrequenzablation zur Verfügung. Da HCC einer der am meisten chemoresistenten Tumoren ist, sind kaum systemische Therapeutika verfügbar. Lediglich für Sorafenib wurde in klinischen Studien die Wirksamkeit nachgewiesen. Jedoch vermochte Sorafenib die mediane Überlebenszeit nur auf 10,7 Monate gegenüber 7,9 Monaten in der Kontrollgruppe zu verlängern [EASL-EORTEC 2012].

Insgesamt ist die Prognose bei diagnostiziertem HCC eher ungünstig: So liegt die mediane Überlebenszeit in den USA beispielsweise bei 8 Monaten, die 1-Jahres-Überlebensrate bei 20% und die 3-Jahres-Überlebensrate bei 5% [El-Serag, Marrero et al. 2008].

2.2 DER SONDERFORSCHUNGSBEREICH/TRANSREGIO 77

Ein Sonderforschungsbereich/Transregio wird gemäß der Programmdefinition der DFG von in der Regel bis zu drei Hochschulen gemeinsam beantragt [Deutsche Forschungsgemeinschaft o. J.]. Im Fall des SFB/TRR77 sind dies am Standort Heidelberg die Universität Heidelberg und das Deutsche Krebsforschungszentrum und in Hannover die Medizinische Hochschule Hannover zusammen mit dem Helmholtz-Zentrum für Infektionsforschung in Braunschweig.

Der SFB/TRR77 hat zum Ziel, die molekularen Mechanismen, die Leberkrebs hervorrufen und begünstigen zu erforschen und ihr Potenzial zur Anwendung in Prävention, Diagnose und Therapie zu untersuchen. Somit handelt es sich um Forschung im Kontext der translationalen Medizin, die bestrebt ist, Ergebnisse aus der Grundlagenforschung möglichst schnell für den Einsatz am Patienten nutzbar zu machen. Hierzu ist der Verbund in insgesamt 26 Projekte untergliedert, von denen fünf Projekte assoziiert sind. Inhaltlich ist der SFB/TRR77 in vier Projektbereiche aufgeteilt:

PROJEKTBEREICH A untersucht die Vorgänge beim Übergang zwischen der chronischen Lebererkrankung zum Tumor

PROJEKTBEREICH B betrachtet grundlegende molekulare Mechanismen und neue Zielstrukturen

PROJEKTBEREICH C erforscht neue Therapieansätze

PROJEKTBEREICH Z stellt zentral querschnittliche Dienste einschließlich der IT-Plattform bereit

Die erste Förderperiode des SFB/TRR77 läuft vom 1. Januar 2010 bis zum 31. Dezember 2013 und hat ein Gesamtfördervolumen von rund 10 Millionen Euro [Woll et al. 2013].

2.3 BIOMEDIZINISCHE DATEN

Zum Erreichen der Ziele eines biomedizinischen Forschungsverbundes ist die Erzeugung und Verarbeitung vieler verschiedener biomedizinischer Daten erforderlich. Eine wichtige Quelle für die im Verbund zur Verfügung stehenden Daten ist ein rund 80 Patienten umfassendes anonymisiertes Kollektiv, für das HCC-Proben zur Verfügung stehen. Zu den Patienten liegen klinische Daten vor, die das Krankheitsstadium und, wenn bekannt, die Krankheitsursache charakterisieren. Bei rund 40% der Patienten konnte eine Infektion mit HBV, HCV oder beiden Viren nachgewiesen werden. Weiterhin stehen zu Kontrollzwecken zehn Leberproben, die nicht mit Tumormaterial belastet sind, sowie Leberkrebszelllinien zur Verfügung.

In den folgenden Abschnitten werden einige Verfahren vorgestellt, die bei der Analyse des Probenmaterials angewandt werden. Dabei werden insbesondere die Daten berücksichtigt, die bei den Analysen entstehen.

2.3.1 *Microarrays*

Zur Erforschung der molekularen Mechanismen sind genomische Daten von besonderem Interesse. Diese werden im Rahmen des SFB/TRR77 mit *Microarrays* gewonnen. Microarrays sind für verschiedene Analysen erhältlich. Im SFB/TRR77 werden Arrays zur Analyse der Genexpression, Copy-Number Variations (CNV), Non-coding RNA (ncRNA) und Methylierung verwendet.

Das den Microarrays zugrunde liegende Prinzip ist die Nukleinsäurehybridisierung, wie sie zwischen zwei Deoxyribonucleic Acid (DNA)-Strängen stattfindet. Für die Analysen werden einsträngige Nukleinsäure-Fragmente synthetisch hergestellt und als Sonden auf einem Trägermaterial, wie beispielsweise einem Glasträger, in Form einer Matrix aufgebracht. Die Sonden bestehen dabei aus Nukleotidsequenzen, die komplementär zu den zu detektierenden Sequenzen aus der Probe sind. Aus der Zusammenstellung der Sondensequenzen ergibt sich das Profil des Microarrays. Zur Vorbereitung des Microarray-Experiments wird die Probe entsprechend

gereinigt, aufbereitet und mit einem fluoreszierenden Farbstoff markiert. Die einsträngigen Nukleotidsequenzen der Probe werden anschließend auf das Microarray aufgebracht. Die Sequenzen der Probe binden nun an die komplementären Sequenzen der Sonden, wobei die Bindung umso stärker ist, je besser die beiden Komplemente einander ergänzen. Nach Abschluss der Hybridisierung wird das Array gewaschen und so von nicht gebundenem Probenmaterial befreit. Im letzten Schritt wird das Array zur Auswertung in einen Scanner gegeben. Dort werden die fluoreszierenden Stellen der Sondenmatrix gemessen. Stellen, deren Sonden in hohem Maße markiertes Probenmaterial binden konnten, liefern ein stärkeres Signal als solche, zu denen keine oder nur wenige komplementäre Sequenzen in der Probe vorhanden waren. Die so erzeugten Rohdaten werden mit Methoden der Bioinformatik weiter ausgewertet.

In den folgenden Abschnitten werden die im SFB/TRR77 verwendeten Microarray-Typen vorgestellt.

Array comparative genomic hybridization

Die Array comparative genomic hybridization (aCGH) ist ein Verfahren zur Erkennung von Veränderungen in der Kopienzahl von DNA-Abschnitten in einer Probe. Bei der Entstehung von Krebs finden häufig Veränderungen in der Struktur der Chromosomen statt, sodass bestimmte Sequenzen, die auch Gene enthalten können, im Krebsgenom häufiger oder seltener vorhanden sind als im gesunden Genom des Patienten.

Bei Microarrays für aCGH werden die DNA aus dem Tumor und die Referenz-DNA (aus gesundem Gewebe) mit Fluoreszenzfarbstoffen unterschiedlicher Farbe markiert und gemeinsam auf das entsprechend ausgelegte Microarray aufgetragen. Die Nukleotidsequenzen der Sonden sind dabei mit 100 bis 200 Kilo-Basenpaaren so lang, dass ein oder mehrere Gene darin enthalten sein können. Aus dem vom Scanner abgetasteten Farbsignal kann dann auf die jeweilige Kopienzahl geschlossen werden. Nach der bioinformatischen Aufbereitung der Daten entsteht eine Liste, die für jede Sondenposition angibt, ob dort eine Abweichung in der Kopienzahl festgestellt wurde und falls ja, ob es sich um den Gewinn oder Verlust von Sequenzmaterial handelt. Dies wird häufig durch die Werte -1, 0 und 1 kodiert.

Genexpression

Bei der Genexpressionsanalyse ist es das Ziel festzustellen, welche Gene in einer Zelle zu einem bestimmten Zeitpunkt aktiv sind, das heißt, welche Gene gerade von der DNA zur Messenger RNA (mRNA) transkribiert werden. Um dies zu messen, wird ein Microarray-Typ verwendet, der in der Lage ist, die Menge bestimmter mRNA-Sequenzen zu bestimmen. Dazu wird zunächst die mRNA aus der Probe extrahiert und in Complemen-

tary DNA (cDNA) transkribiert. In der Folge kann analog zu den oben beschriebenen DNA-Microarrays verfahren werden.

Nach dem Scannen des Arrays und der bioinformatischen Aufbereitung liegt in der Regel eine Liste mit den betrachteten Genen und jeweils einem Messwert vor. Mit diesen Messwerten können dann weitere Analysen durchgeführt und die Expressionsniveaus verschiedener Proben verglichen werden.

Non-coding RNA

ncRNA ist ein Ribonucleic Acid (RNA)-Molekül, welches nicht in ein Protein übersetzt wird. Allerdings nimmt ncRNA wichtige Aufgaben bei der Steuerung in zellulären Vorgängen wahr. Das Prinzip der Microarrayanalyse ist bei der ncRNA analog zur Genexpressionsanalyse. Auch hier wird nach Abschluss der bioinformatischen Aufbereitung eine Liste mit ncRNA-Namen und den zugehörigen Messwerten erstellt.

Methylierung

Die Nukleotide der DNA können chemisch modifiziert werden. Hierzu werden einzelne Nukleotide mit einer Methylgruppe ergänzt. Eine wesentliche Funktion der Methylierung ist es, die Aktivität eines stromabwärts gelegenen Gens selektiv zu regulieren: Die Methylierung bewirkt, dass das Gen weniger oft ausgelesen und somit in seiner Aktivität gehemmt wird, ohne dass die Abfolge der Basenpaare im DNA-Strang verändert werden muss. Auch zur Erfassung des Methylierungsprofils einer Probe stehen spezifische Microarrays zur Verfügung [Schumacher et al. 2006]. Das Methylierungsprofil wird als Liste, bestehend aus den Namen der assoziierten Gene und den Messwerten der Methylierung, angegeben.

2.3.2 *Tissue Microarrays*

Tissue Microarrays (TMA) unterscheiden sich von den bisher vorgestellten Microarrays insofern, als dass dabei keine genomischen Daten gewonnen werden. Vielmehr handelt es sich dabei um histologische Schnitte mit einem Durchmesser von rund 0,6 mm die gemeinsam auf einem Glasträger aufgebracht werden. Diese Schnitte sind in einer Matrix angeordnet und erinnern so an genomische Microarrays. Durch die kleine Menge an Probenmaterial tragen Tissue Microarray (TMA) dazu bei, mit wertvollen Tumorproben sparsam umzugehen. TMA werden wie herkömmliche Objektträger mikroskopisch befundet.

2.3.3 *Pathways*

Pathways beschreiben biologische Prozesse in Zellen. In Krebszellen sind die Pathways in der Regel gegenüber gesunden Zellen verändert, das heißt, es werden Prozesse aktiviert oder unterdrückt, die für einen gesunden Lebenszyklus der Zelle erforderlich sind. Die Kommunikation zwischen den Zellen wird mit sogenannten Signaling Pathways beschrieben. Auch hierbei kommt es in Krebszellen oft zu Störungen, sodass sie beispielsweise nicht mehr auf Signale reagieren, die Apoptose und damit den kontrollierten Zelltod einzuleiten. Von vielen Genen ist bekannt, dass sie mit Pathways in Verbindung stehen. Daher werden die Pathways in entsprechenden Datenbanken dokumentiert. Die bekanntesten Pathwaydatenbanken sind Kyoto Encyclopedia of Genes and Genomes (KEGG) und Gene Ontology (GO) [Ashburner et al. 2000; Kanehisa, Goto 2000].

2.3.4 *Datenhaltung*

Die gemeinsam genutzten Daten eines biomedizinischen Forschungsverbundes können entweder zentral oder dezentral gespeichert und verwaltet werden. Beide Varianten haben spezifische Vor- und Nachteile, die entsprechend der konkreten Anforderungen zu bewerten sind. Werden die Daten dagegen nur lokal von einzelnen Forschern genutzt, so werden die Daten häufig in Tabellenform gespeichert.

Zentrale Datenhaltung

Bei der zentralen Datenhaltung werden alle Daten eines Verbundes in einer gemeinsamen Datenbank zusammengeführt. Dadurch ist es einfach, die Struktur der Daten zu kontrollieren, indem diese beispielsweise beim Importvorgang auf ein zentrales Format abgebildet werden. Eine projektübergreifende Auswertung der Daten ist ohne Schwierigkeiten möglich, da sich diese bereits in derselben Datenbank befinden.

Andererseits ist es schwieriger, die Zugriffsberechtigungen auf die Daten der Projekte individuell und gleichzeitig transparent zu regeln. Die Projekte verlieren in diesem Fall die unmittelbare und physische Kontrolle über die Daten, die sie dem Verbund zur Verfügung stellen. Sie müssen daher der zentralen Plattform vertrauen, dass die Daten auch verbundintern nur in berechtigter Weise zugänglich gemacht werden.

Dezentrale Datenhaltung

Bei einem dezentralen Ansatz zur Datenhaltung werden die Daten in verschiedenen Datenbanken gespeichert. Für eine übergreifende Auswertung werden dann die Daten einiger oder aller Datenbanken gemeinsam abgefragt. Man spricht in diesem Fall von *föderierten Datenbanken*. Durch das

Zusammenschließen der Datenbanken zur Abfragezeit entsteht an dieser Stelle ein erhöhter Verwaltungsaufwand, da gegebenenfalls dynamische Konvertierungen zwischen den Formaten der Datenbanken erforderlich sind. Weiterhin ist eine Infrastruktur erforderlich, die das Auffinden der Datenbanken ermöglicht, sowie die semantische und syntaktische Beschreibung des Datenformats bereitstellt.

Da die Daten dezentral vorgehalten werden, ist es bei diesem Ansatz jedem Projekt möglich, eine eigene Datenbank zu betreiben und die Daten dem Verbund bei voller Kontrolle über die eigenen Daten bereitzustellen. Die Daten können auch physikalisch beim Projekt verbleiben, sodass der Kontrollaspekt auch von weniger technisch orientierten Projektmitarbeitern gut nachzuvollziehen ist.

Excel

In der biomedizinischen Forschung werden Daten häufig in Form von Tabellen bearbeitet. Weit verbreitet sind dabei Dateien im Format der Tabellenkalkulationssoftware Microsoft Excel. Dabei werden oft die Ergebnisse der Analysen mehrerer Proben einander gegenübergestellt. Bei Microarray-Daten werden in der Regel die Messwerte der einzelnen Messpunkte in mehreren Tausend Zeilen aufgetragen. Die einzelnen Proben, welche beispielsweise von unterschiedlichen Patienten stammen können, werden in Spalten dargestellt. Bisweilen werden noch zusätzliche Spalten angefügt, die Annotationen zum Microarray enthalten. Dies können beispielsweise die Position eines Gens auf dem Chromosom, assoziierte Pathways oder Referenzen zu Bioinformatikdatenbanken sein.

2.4 ZELLLINIEN

Zelllinien sind etablierte *in vitro* Werkzeuge der biomedizinischen Forschung. Gewöhnlich werden Zelllinien aus menschlichen oder tierischen Zellen gewonnen, die einen unsterblichen Phänotyp entwickelt haben. Im Gegensatz dazu sind Primärzellen nicht unsterblich und können daher *in vitro* nur begrenzt kultiviert werden [Hayflick, Moorhead 1961]. Unsterbliche Zellen werden entweder aus Krebsgewebe gewonnen oder sie werden durch biotechnische Verfahren erzeugt, indem beispielsweise die Verkürzung der Telomere an der DNA unterbunden wird [Ouellette 2000].

Verschiedene Zelllinien, die von einem bestimmten Zelltyp abstammen, besitzen gewöhnlich ähnliche biologische Eigenschaften. Daher sind Zelllinien standardisierte Materialien für die *in vitro*-Forschung. Da Standardisierung und Zuverlässigkeit die Basis für biologische und medizinische Forschung darstellen, werden vereinheitlichte Zelllinien von Zellbanken vertrieben. Die international bekannteste Zellbank ist die American Type Culture Collection (ATCC).

Im Gegensatz zu den Zelllinien selbst erscheinen die Metadaten, mit deren Hilfe die Zelllinien beschrieben werden, sehr uneinheitlich. Gleichwohl gibt es einige Aktivitäten mit dem Ziel, die Namen der Zelllinien aus den Katalogen der Zellbanken zu extrahieren und in biomedizinischen Ontologien verfügbar zu machen [Romano et al. 2009; Sarntivijai et al. 2008]. Allerdings ist die Kenntnis des Zelllinniennamens nicht ausreichend, vielmehr müssen auch die Kulturbedingungen, wie Wachstumsmedium, Kulturzusätze, Inkubationstemperatur und zugehörige Protokolle betrachtet werden. Diese Parameter werden im Zuge der Etablierung einer Zelllinie in einem Labor bisweilen den lokalen Erfordernissen angepasst. Um die Reproduzierbarkeit durchgeführter Experimente zu gewährleisten, müssen solche angepassten Parameter kontrolliert und dokumentiert werden. Während die Datenbanken der Zellbanken gewöhnlich die Standardparameter enthalten, werden sie von Ontologien mit Bezug zu Zelllinien nur sporadisch berücksichtigt.

Werden Daten zu Zelllinien nicht in einer gewöhnlichen Datenbank gespeichert sondern in Form einer Ontologie, so wird dadurch die semantische Integration dieser Daten mit anderen Forschungsdaten unterstützt. Somit können durch den Einsatz von Reasoner-Programmen automatisiert neue Erkenntnisse abgeleitet werden. Zwei Beispiele für derart abgeleitetes Wissen sind die folgenden:

- Identifikation von Zellkulturbedingungen für einen neuen Versuch auf Basis der Kulturbedingungen verwandter Zelllinien
- Identifikation von Zelllinienindividuen, die für ein bestimmtes Experiment geeignet sind, aus den Ergebnissen früherer Experimente

2.5 SERVICEORIENTIERTE ARCHITEKTUREN (SOA)

Eine SOA ist ein Entwurfsansatz für die IT-Architektur von größeren verteilten Anwendungen mit dem Ziel, die Funktionen der Anwendung in eigenständige Komponenten aufzugliedern. Damit unterstützt eine SOA das Bestreben, in der Unternehmens-IT monolithische Systeme, die für einen spezifischen Zweck erstellt wurden, durch flexible, auf abgeschlossenen Komponenten basierende IT-Systeme zu ersetzen.

Je nach der Intention der Autoren treten unterschiedliche Aspekte bei der Definition des Begriffs SOA in den Vordergrund. Im SOA-Referenzmodell werden die zentralen Eigenschaften dieses Ansatzes folgendermaßen beschrieben: „**Service Oriented Architecture** (SOA) is a paradigm for organizing and utilizing distributed **capabilities** that may be under the control of different ownership domains.“ [OASIS SOA-RM, 2006] Die Funktionen der verteilten Anwendung werden dementsprechend nicht notwendigerweise aus einer Hand bereitgestellt, sondern von verschiede-

nen Organisationseinheiten. Solche Funktionen werden als *Dienste* oder englisch als *Services* bezeichnet.

Voraussetzung für das Verknüpfen von Diensten verschiedener Anbieter ist die präzise fachliche Beschreibung der Leistungen der Dienste. Eine feste Verknüpfung der Dienste miteinander erfolgt dabei nicht, vielmehr erfolgt die Auswahl und Einbindung der konkreten Serviceinstanz erst zu dem Zeitpunkt, an dem der Dienst tatsächlich genutzt wird. Man spricht daher auch von *loser Kopplung* der Dienste, da die Komponenten nicht gemeinsam kompiliert werden und austauschbar sind, sofern Dienste von alternativen Anbietern verfügbar sind [Richter et al. 2005].

Somit ergibt sich, wie in Abbildung 1 dargestellt, eine der einfachsten und ältesten Darstellungen des Konzeptes einer SOA. Diese Darstellung besteht aus drei Elementen [Bih 2006; Endrei et al. 2004]:

- Service-Konsument: Nutzt einen konkreten Dienst, dessen Adresse zur Laufzeit aus der Registratur abgerufen wird
- Service-Anbieter: Bietet einen Dienst an, dessen Beschreibung und Endpunktadresse in der Registratur veröffentlicht wird
- Service-Registratur: Hält Beschreibungen zu Diensten bereit und vermittelt so zwischen Anbieter und Konsument

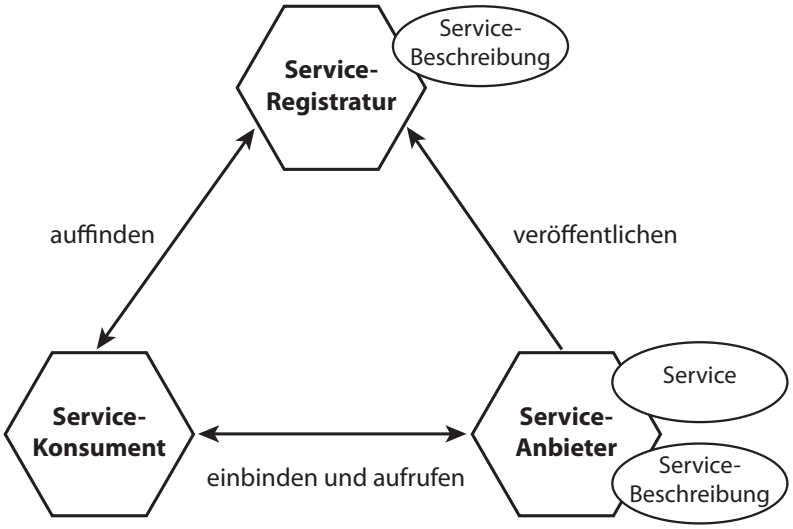


Abbildung 1: Ursprüngliches SOA-Paradigma nach [Endrei et al. 2004]

Dieses ursprüngliche Modell einer SOA wird jedoch von Stal als nicht hinreichend angesehen, da es die SOA nicht ausreichend von anderen ver-

teilten Middleware-Ansätzen abgrenzt [Stal 2006]. Unter *Middleware* versteht man bei verteilten Anwendungen die Software, welche das Bindeglied zwischen Clients und Servers darstellt [Tresch 1996]. Stal schlägt vor, folgende Aspekte im Sinne einer SOA genauer zu spezifizieren [Stal 2006]:

- Schnittstellen und Verträge: Wie können Dienstanbieter die Dienste und zugehörigen Verträge zur Nutzung beschreiben? Wie können Nutzer diese Beschreibungen und Verträge verstehen und die Dienste nutzen?
- Kommunikation: Welche Kommunikationsstile werden für die Interaktion mit den Diensten angeboten und was sind Inhalt und Bedeutung der Kommunikation?
- Auffinden und Registrieren der Dienste: Wie können Anbieter ihre Dienste bekannt machen und Nutzer diese auffinden?
- Zustand und Aktivierung: Wie geht die Infrastruktur mit Fragen des Zustandes und der Aktivierung um, vor allem im Hinblick auf zustandslose Kommunikationsprotokolle?
- Prozesse und ihre Implementierungen: Wie kann die SOA-Infrastruktur die Nutzer der Dienste durch Koordination und Orchestrierung unterstützen, verschiedene unabhängige Dienste zu einem vollständigen Prozess zu verknüpfen?

Im Verlauf der weiteren SOA-Entwicklung wurde das Architekturparadigma kontinuierlich fortgeschrieben. In der neueren Literatur beschreibt Erl acht Prinzipien, die aus seiner Sicht die theoretische Grundlage der SOA bilden [Erl 2008]:

1. Standardisierter Servicevertrag: Mit *Vertrag* ist hier kein rechtliches Dokument, sondern eine verbindliche technische Beschreibung der Diensteigenschaften gemeint. Als grundlegender Teil der Serviceorientierung muss die öffentliche Schnittstelle, welche die Beschreibung der Funktionalität, der Datentypen und des Datenmodells beinhaltet, in geeigneter Granularität definiert und fortgeschrieben werden.
2. Lose Kopplung: Um die unabhängige Entwicklung von Dienstinstanzen zu fördern, ist es erforderlich, die Abhängkeiten zwischen Dienstkonsumenten auf ein sinnvolles Maß zu beschränken.
3. Abstraktion von Services: Um eine lose Kopplung zu erreichen ist die Abstraktion der Dienste erforderlich, indem die internen Implementierungsdetails der Dienste vor den Nutzern verborgen werden.

4. Wiederverwendbarkeit von Services: Die Wiederverwendbarkeit von Diensten ist ein zentraler Bestandteil der SOA. Um diese zu ermöglichen, müssen die Dienste in einer geeigneten Granularität entworfen werden und in andere Kontexte übertragbar sein.

5. Autonomie von Services: Damit Dienste in verschiedenen Kontexten funktionieren können, müssen sie unabhängig sein. Dazu müssen sie die Kontrolle über die von ihnen benötigten Ressourcen ausüben können.

6. Zustandslosigkeit von Services: Dienste sollten möglichst zustandsfrei sein, da der mit der Zustandsverwaltung einhergehende Verwaltungsaufwand die Skalierbarkeit der Dienste beeinträchtigt.

7. Auffindbarkeit von Services: Um die Wiederverwendung von Diensten voranzutreiben müssen diese leicht auffindbar sein.

8. Kompositionsfähigkeit von Services: Die Fähigkeit, zu Prozessketten zusammengefügt zu werden ist zentraler Bestandteil der SOA-Idee. Daher muss die Kompositionsfähigkeit bei jedem Dienst vom Entwurf an berücksichtigt werden.

2.5.1 *Technische Umsetzung*

Die technische Kommunikation zwischen den Diensten wird durch die Verwendung von standardisierten Mechanismen gekapselt. Obwohl eine SOA keine bestimmte Technologie vorschreibt und daher auch mit Methoden wie Common Object Request Broker Architecture (CORBA) [ISO/IEC 19500-2, 2012] umgesetzt werden könnte, ist die SOA eng mit den so genannten Webservices verbunden. Unter Webservices versteht man eine Familie von Standards, welche Schnittstellen, Datenaustauschformate und Eigenschaften ihrer Implementierung beschreiben. Sie umfassen weiterhin Aspekte der Sicherheit, der Registrierung und Komposition der zugehörigen Komponenten [Kossmann, Leymann 2004].

Webservices werden durch die Empfehlungen des World Wide Web Consortium (W3C) spezifiziert. Wesentliche Basis für Webservices ist die Extensible Markup Language (XML) [W3C XML, 2006], auf deren Basis sowohl die Beschreibungen der Dienste als auch die Datenaustauschformate beruhen. Für die Interaktion zwischen Serviceanbieter und Servicekonsument werden Simple Object Access Protocol (SOAP)-Nachrichten ausgetauscht [W3C SOAP, 2007]. Dabei handelt es sich um standardisierte XML-Dokumente, welche die Serviceanfrage oder die zugehörige Antwort enthalten. Die Übertragung der SOAP-Nachrichten erfolgt meist über das Hypertext Transfer Protocol (HTTP) unter Verwendung eines Uniform Resource Identifier (URI), jedoch sind auch andere Mechanismen zur Nachrichtenübermittlung wie beispielsweise E-Mail möglich.

Weiterhin existieren Representational State Transfer (REST)-konforme Webservices, die ebenfalls HTTP als Übertragungsprotoll und URI zum Auffinden der entsprechenden Ressourcen verwenden [Fielding 2000]. REST-Webservices besitzen Schnittstellen mit einheitlicher Semantik, welche im Wesentlichen die Funktionen Erstellen, Abrufen, Aktualisieren und Löschen für die zu Grunde liegenden Objekte bereitstellen [W3C WSA, 2004].

Serviceorientierung bezieht sich aber nicht notwendigerweise ausschließlich auf IT. Vielmehr kann der Begriff auf die Bereitstellung von Ressourcen oder Dienstleistungen im Allgemeinen ausgeweitet werden.

2.5.2 *Dienstekomposition*

Aus der Kompositionsfähigkeit der Dienste einer SOA ergibt sich das Erfordernis, den Vorgang der dynamischen *Dienstekomposition* mit Prozessen und Werkzeugen zu unterstützen [Milanovic, Malek 2004]. Im Bereich der biomedizinischen Forschung werden wissenschaftliche Workflow-Systeme verwendet, um beispielsweise die Prozesse zur Analyse von genomischen Daten in einer Pipeline zusammenzufassen. Oft werden solche Pipelines von Rechnerverbünden, auch *Grids* oder *Cluster* genannt, abgearbeitet [Foster et al. 2001; Oster et al. 2007]. Dabei werden die Aufgaben automatisch auf verschiedene Rechnerressourcen verteilt und, wenn möglich, parallel abgearbeitet.

[Cui Lin et al. 2009] beschreiben eine Referenzarchitektur für wissenschaftliche Workflow-Systeme und bewerten verfügbare Systeme anhand der Referenzarchitektur. Auch [Talia 2013] betrachtet aktuelle wissenschaftliche Workflow-Systeme und schließt dabei auch die zur Familie der Webservices-Standards gehörige Business Process Execution Language (BPEL) mit ein [OASIS BPEL, 2007].

2.5.3 *Sicherheit in serviceorientierten Architekturen*

Die in einer SOA zu berücksichtigenden Sicherheitsaspekte sind vielschichtig. Von besonderer Bedeutung sind jedoch die gesicherte *Datenübertragung*, sowie die *Authentisierung* und die *Autorisierung* der Nutzer.

Die gesicherte Datenübertragung wird dabei entweder durch Verschlüsselung der ausgetauschten Nachrichten selbst oder des Übertragungskanals zwischen den Kommunikationspartnern erreicht [Naedele 2003]. Die Verschlüsselung der Nachrichten im Rahmen von Webservices wird durch die Spezifikation der Web Services Security festgeschrieben [OASIS WSS, 2006].

Da SOA-Systeme verteilt sind, können Authentisierungs- und Autorisierungsentscheidungen ebenfalls oft nicht an einer Stelle getroffen werden,

vielmehr werden diese Entscheidungen in der Regel von technisch und gegebenenfalls auch organisatorisch getrennten Komponenten der SOA-Umgebung getroffen. In einem gängigen Szenario melden sich die Benutzer eines SOA-Systems an einem Zugangspunkt des Systems, beispielsweise einem Webportal, an. Dort wird der Benutzer *authentisiert* indem seine Identität beispielsweise anhand von Benutzername und Passwort überprüft wird. Möchte der Benutzer nun einen anderen Dienst des SOA-Systems nutzen, so wird seine beglaubigte Identität in Form eines kryptografischen Tokens an diesen Dienst weitergeleitet [OASIS SAML, 2005].

Der aufgerufene Dienst vertraut dem Aussteller dieses Tokens bezüglich der beglaubigten Identität. Die Entscheidung, ob die entsprechende Person den Dienst nutzen darf, also *autorisiert* ist, trifft jedoch eine organisatorisch zum Eigentümer des Dienstes gehörige SOA-Komponente, der *Policy Decision Point (PDP)* [Dürbeck et al. 2011; OASIS XACML, 2005]. Der PDP meldet seine Entscheidung an den Dienst zurück. In diesem Zusammenhang stellt der Dienst auch einen *Policy Enforcement Point (PEP)* dar, indem er die Anfrage des Nutzers akzeptiert oder abweist. Zur Verwaltung der Zugriffsregeln wird ein *Policy Administration Point (PAP)* verwendet, der die Regeln für den PDP bereitstellt.

Somit kann eine SOA bezüglich der Zugriffskontrolle so gestaltet werden, dass die im Verbund angebotenen Dienste organisatorisch getrennt sind. Die Entscheidung, welche Benutzer Zugriff zu einem Dienst bekommen, kann ebenfalls pro Dienst unterschiedlich sein, sodass die Eigentümer der Dienste die Hoheit über ihren Beitrag zum SOA-System behalten können. Die in diesem Abschnitt aufgezeigten Sicherheitsmechanismen sind jedoch nur ein Ausschnitt der Aspekte, die für eine sichere SOA berücksichtigt werden müssen. Eine umfangreiche Betrachtung kann beispielsweise dem SOA-Security-Kompendium entnommen werden [Bundesamt für Sicherheit in der Informationstechnik 2009].

2.5.4 *Frameworks für die medizinische Forschung*

Speziell für den Einsatz in der (bio)medizinischen Forschung wurden Frameworks entwickelt, die bei der Erstellung von SOA-Systemen unterstützen. In den folgenden Abschnitten werden zwei besonders relevante Frameworks vorgestellt.

i2b2

Das i2b2-Framework zielt darauf ab, klinische Daten mit Forschungsdaten zu verknüpfen und so für die translationale Forschung nutzbar zu machen [Ganslandt et al. 2011]. Nach außen hin bietet i2b2 Schnittstellen zur Einbindung in eine SOA-Umgebung an. Als Kern besitzt i2b2 ein zentrales *Data Warehouse*, es verfolgt somit den Ansatz der zentralen Datenhaltung

(siehe Abschnitt 2.3.4). In einer solchen Datenbank werden die Daten aus den verschiedenen zu berücksichtigenden Quellen zusammengeführt und gespeichert, wobei eine Optimierung hinsichtlich der Auswertbarkeit der Daten angestrebt wird [Gluchowski 1997]. Hierzu werden besondere, nicht normalisierte Datenbankschemata wie das Sternschema verwendet [Ponniah 2010]. Zur Verwendung im Data Warehouse werden die Daten aus dem Quellsystem *extrahiert*, in das Schema des Data Warehouses *transformiert* und anschließend in das Data Warehouse geladen. Man spricht daher auch vom Extract, Transform, and Load (ETL)-Prozess, der zyklisch wiederholt wird, um die Daten zu aktualisieren [Ponniah 2010].

Um das Data Warehouse herum werden weitere i2b2-Funktionen in Form von Webservices bereitgestellt. Die so entstehende SOA wird als i2b2-*Hive* bezeichnet. Weitere Dienste mit zusätzlichen Funktionen können auch von externen Entwicklern als Zellen hinzugefügt werden [Murphy et al. 2010].

caBIG

caBIG ist ein Projekt, das vom National Cancer Institute (NCI) in den Vereinigten Staaten von Amerika in den Jahren 2004 bis 2012 durchgeführt wurde. Ziel des Projektes war es, die Zusammenarbeit zwischen verteilten Forschungsorganisationen zu fördern. Im Rahmen von caBIG wurden über 70 verschiedene quelloffene Programme für verschiedene biomedizinische Fragestellungen entwickelt [Fenstermacher et al. 2005]. Nicht alle von diesen Programme wurden für caBIG neu entwickelt. Einige bestehende Anwendungen wurden mit Hilfe von caBIG-Entwicklungswerkzeugen so angepasst, dass sie nun als Dienste in der SOA-Umgebung von caBIG zur Verfügung stehen [Krikov et al. 2011]. Während caBIG als Programm von seinem Nachfolger, dem National Cancer Informatics Program (NCIP) abgelöst wurde, ist die entstandene Software weiterhin verfügbar [Varmus 2012].

caBIG verfolgt die Strategie, Ressourcen in der Krebsforschung in einem *föderierten* Modell nutzbar zu machen. Somit ist die Datenhaltung in caBIG verteilt und die Verantwortung für die Integrität und Sicherheit der Daten verbleibt bei den ursprünglichen Besitzern der Daten [Komatsoulis 2011; Saltz, Oster, Hastings, Langella, Ferreira et al. 2008]. Dies wird durch die Etablierung einer SOA ermöglicht. Um die gemeinsame Nutzung von Rechnerressourcen zu unterstützen, baut caBIG auf das Cancer Grid (caGRID) auf [Oster et al. 2007; Saltz, Oster, Hastings, Langella, Kurc et al. 2006]. Über diesen Verbund können Rechenaufträge abgearbeitet werden und föderierte Datenabfragen in der caGrid Query Language (CQL) abgearbeitet werden [McCusker et al. 2009]. Datenquellen werden als *Dataservices* mit einer definierten Webservice-Schnittstelle bereitgestellt. Nicht alle caBIG-Werkzeuge unterstützen die Verwendung der caGRID-Infrastruktur. Um-

gekehrt können diejenigen Programme, welche caGRID nutzen können in der Regel auch ohne das Grid verwendet werden.

Um die semantische Interoperabilität zwischen den Diensten zu gewährleisten, wird ein zentraler Verzeichnisdienst, das Cancer Data Standards Registry and Repository (caDSR) betrieben [Kunz et al. 2009]. Das caDSR ist konform mit der ISO-Spezifikation zu Metadatenverzeichnissen in der Edition 2 [ISO/IEC 11179-1:2004(E), 2004]. Dort werden sämtliche Metadaten abgelegt, die zum Auffinden und Verknüpfen der caBIG-Dienste erforderlich sind. Zur vereinfachten Entwicklung von Diensten und deren Integration in das caDSR wird das Datendienst-Framework *Cancer Common Ontologic Representation Environment (caCORE) Software Development Kit (SDK)* bereitgestellt [Komatsoulis et al. 2008]. Beispielsweise werden in caDSR auch umfangreiche Beschreibungen von gemeinsamen Datenelementen und Modelle zur Dokumentation der Beziehungen zwischen den Datenelementen verwaltet. Der caDSR-Dienst wird ergänzt durch ein Werkzeug namens caAdapter zur Durchführung der erforderlichen Abbildung der Metadaten auf das Modell des jeweiligen Dienstes.

Die SOA-Sicherheitsmechanismen werden in caGRID durch die *Grid Authentication and Authorization with Reliably Distributed Services (GAARDS)* umgesetzt [Langella, Hastings et al. 2008; Langella, Oster et al. 2007]. Dieses System ermöglicht die effiziente Verwaltung von Benutzeridentitäten und Identitäten im föderierten Umfeld von caBIG. Auch GAARDS muss nicht zwingend, beziehungsweise nicht für jede Software, verwendet werden, die im Rahmen des caBIG-Projektes bereitgestellt wird. Weiterhin stellt caCORE SDK das Common Security Module (CSM) bereit [Langella, Oster et al. 2007]. Dieses Modul ist eine eigenständige Komponente, die sowohl die Authentisierung als auch die Autorisierung der Benutzer auf lokaler Ebene durchführen kann.

2.5.5 *Portalserver*

Portalserver werden häufig als Benutzerschnittstelle für SOA-Systeme verwendet. Die Anwender benutzen dabei Webbrowser um den Portalserver über die Netzwerkprotokolle des Internets zu erreichen. Portale bieten die Möglichkeit, einzelne Komponenten, sogenannte *Portlets*, für spezielle Aufgaben zu entwickeln. Der Portalserver übernimmt dann die Anzeige der Portlets und erlaubt es dem Anwender, mehrere Portlets auf einer Seite zusammenzustellen. Je nach Ausprägung der Portlets können diese durch den Anwender in mehr oder weniger großem Umfang konfiguriert und so den eigenen Bedürfnissen angepasst werden.

Portlets sind standardisierte Java-Programme, die durch den Java Specification Request (JSR) normiert sind [JSR 286, 2008]. Sie können in verschiedenen Portalserver-Produkten, die diesen Standard umsetzen, ausgeführt werden.

2.6 METADATEN

In [Haber et al. 2007] wird die große Vielfalt an verwendeten Begriffen für ähnliche oder identische Konzepte als großes Problem in der biomedizinischen Forschung benannt. Um diesem Problem zu begegnen, ist es erforderlich, die Annotation der Daten und Datenquellen zu standardisieren. Die standardisierte Annotation ist gleichzeitig eine Voraussetzung für die automatisierte Verknüpfung sowie das Auffinden geeigneter Dienste in einer SOA.

In den folgenden Abschnitten werden die für eine Standardisierung der Metadaten erforderlichen Begriffe eingeführt.

2.6.1 *Definitionen*

In diesem Abschnitt werden Begriffe, die im Zusammenhang mit Metadaten relevant sind, definiert. Wo immer möglich wird hierzu die Definition des Deutschen Instituts für Normung (DIN) bzw. der International Organization for Standardization (ISO) herangezogen.

Gegenstand, Begriff und Benennung

Ein *Gegenstand* oder *Objekt* ist gemäß der DIN 2342 ein „Beliebiger Ausschnitt aus der wahrnehmbaren oder vorstellbaren Welt“ [DIN 2342, 1992]. Ein Gegenstand muss nicht materieller Natur sein, sondern kann zum Beispiel auch ein Sachverhalt oder ein Ereignis sein.

Weiterhin definiert die DIN 2342 den *Begriff* als „Denkeinheit, die aus einer Menge von Gegenständen unter Ermittlung der diesen Gegenständen gemeinsamen Eigenschaften mittels Abstraktion gebildet wird“ [DIN 2342, 1992]. Begriffe sind abstrakte Konzepte, welche die gedankliche Vorstellung von Objekten beschreiben, jedoch nicht direkt eine sprachliche *Benennung* darstellen.

Die *Bezeichnung* ist nach DIN 2342 die „Repräsentation eines Begriffs mit sprachlichen oder anderen Mitteln“ [DIN 2342, 1992]. Somit kann die Bezeichnung sowohl durch Symbole als auch durch Worte erfolgen. Im Rahmen dieser Arbeit ist die verbale Bezeichnung von besonderer Bedeutung. Die DIN 2342 definiert dazu die *Benennung* als „Aus einem Wort oder mehreren Wörtern bestehende Bezeichnung“ [DIN 2342, 1992]. Die Gesamtheit aller Begriffe und Benennungen eines Fachgebietes ist die *Terminologie* [DIN 2342, 1992].

Kontrolliertes Vokabular

Das *kontrollierte Vokabular* stellt nach dem American National Standards Institute (ANSI)/National Information Standards Organization (NISO) [NISO 2005] ein Werkzeug dar, um Informationen zu organisieren. In

einem solchen, von einer Organisation herausgegebenen und gepflegten Vokabular werden alle zulässigen Benennungen explizit aufgeführt und mit einer möglichst eindeutigen, nicht-redundanten Definition versehen. Sollten Homonyme oder Polyseme zu den Begriffen des Vokabulars existieren, so müssen diese entsprechend berücksichtigt werden, sodass Mehrdeutigkeit vermieden wird.

Die Benennungen des kontrollierten Vokabulars können nach dem ANSI beziehungsweise NISO-Standard klassiert und in einer *Taxonomie* hierarchisch angeordnet werden [NISO 2005]. Dadurch werden die Begriffe derart zueinander in Beziehung gesetzt, dass die hierarchisch tiefer stehenden Begriffe jeweils Spezialisierungen der übergeordneten Begriffe sind (Hierarchierelation). Sind für einen Begriff synonyme Benennungen vorhanden, so müssen diese erfasst und der Begriff durch eine Vorzugsbenennung (*Deskriptor*) eindeutig bezeichnet werden. Eine Taxonomie ist somit eine spezielle Form eines kontrollierten Vokabulars.

Thesaurus

Weiterhin definiert die DIN 1463 den *Thesaurus*: „Ein Thesaurus im Bereich der Information und Dokumentation ist eine geordnete Zusammenstellung von Begriffen und ihren (vorwiegend natürlichsprachigen) Bezeichnungen, die in einem Dokumentationsgebiet zum Indexieren, Speichern und Wiederauffinden dient“ [DIN 1463, 1987]. Ein Thesaurus hat die gleichen Eigenschaften wie eine Taxonomie, zusätzlich zur Hierarchierelation können jedoch noch die Relationsarten *Abstraktion, Äquivalenz, Assoziation* und *Bestand* zwischen den Begriffen definiert werden [DIN 1463, 1987].

Ontologie

In einer *Ontologie* schließlich können die Begriffe eines Vokabulars auch nicht hierarchisch in einem Netzwerk miteinander verknüpft werden [Gruber 1993; Guarino et al. 2009]. Zusätzlich können allgemeingültige Aussagen (Axiome), die nicht aus den Begriffen abgeleitet werden können, formuliert und in der Ontologie abgelegt werden. Somit stellt die Ontologie umfangreichere Werkzeuge zur Modellierung des Wissens in einem Themengebiet bereit.

Eine übergeordnete Ontologie, (englisch *upper [level] ontology*), stellt grundlegende Konzepte bereit, die auch über ein spezifisches Themengebiet hinaus gültig sind. Durch den Einsatz übergeordneter Ontologien lässt sich ein Rahmen schaffen, der die Integration und Zusammenführung themenspezifischer Ontologien erleichtert.

2.6.2 *Metadatenformate*

Für die standardisierte Darstellung von Metadaten stehen verschiedene Standards zur Verfügung. In den folgenden Abschnitten werden einige für diese Arbeit relevante Formate vorgestellt.

Resource Description Framework

Das Resource Description Framework (RDF) ist ein Standard des W3C, welcher ein Modell zum Austausch von Daten über das World Wide Web spezifiziert [W3C RDF, 2004]. Das Datenmodell basiert auf Aussagen, die aus drei Einheiten bestehen: Je einem Subjekt, Prädikat und Objekt. Durch diese Tripel entstehen gerichtete Grafen, deren Knoten zwei Entitäten sind und deren Kanten bezeichnete Beziehungen zwischen den Entitäten sind.

Simple Knowledge Organization System

Besonders relevant unter den Metadatenformaten ist das Simple Knowledge Organisation System (SKOS) [W3C SKOS, 2009a]. Das SKOS-Datenmodell stellt eine relativ einfache, standardisierte Spezifikation bereit, um ein Vokabular formal darzustellen und auszutauschen. SKOS baut auf RDF auf, daher werden Konzepte als RDF-Ressourcen gespeichert.

Web Ontology Language

Die Web Ontology Language (OWL) ist eine Beschreibungssprache, mit der die Begriffe einer Domäne sowie ihre Beziehungen untereinander formal so beschrieben werden können, dass sie auch maschinell ausgewertet werden können. OWL wird ebenfalls vom W3C spezifiziert [W3C OWL, 2009]. Technisch greift OWL wie auch SKOS auf die RDF-Syntax zurück, sodass die Ontologie in XML-Form vorliegt.

2.6.3 *Metadatenverzeichnis*

Ein Metadatenverzeichnis ist eine zentrale Datenbank in einer Organisation, mit der die Metadaten verwaltet und bereitgestellt werden. In einem solchen Verzeichnis werden sowohl semantische Informationen wie auch technische Informationen, die zum Beispiel das Datenformat betreffen, abgespeichert. Die ISO spezifiziert in der Standard-Reihe 11179 ein umfangreiches Framework für Metadatenverzeichnisse [ISO/IEC 11179-1:2004(E), 2004]. Die Standards umfassen dabei unter anderem Aspekte der Datendefinition, des Metamodells und der Registrierung.

Eine ISO 11179 Edition 2 konforme Software zum Erstellen eines Metadatenverzeichnisses ist das caDSR aus dem caBIG-Projekt (siehe Abschnitt 2.5.4). caDSR ist eine Open-Source-Software, die allerdings zum

Betrieb zwingend das Vorhandensein bestimmter kommerzieller Softwarekomponenten voraussetzt. Weiterhin kann die vom NCI betriebene Instanz des caDSR genutzt werden, wobei eigene Einträge dann nur über den Kuratierungsprozess des NCI möglich sind.

In Deutschland wird derzeit ein durch das Bundesministerium für Bildung und Forschung (BMBF) gefördertes nationales Metadatenverzeichnis zur Sammlung der in klinischen Studien genutzten Merkmalsarten entwickelt und etabliert. Für dieses Verzeichnis ist die Implementierung der Edition 3 des ISO-Standards 11179 vorgesehen [Stausberg et al. 2009].

2.6.4 *Vokabularserver*

Für die Verwaltung und Bereitstellung eines kontrollierten Vokabulars ist ein entsprechendes IT-gestütztes Werkzeug hilfreich. Dieses hat zum einen die Aufgabe, das Vokabular benutzerfreundlich darzustellen und so den Kuratierungsprozess zu unterstützen. Zum anderen muss das Vokabular in verschiedenen Formaten und über verschiedene Schnittstellen zur technischen Integration mit IT-Systemen bereitgestellt werden.

Bezüglich der Schnittstellen haben sich die Webservices Schnittstellen SOAP und REST etabliert (siehe auch Abschnitt 2.5.1). Als Formate für den Export des Vokabulars sind verschiedene XML-Derivate gebräuchlich.

Ein einfach zu verwendender Vokabularserver ist die Open-Source-Software TemaTres [Gonzales-Aguilar et al. 2012]. TemaTres kann über einen Webbrowser bedient werden. Benennungen können sowohl direkt eingegeben, als auch über Textdateien importiert werden. Für die Veröffentlichung des Vokabulars steht neben dem REST-Protokoll auch der Export im SKOS-core, Dublin Core [Weibel 1997] und weiteren Formaten zur Verfügung.

Im caBIG-Verbund wird der Enterprise Vocabulary Service (EVS) als Vokabularserver verwendet [Goncalves et al. 2011]. Der Dienst basiert auf dem Open-Source-Software-Paket *LexEVS*, das für die Veröffentlichung von Terminologien erstellt wurde. Es unterstützt verschiedene Ontologie-Formate einschließlich OWL [Ethier et al. 2013].

2.6.5 *Protégé*

Protégé ist ein Software-Werkzeug zur Erstellung und Pflege von Ontologien und anderen elektronischen Wissensbasen [Noy, Crubezy et al. 2003]. Die Java-basierte Software wird von der Stanford University entwickelt und gewartet.

Protégé erlaubt es dem Benutzer, die Klassen der Ontologie zu definieren, hierarchisch zu organisieren und mit Relationen zu verknüpfen. Das so entstehende Netzwerk kann mithilfe verschiedener Visualisierungsmodule grafisch dargestellt werden. Zur Validierung von Ontologien und zum

automatischen Schließen können Reasoner-Programme eingebunden und ausgeführt werden.

2.6.6 *UMLS*

Das Unified Medical Language System (UMLS) ist ein Projekt der National Library of Medicine (NLM), welches im Jahr 1968 gegründet wurde um die computergestützte Verarbeitung medizinischen Wissens voranzubringen [Lindberg et al. 1993]. Zu diesem Zweck werden kontrollierte Vokabulare aus verschiedenen Quellen zusammengeführt und aufeinander abgebildet. Im Juli 2013 umfasst das UMLS 168 Vokabulare. Somit ist es prinzipiell möglich, die unter Verwendung eines UMLS-Vokabulars erzeugten Metadaten unter Verwendung des UMLS-Abbildungsmodells in ein anderes Vokabular zu übersetzen.

Der Kern des UMLS, der Metathesaurus, besteht aus ungefähr 100 kontrollierten Vokabularen [Tuttle et al. 1988]. Im Rahmen dieser Arbeit sind die beiden folgenden im Metathesaurus enthaltenen Vokabulare von besonderer Bedeutung:

- Medical Subject Headings (MeSH): Ein kontrolliertes Vokabular der NLM zum Indizieren, Katalogisieren und Suchen biomedizinischer Dokumente und Informationen [Lowe, Barnett 1994].
- NCI Thesaurus (NCIt): Die Referenzterminologie des NCI. Es deckt Terme für die klinische Praxis, translationale Forschung, Grundlagenforschung und andere Gebiete ab [Coronado et al. 2004; Golbeck et al. 2003].

Die NLM stellt weiterhin die UMLS-Terminology Services (UTS) als eine Schnittstelle zum Zugriff auf UMLS-Daten bereit [National Institutes of Health o. J.]. Die UTS stellen ein aus mehreren Paketen bestehendes Application Programming Interface (API) bereit, welches dazu genutzt werden kann, um programmatisch auf Daten des UMLS-Metathesaurus und die enthaltenen Datenbanken wie MeSH zuzugreifen. Technisch betrachtet wird das API als Webservice zum Zugriff über das Internet bereitgestellt.

2.7 ANFORDERUNGSANALYSE

Für die Bereiche Systems- und Softwareengineering wird der Begriff *Anforderung* durch den Standard [ISO/IEC/IEEE 24765, 2010] definiert als:

1. eine Vorbedingung oder Fertigkeit die von einem Anwender benötigt wird, um ein Problem zu lösen oder ein Ziel zu erreichen.
2. eine Vorbedingung die erfüllt sein oder eine Fertigkeit, über die ein System, eine Systemkomponente, ein Produkt oder eine Dienst-

leistung verfügen muss um einer Vereinbarung, einem Standard, einer Spezifikation oder anderen formal verpflichtenden Dokumenten gerecht zu werden. Anforderungen schließen die quantifizierten und dokumentierten Bedürfnisse, Wünsche und Erwartungen von Auftraggeber, Kunde und anderen Interessensvertretern mit ein.

3. eine dokumentierte Darstellung von Bedingungen oder Fähigkeiten, wie sie in 1. oder 2. definiert sind.

Auch die *Spezifikationen der Anforderungen* müssen bestimmte Eigenschaften aufweisen, damit sie als Grundlage für den Entwicklungsprozess eines Anwendungssystems geeignet sind. Im Institute of Electrical and Electronics Engineers (IEEE)-Standard 830 wird eine Reihe solcher Eigenschaften vorgeschlagen [IEEE 830, 1998]. Folgende Eigenschaften aus dem Standard werden im Zusammenhang mit der vorliegenden Arbeit als relevant erachtet:

Eine Anforderungsspezifikation bedarf der

- Korrektheit: Eine Anforderungsspezifikation ist nur dann korrekt, wenn die Anforderung tatsächlich von dem betrachteten System erfüllt werden muss.
- Eindeutigkeit: Eine Anforderungsspezifikation ist eindeutig, wenn sie nur eine Interpretation bezüglich der Anforderung zulässt.
- Vollständigkeit: Um vollständig zu sein, muss die Anforderungsspezifikation alle maßgeblichen Anforderungen funktionaler und nicht-funktionaler Art berücksichtigen.
- Konsistenz: Eine Anforderungsspezifikation ist konsistent, wenn sich keine der in ihr enthaltenen Anforderungen widersprechen.
- Einordnung bezüglich Bedeutsamkeit: Die einzelnen Anforderungen sollten nach der Wichtigkeit ihrer Erfüllung geordnet werden.
- Nachprüfbarkeit: Eine Anforderungsspezifikation ist nachprüfbar, wenn es für alle enthaltenen Anforderungen einen realistischen Weg gibt, zu prüfen, ob ein System die Anforderungen erfüllt.

Bisweilen wird in der Literatur zwischen *Zielen* und *Anforderungen* unterschieden [Ammenwerth 2000; Pohl 2007]. Dabei wird das Ziel im Wesentlichen als allgemeinere, weniger konkrete Anforderung beschrieben. So definiert [ISO/IEC/IEEE 24765, 2010] beispielsweise ein Ziel einfach als ein *gewolltes Resultat*. Im Rahmen dieser Arbeit werden daher unter Zielen die Anforderungen höchster Ebene, also die allgemeinsten Anforderungen an das System verstanden.

2.8 REFERENZMODELL

Ein *Modell* ist gemäß [ISO/IEC/IEEE 24765, 2010] die Darstellung eines realen Prozesses, Gerätes oder Konzeptes. Für den Begriff *Referenzmodell* existiert hingegen in der Literatur eine Reihe unterschiedlicher Definitionen [Becker et al. 1995; Fettke, Loos 2004; Thomas 2006]. Fettke und Loos stellen verschiedene Deutungsmöglichkeiten für den Referenzmodellbegriff in der Wirtschaftsinformatik zusammen [Fettke, Loos 2004]. Von den dort zusammen getragenen Definitionen lässt sich die folgende aus dem Bereich der Unternehmensmodellierung auf die Modellierung von Forschungsverbünden übertragen:

> „Referenzmodell als *Menge genereller Aussagen:* In diesem Fall beschreibt das Referenzmodell nicht ein *bestimmtes* Unternehmen, sondern eine *Klasse* von Unternehmen." [Fettke, Loos 2004]

Diese Definition steht im Einklang mit der formaleren Darstellung von Winter:

> „Sei eine Klasse S von Sachverhalten gegeben. Ein Modell R ist *Referenz für eine Klasse S* oder R ist ein *Referenzmodell für die Klasse S*, genau dann wenn R ein allgemeines Modell ist, das
>
> - als Grundlage für die Konstruktion spezieller Modelle für Sachverhalte der Klasse S
>
> oder
>
> - als Vergleichsobjekt für Modelle von Sachverhalten der Klasse S
>
> dienen kann." [Winter et al. 1999]

Ein Referenzmodell für die Anforderungen von Forschungsverbünden ist dementsprechend ein Anforderungsmodell, das als Ausgangsbasis für die Formulierung von Anforderungen eines spezifischen Forschungsverbundes dienen kann. Es enthält für Forschungsverbünde relevante Elemente, die jedoch in der Regel nicht deckungsgleich mit dem zu betrachtenden Verbund sind: Eine Teilmenge der Elemente könnte nicht anwendbar sein, eine weitere Teilmenge könnte Elemente enthalten, die modifiziert anwendbar sind und es könnten weitere Elemente fehlen.

3 METHODIK

Um die Ziele zu erreichen, werden zunächst die Anforderungen von Forschungsverbünden an eine IT-Plattform in Form eines *Referenzmodells für Anforderungen* erarbeitet. Nach dem Referenz-Anforderungsmodell wird ein für den SFB/TRR77 spezifisches, konkretes, Anforderungsmodell beschrieben, das die Grundlage für die Entwicklung der *IT-Architektur* für den Forschungsverbund bildet. Dieser Entwicklungsprozess wird in Abschnitt 3.2 dargestellt. Weiterhin wird das Vorgehen zur Spezifikation von *Metadaten* beschrieben, die zur projektspezifischen Annotation von Forschungsdaten erforderlich sind. Abbildung 2 enthält eine Übersicht der verwendeten Methoden und setzt sie bezüglich ihres Abstraktionsgrades und ihrer Allgemeingültigkeit in Beziehung.

Zur Erfassung der Anforderungen und der Definition der Metadaten werden übergreifende Erhebungen durchgeführt, die Erkenntnisse zu beiden Aspekten liefern: Allen Projektleitern des SFB/TRR77 wird ein Fragebogen vorgelegt, der darauf abzielt, die Grundlagen zu beiden Themenkomplexen zu erheben. In den folgenden Abschnitten werden daher die das jeweilige Thema betreffenden Fragen aus der Erhebung dargestellt.

Nach Auswertung der Fragebogen werden die Projektleiter einer Auswahl aller am Verbund beteiligten Projekte besucht, um die Ergebnisse aus den Fragebogen zu verfeinern. Das Vorgehen bei diesen Interviews wird ebenfalls in den entsprechenden Abschnitten dargelegt. In Anhang B ist eine Liste der Fragen aufgeführt, anhand derer die Interviews durchgeführt werden.

3.1 REFERENZMODELL FÜR ANFORDERUNGEN

Die Erhebung der Anforderungen erfolgt in Anlehnung an die von Abran im Kapitel *Software Requirements* vorgestellte Methodik [Abran 2004]. Das dort beschriebene Vorgehen wird für die Erfassung der Referenzanforderungen entsprechend angepasst. Im Einzelnen werden folgende Schritte durchgeführt:

1. Quellenanalyse
2. Erhebung der Ziele und Anforderungen
3. Dokumentation der Ziele und Anforderungen

Das von Abran vorgeschlagene Vorgehen wurde ursprünglich für die Erfassung projektspezifischer Anforderungen entwickelt. Für die Entwick-

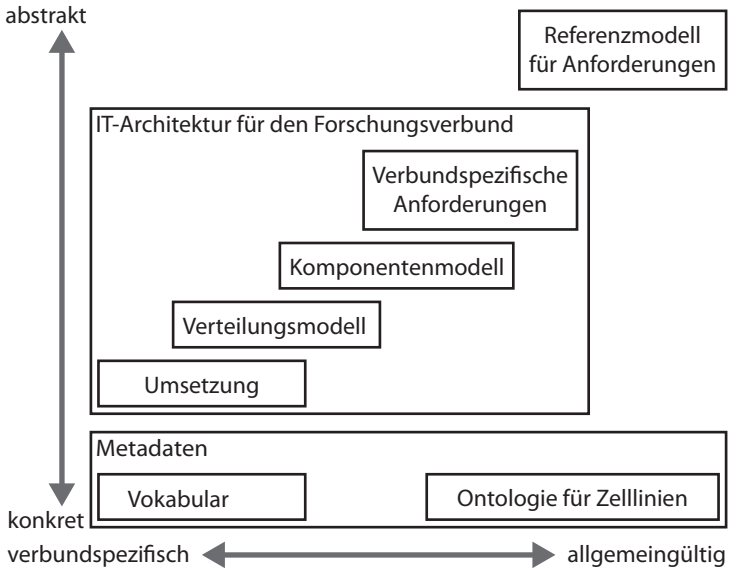


Abbildung 2: Die Abbildung zeigt einen Überblick zur angewandten Methodik. Die Elemente werden zum einen nach ihrem Abstraktionsgrad, zum anderen nach ihrer Allgemeingültigkeit eingeordnet.

lung eines *Referenzmodells* für Anforderungen wird das Vorgehen daher um eine zweite Stufe ergänzt: Nach der quellenspezifischen Erhebung der Ziele und Anforderungen werden diese konsolidiert und abstrahiert. In den folgenden beiden Abschnitten 3.1.1 und 3.1.2 wird das Vorgehen zur Quellenanalyse sowie zur Erhebung der Ziele und Anforderungen dargestellt. In Abschnitt 3.1.3 wird beschrieben, wie diese Erkenntnisse in abstrahierter Form als Referenzmodell für Anforderungen dokumentiert werden.

3.1.1 *Quellenanalyse*

Anforderungen können aus verschiedenen Quellen gewonnen werden, die jeweils eine bestimmte Ebene und einen eigenen Blickwinkel auf die Anforderungen des Systems repräsentieren [Abran 2004]. Für einen Forschungsverbund sind in der Regel die folgenden Quellen verfügbar:

- Projektdokumentation
- Mitarbeiter der Projekte
- Projektkontext

Jede Quelle erfordert, wie im folgenden Abschnitt dargestellt, einen eigenen methodischen Ansatz, um zur Modellierung der Ziele und Anforderungen beizutragen.

Tabelle 2: Auszug von Fragen aus dem Vorerhebungsbogen, welche die Erfassung von Anforderungen berühren.

ERHEBUNGSFRAGE
Welche Datenarten verwenden Sie?
In welchem Format werden die Daten gespeichert?
Wann werden die Daten erzeugt?
Sind Ihre Daten vertraulich?
Welche anderen Projekte könnten an Ihren Daten interessiert sein?
Welche Daten anderer Projekte würden Ihnen bei der Arbeit helfen?
Welche übergreifenden Forschungsfragen interessieren Sie?

3.1.2 *Erhebung der Ziele und Anforderungen*

Jeder Forschungsverbund wird mit der Aufgabe, eine bestimmte Forschungsfragestellung zu bearbeiten, ins Leben gerufen. Für den SFB/TRR77 werden die Projektbeschreibungen aus den Antragsunterlagen zur Erfassung der Ziele herangezogen [Universität Heidelberg, Medizinische Hochschule Hannover 2009]. Weitere Anforderungen können sich aus den dort dokumentierten operationalen und organisatorischen Rahmenbedingungen ergeben.

Hauptquelle für die Definition der Anforderungen sind die Projektleiter der am Forschungsverbund beteiligten Projekte. Sie sollten in der Lage sein, die Beziehungen zwischen den Projekten und die daraus resultierenden Datenintegrationserfordernisse zu formulieren. Insofern repräsentieren sie die Gruppe der späteren Benutzer der Datenintegrationsplattform. Ihre Anforderungen werden in einem zweistufigen Vorgehen erhoben. In der ersten Stufe werden Anforderungen im Rahmen des Fragebogens zur Vorerhebung erfasst. Die den Aspekt der Anforderungsanalyse betreffenden Fragen sind in Tabelle 2 dargestellt.

Im Anschluss werden in einer zweiten Stufe Interviews mit Projektleitern der Verbundprojekte durchgeführt. Die Interviews orientieren sich an einem für den SFB/TRR77 entwickelten Leitfaden (siehe Anhang B). Für die Anforderungsmodellierung besonders relevant sind Fragestellungen, die zur Erfassung funktionaler Aspekte führen. Tabelle 3 zeigt eine Auswahl solcher Fragen, die zur Erhebung der Anforderungen beitragen. Dabei steht die Betrachtung von *Vorgängen* in Zusammenhang mit den Projekten im Vordergrund.

Des Weiteren ist es für den Entwicklungsprozess erforderlich, das zum Verständnis des Forschungsverbundes notwendige Domänenwissen aus dem Projektkontext zu erschließen. Dieses Wissen ermöglicht es erst, die

Tabelle 3: Auszug von Fragen aus dem Interviewleitfaden für die Erfassung von Anforderungen (siehe auch Anhang B)

INTERVIEWFRAGE
Welche Vorgänge/Prozesse finden im Projekt statt?
Welche Auswertungen finden statt?
Wie ist der Lebenszyklus der Daten?

formulierten Anforderungen zu verstehen. Weiterhin ergeben sich aus dem Hintergrundwissen implizit weitere Anforderungen. Um das Domänenwissen zu überprüfen und die sich daraus ergebenden Anforderungen zu konkretisieren, wird ein erster Prototyp eines IT-Systems auf der Basis einer SOA erstellt. Dieser Prototyp wird einer Auswahl an Projekten des Forschungsverbundes vorgestellt. Mit Hilfe des Feedbacks zum Prototypen werden die bisher erfassten Anforderungen verfeinert und gegebenenfalls ergänzt.

3.1.3 *Dokumentation der Ziele und Anforderungen*

Die erfassten Anforderungen werden systematisch dokumentiert. Dies geschieht einerseits mit Hilfe von Anforderungsdiagrammen in Unified Modeling Language (UML)-Notation. Diese Diagramme sind geeignet, einen Überblick über die einzelnen Anforderungen sowie ihre Beziehungen untereinander zu geben. Dabei kommen die Beziehungsarten *Aggregation*, *Generalisierung* und *Assoziation* zur Anwendung. Abbildung 3 zeigt die Elemente und Beziehungen, die im Anforderungsdiagramm zur Anwendung kommen. Die Anforderungsdiagramme werden, ebenso wie alle anderen UML-Diagramme in dieser Arbeit, mit der Modellierungssoftware *Enterprise Architect Professional Version 9.3* (*http://www.sparxsystems.com*, Stand: 4.12.2013 15:31) erstellt.

Weiterhin werden die Eigenschaften jeder Referenzanforderung nach dem in Tabelle 4 dargestellten Muster dokumentiert. In den Tabellen werden die Nummern und Bezeichnungen aus dem Anforderungsdiagramm aufgegriffen. Zusätzlich enthält die Tabelle eine Beschreibung der Anforderung sowie deren Gewichtung. Damit werden die in Abschnitt 2.7 empfohlenen Eigenschaften für den Einsatz in Forschungsverbünden angemessen repräsentiert.

3.2 IT-ARCHITEKTUR FÜR DEN FORSCHUNGSVERBUND

Das im ersten Teil beschriebene Referenzmodell bildet die Ausgangsbasis für die Entwicklung der IT-Architektur für den Forschungsverbund. Im

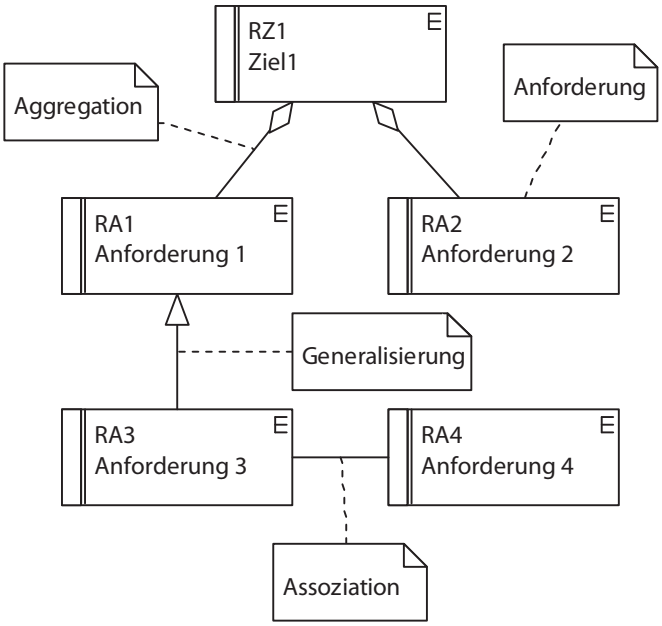


Abbildung 3: Überblick über die Elemente eines Anforderungsdiagramms in UML-Notation. Die Referenzanforderungen und -ziele werden im Diagramm durch eine eindeutige Nummer (wie beispielsweise RZ1 oder RA1) sowie die Bezeichnung des Elements beschrieben. Die Verbindungen zwischen den Elementen symbolisieren die Art der Beziehung.

nächsten Schritt werden zu den allgemeinen Anforderungen und Zielen des Referenzmodells die *verbundspezifischen* Instanzen formuliert. Diese Anforderungen werden sodann auf IT-Komponenten abgebildet, welche diese Anforderungen durch geeignete Funktionen erfüllen. Somit wird zunächst ein produktunabhängiges Komponentenmodell erstellt. Für jede abstrakte Komponente gibt es verschiedene Möglichkeiten, diese konkret zu implementieren, daher wird für jede Komponente eine Umsetzungsvariante festgelegt. Schließlich wird ein Verteilungsmodell entwickelt, welches beschreibt, wie die Komponenten auf Systemressourcen abgebildet werden. Begleitet wird die Entwicklung der IT-Architektur von der Berücksichtigung sicherheitsrelevanter Aspekte.

Die so entwickelte IT-Architektur wird für den SFB/TRR77 als lauffähiges System umgesetzt. Das Gesamtsystem trägt den Namen *Platform Enhancing Liver Cancer Networked Research (PELICAN)*. Bei PELICAN handelt es sich um ein Java-basiertes SOA-System, welches von den Benutzern mittels eines Webbrowsers bedient wird.

Tabelle 4: Muster zur Beschreibung von Referenzanforderungen

EIGENSCHAFT	ERKLÄRUNG
Nummer	Eindeutige Nummer zur Identifikation der Anforderung
Bezeichnung	Benennung der Anforderung
Beschreibung	Erläuterung der Anforderung
Gewichtung	Bedeutung der Anforderung zur Erfüllung der Ziele (niedrig, mittel, hoch)

3.2.1 *Verbundspezifische Anforderungen*

Während die nach Abschnitt 3.1 erstellten Referenzziele und Referenzanforderungen einen allgemeingültigen Charakter haben, weichen die Ziele und Anforderungen eines *konkreten* Forschungsverbundes möglicherweise von diesen ab oder konkretisieren sie. Daher werden die Elemente des Referenzmodells für Anforderungen für den SFB/TRR77 auf ihre Wichtigkeit hin überprüft. Die Dokumentation des verbundspezifischen Anforderungsmodells erfolgt analog zum Referenzmodell für Anforderungen nach dem in Tabelle 4 dargestellten Muster. Zur eindeutigen Unterscheidung zwischen Referenzanforderungen und konkreten Anforderungen erhalten letztere Identifikationsnummern mit dem Präfix Z beziehungsweise A statt RZ beziehungsweise RA.

Zu den im Abschnitt 3.1 erfassten Referenzzielen und Referenzanforderungen werden die verbundspezifischen Instanzen dargestellt. Einen wesentlichen Anteil am verbundspezifischen Anforderungsmodell besitzen die Anforderungen, welche die Projekte des Verbundes an die Verarbeitung ihrer Daten stellen. Zum genauen Verständnis dieser Anforderungen ist zunächst die Erfassung der im Verbund verwendeten Datenarten erforderlich. Anschließend werden die zugehörigen Anforderungen betrachtet.

Erhebung der Datenarten

Die Datenarten werden mit Hilfe des Fragebogens (siehe Anhang A) bei den Projektleitern des SFB/TRR77 erhoben. Daraus sind die folgenden beiden Fragen bezüglich der Datenarten relevant:

1. *What kind of data do you use? (Frage 2)* Vorgegebene Antwortmöglichkeiten sind *Microarray-Daten, Bilder, Texte* und *Sonstiges*. Zu jedem Punkt besteht die Möglichkeit der Präzisierung als Freitexteintrag.

2. *In which format are data stored? (Frage 3)* Hier stehen die Formate *Excel-Datei, XML-Datei, Textdatei, Datenbank, Character Separated*

Value (CSV)-Datei, Bilddatei und *Sonstiges* als mögliche Antworten zur Auswahl. Zu den Antworten *Datenbank* und *Sonstiges* besteht die Möglichkeit des Freitextkommentars.

Im Zuge der Auswertung des Fragebogens wird somit ein Überblick sowohl über die im SFB/TRR77 auftretenden Datenarten als auch deren Speicherformate gewonnen. Für alle Projekte werden die jeweils relevanten Datenarten sowie die korrespondierende Datenstruktur möglichst vollständig erfasst. Die Daten können dabei in IT-verwertbarer Form wie zum Beispiel einer Datei eines Tabellenkalkulationsprogramms vorliegen oder auch analog in Form eines Westernblots vorliegen. Weiterhin können die Daten einen unterschiedlichen Verarbeitungsgrad aufweisen: Bei Microarrays können die Lesewerte des Scanners als Rohdaten vorliegen, aber auch deren qualitätsgesicherte Interpretationen. Ähnlich dazu kann von einem Blot eine (digitale) Fotografie vorliegen oder die Interpretation des Blots.

Die Datenarten werden formalisiert erfasst, sodass analoge Datenstrukturen über die Projektgrenzen hinweg gefunden und standardisiert werden können. Durch die Beschreibung solch harmonisierter Datenelemente wird der Transformationsaufwand bei einer späteren Verknüpfung der Daten minimiert. Die Relationen zwischen einzelnen Datenelementen werden dabei in Datenmodellen beschrieben. Dies ist eine Voraussetzung, um Daten verschiedener Art auf eine sinnvolle Weise miteinander zu verknüpfen.

Datenschutz und Vertraulichkeit

Zum Themengebiet Datenschutz und Vertraulichkeit werden den Projektleitern ebenfalls zwei Fragen gestellt. Die Frage *Are you concerned about the confidentiality of your data? (Frage 5)* zielt auf den generellen Schutzbedarf, den die Projektleiter für ihre Daten sehen. Die vorgegebenen Antworten reichen von sehr hohem (Kein anderes Projekt darf auf die Daten zugreifen) bis zu geringem Schutzbedarf (Jeder, auch von außerhalb des SFB/TRR77, kann auf die Daten zugreifen).

Die zweite Frage *Would you provide your data in a common TRR77 information system? (Frage 6)* erfasst die konkrete Bereitschaft der Projektleiter, Daten zur Verfügung zu stellen. Hier werden die Antwortmöglichkeiten

- *ja*,
- *Nur wenn meine Vertraulichkeitsanforderungen erfüllt sind* und
- *nein*

vorgegeben.

Analyse

Die mit den Daten durchzuführenden Analysen haben möglicherweise Einfluss auf das Design von PELICAN. Um einen Überblick zu bekommen,

Tabelle 5: Muster zur Dokumentation der Abbildung der Anforderungen auf Systemeigenschaften und Komponenten

ANFORDERUNG	EIGENSCHAFT	KOMPONENTE
A*n* Anforderung *n*	Beschreibung der Systemeigenschaft	Name der verantwortlichen Komponente

welche Analysen die Projektleiter mit ihren Daten planen, wird ihnen die Frage 7 des Fragebogens gestellt: *How do you analyze your data?* Diese Frage hat keine vorgegebenen Antwortmöglichkeiten sondern ist als Freitext gestaltet.

3.2.2 *Technologieauswahl*

Für die Umsetzung der IT-Architektur werden im Rahmen dieser Arbeit zwei etablierte Bioinformatik-Frameworks betrachtet, die unterschiedliche Ansätze zur Datenhaltung verfolgen: caBIG verfolgt einen eher dezentralen, föderierten Ansatz, während i2b2 ein zentrales Datawarehouse bereit stellt. Die beiden Frameworks werden dahingehend untersucht, ob sie zur Abdeckung der geforderten Systemeigenschaften geeignet sind. Die Kriterien für die Technologieauswahl ergeben sich aus den Anforderungen.

3.2.3 *Abbildung der Anforderungen auf Systemeigenschaften*

Jede Anforderung an das PELICAN-System muss durch eine Systemeigenschaft umgesetzt werden. Da das zu erstellende System aus mehreren Komponenten besteht, werden diese Eigenschaften ebenfalls auf die Komponenten abgebildet. Die Komponenten werden zunächst mit allgemeinen, nicht produktspezifischen Begriffen beschrieben.

Die Dokumentation der Abbildung erfolgt nach dem Muster von Tabelle 5. Dabei wird für alle Anforderungen, die einem Ziel zugeordnet sind, jeweils eine gemeinsame Tabelle erzeugt.

3.2.4 *Komponentenmodell*

Zur Umsetzung des SOA-Systems ist es erforderlich, die generischen Komponenten zu konkretisieren. Dies kann zum einen dadurch geschehen, dass der Komponente ein existierendes Produkt zugeordnet wird, zum anderen dadurch, dass die Komponente als neu zu entwickeln gekennzeichnet wird. Als Mischform kann einer Komponente ein Framework zugeordnet wer-

den, welches zur vereinfachten Entwicklung der jeweiligen Komponente verwendet wird.

Datendienste

Datendienste werden in PELICAN mit der caCORE SDK-Komponente umgesetzt. Zur Erstellung eines Dienstes werden dessen Eigenschaften mit Hilfe eines UML-Modells beschrieben. Das Modell wird anschließend im XML Metadata Interchange (XMI)-Format gespeichert und dient so als Eingabe für den Dienstgenerator des caCORE SDK. Die Erstellung des Modells erfolgt manuell für jeden Dienst unter Verwendung der Software *Enterprise Architect Professional Version 9.3*.

Das zu erstellende Modell umfasst im Wesentlichen zwei Teile: Zum einen ein logisches Modell, welches die für den Dienst erforderlichen Klassen und deren Beziehungen modelliert. Da das caCORE SDK und auch die übrigen Komponenten von PELICAN in Java implementiert sind, handelt es sich dabei um ein Java-Klassendiagramm in UML-Notation. Zum anderen wird ein physisches Datenmodell erzeugt, welches die Struktur der vom Datendienst bereitgestellten Daten in der Datenbank beschreibt. Auch hierbei handelt es sich um ein UML-Klassendiagramm, das allerdings das Datenmodell der zu Grunde liegenden MySQL-Datenbank (siehe *http://www.mysql.com/*, Stand: 01.08.2012, 14:01) abbildet. In diesem Modell sind sämtliche Datenbankaspekte, wie beispielsweise Fremdschlüssel, Bedingungen oder Trigger, enthalten.

Weiterhin enthält das erstellte Modell Beziehungen, welche die logischen und die physikalischen Aspekte miteinander verknüpfen. Zur detaillierten Beschreibung des Modells für den caCORE SDK-Generator müssen die Attribute der modellierten Klassen sowie bestimmte Assoziationen mit sogenannten *Custom Tag Values* annotiert werden. Diese Tag Values erfüllen komplexe Steuerungsaufgaben im Dienstgenerator und müssen präzise eingesetzt werden. Die verfügbaren Tag Values werden im Kapitel 3 des *caCORE SDK Version 4.3 Object Relational Mapping Guide* [Wiley, Gagne 2012] beschrieben.

Aus dem XMI-Modell erzeugt und testet der caCORE SDK-Generator den Dienst in Form einer Java-Web-Applikation. Diese Applikation kann sodann auf einem entsprechenden Applikationsserver gestartet werden. Im Rahmen von PELICAN wird der Applikationsserver *Tomcat Version 5.5.20* verwendet (Siehe *http://tomcat.apache.org*, Stand: 01.08.2012, 14:23). Das modellierte Datenbankschema wird mit Hilfe der Enterprise Architect-Software aus dem Modell exportiert und auf dem MySQL-Datenbankserver installiert. Anschließend werden die Daten selbst unter Verwendung der ETL-Software *Talend Open Studio for Data Integration Version 5.0.0* (siehe *http://de.talend.com*, Stand: 09.12.2013, 22:28) in das definierte Datenformat transformiert und in die Datenbank geladen. Mit Talend Open Studio wird

dafür eine entsprechende Transformationsvorschrift unter Verwendung der grafischen Benutzeroberfläche erstellt. Das verwendete Eingangsformat ist dabei eine Datei im Excel-Format, deren Inhalt nach der Transformation direkt in die Datenbank geladen wird.

Auf den Dienst kann anschließend über Webservices-Schnittstellen zugegriffen werden. Zur einfachen Prüfung des Modellierungs- und Generierungsprozesses steht weiterhin eine Webschnittstelle zur Verfügung, über welche die Daten des Dienstes mit Hilfe eines Webbrowsers abgerufen werden können.

Portlets

Portlets werden in Java mit Hilfe des Liferay Plugins SDK entwickelt, das in die Java-Entwicklungsumgebung *Eclipse Juno* (siehe *http://www.eclipse.org/juno/*, Stand: 01.12.2013, 22:09) integriert wird [Sezov 2012]. Im Zuge der Entwicklung werden Java-Klassen entwickelt, welche die Portletspezifikation nach [JSR 286, 2008] implementieren. In der Regel folgen Portlets dem Model/View/Controller-Entwurfsmuster [Gamma et al. 1995]. Dabei werden die aus den Daten- und Analysediensten gewonnenen Daten in sogenannten *Model*-Objekten gespeichert. Zur Darstellung im Portlet wird für jede Ansicht eine Extensible HyperText Markup Language (XHTML)-Komponente, die *View*-Komponente, programmiert. Während der Erzeugung der vollständigen Portalseite durch den Portalserver ist die View-Komponente für die Bereitstellung der entsprechenden Hypertext Markup Language (HTML)-Fragmente des Portlets verantwortlich.

Zur Vereinfachung der Kommunikation zwischen den Modell-Klassen und der View-Komponente sowie zur Unterstützung der Oberflächenprogrammierung wird in PELICAN die Java-Bibliothek JavaServer Faces (JSF) eingesetzt. Da innerhalb der Portlets auch mit den caBIG-Diensten kommuniziert werden muss, sind die Bibliotheken des caCORE SDK ebenfalls einzubinden. In der Folge entsteht eine komplexe Abhängigkeit, da das Liferay Plugins SDK bereits selbst einige der von caCORE SDK benötigten Bibliotheken in anderer Version mitbringt und diese Konflikte im Rahmen der Entwicklung aufgelöst werden müssen.

3.2.5 *Verteilungsmodell*

Die Software, welche die bei der Komponentenmodellierung identifizierten Elemente umsetzt, muss für den Betrieb geeignet verteilt werden. Hierzu werden die Komponenten auf entsprechende Systemressourcen abgebildet. An Hardwareressourcen stehen zwei Serversysteme mit den in Tabelle 6 dargestellten Leistungsparametern zur Verfügung. Kriterien für die Verteilung der Komponenten sind:

Tabelle 6: Hardwareressourcen für den Betrieb von PELICAN

	SERVER 1	SERVER 2
Modell	IBM Bladeserver HS22 7870-H2G	IBM Bladeserver HS23 7875-C8G
Prozessoren	2 x Intel Xeon 6-Kern-Prozessor X5650, 2,66 GHz	2 x Intel Xeon 8-Kern-Prozessor E5-2680, 2,7 GHz
Hauptspeicher	48 GB	224 GB
Festplattenspeicher	500 GB	900 GB
Gemeinsamer SAN-Speicher	zusammen 500 GB	

- Sicherheit: Bestimmte, sensible Komponenten sollten nicht zusammen mit anderen Komponenten auf einem Knoten betrieben werden.
- Leistung: Die Komponenten müssen so verteilt werden, dass ihnen ausreichende Ressourcen zur Verfügung stehen.
- Verwaltung: Die Komplexität der Verwaltung des Gesamtsystems soll durch eine geeignete Verteilung der Komponenten unterstützt werden.

3.3 VERBUNDSPEZIFISCHE METADATEN

Um eine projektübergreifende Zusammenführung von Daten zu ermöglichen, ist es zunächst erforderlich, die in den Projekten erzeugten und verarbeiteten Datenarten in ihrer Struktur zu erfassen und semantisch zu annotieren. Durch diese Vorarbeiten wird die Möglichkeit eröffnet, die unterschiedlichen Datenquellen automatisiert miteinander in Beziehung zu setzen und somit dynamisch übergreifende Datenanalysen zu ermöglichen. Hierfür wird ein kontrolliertes Vokabular sowie eine Ontologie zur Beschreibung von Zelllinien entwickelt.

3.3.1 *Kontrolliertes Vokabular*

Da zu Beginn des Forschungsverbundes nicht bekannt ist, welche Begriffe in den einzelnen Projekten genutzt werden und ob die Verwendung konsistent über die Projekte hinweg ist, wird im Folgenden ein Vorgehensmodell zur Erstellung eines verbundspezifischen kontrollierten Vokabulars vorgestellt. Um sicherzustellen, dass sich die Begriffe des verbundspezifischen Vokabulars im Einklang mit öffentlich zugänglichen kontrollierten Voka-

bularen befinden, wird weiterhin die Entwicklung eines Werkzeugs zum Abgleich von Vokabularen vorgestellt.

Zur korrekten semantischen Beschreibung der erfassten Datenarten und -elemente ist zunächst ein angemessenes Verständnis der Bedeutung der projektspezifischen Daten sowie der damit assoziierten Vorgänge erforderlich. Hierzu wird die in den einzelnen Projekten verwendete Terminologie erfasst, harmonisiert und anschließend auf einem Vokabularserver bereitgestellt. Bei diesem Vorgang können Relationen wie Synonymie, Hyponymie oder Hyperonymie zwischen den Begriffen erfasst und dargestellt werden.

Im caBIG-Dienst EVS sind bereits die umfangreichen biomedizinischen Terminologien NCIt, NCI Metathesaurus (NCIm) und MeSH verfügbar. Im Zuge der Erhebung müssen daher Daten erfasst werden, die es erlauben, zu bewerten, ob die Terminologien des EVS ausreichen, um die Begriffe der betrachteten Projekte vollständig zu erfassen.

Zur Erfassung des verbundweiten Vokabulars wird ein zweistufiger Prozess, bestehend aus einer Vorerhebung und einer Haupterhebung, definiert. In der Vorerhebung werden zunächst grundlegende Informationen über die in den einzelnen Projekten verarbeiteten Daten erfasst. Diese Informationen dienen zu Vorbereitung auf die anschließende Haupterhebung. Ziel der zweiten Stufe ist es schließlich, die projektspezifischen Begriffe zu erfassen und zusammen mit Kontextinformationen zu dokumentieren.

Vorerhebung

Die Vorerhebungsphase dient dazu, die Wissensdomäne eines Projektes zu erschließen und somit ein tieferes Verständnis für die Forschungsfragestellung im Rahmen der Haupterfassung zu gewährleisten. Dies geschieht zunächst durch die systematische Analyse der im Antragsverfahren veröffentlichten Projektbeschreibungen. Hier kommen Ansätze aus der Qualitativen Inhaltsanalyse, auch Qualitative Datenanalyse (QDA) genannt, zum Einsatz [Mayring 2010]. Insbesondere der Einsatz spezieller QDA-Software ist hilfreich: Hierbei werden elektronische Fassungen der Projektunterlagen mit der Software geöffnet und nach potentiell relevanten Termen durchsucht. Dies kann manuell, aber auch automatisiert geschehen. Identifizierte Begriffe werden direkt im Dokument markiert. Im Rahmen dieser Arbeit kommt die Software MAXQDA (siehe *http://www.maxqda.de*, Stand: 05.12.2013, 11:17) zum Einsatz.

Durch dieses Vorgehen entsteht ein annotiertes Dokument, in dem die Fundstellen der Begriffe auch wieder zurückverfolgt werden können. Die markierten Begriffe können anschließend in Form einer Datei im Excel-Format exportiert und weiterverarbeitet werden.

Als zweiter Bestandteil der Vorerhebungsphase wird der Fragebogen, welcher an die Projektleiter verteilt wird, herangezogen. Tabelle 2 enthält einen Auszug der erhobenen Fragen. Der vollständige Fragebogen ist im

Anhang A enthalten. Hauptsächlich erfasst wird, welche Datenarten im jeweiligen Projekt zur Anwendung kommen. Für die Analyse des Fragebogens werden diese digitalisiert und ebenfalls mit der QDA-Software MAXQDA bearbeitet. Die in den Fragebogen identifizierten Termkandidaten werden analog zur Projektdokumentation in den Fragebögen markiert und anschließend in Excel-Dateien exportiert.

Zum Abschluss der Vorerhebungsphase werden die Termlisten in einer Excel-Datei zusammengeführt und alphabetisch sortiert. Mehrfache Einträge werden mit Hilfe der entsprechenden Excel-Funktion entfernt.

Haupterhebung

Bei der Haupterhebung liegt der Fokus auf der Sammlung der in den Projekten verwendeten Terme. Dies geschieht im Rahmen von Interviews, die mit einem oder mehreren Mitarbeitern der Projekte durchgeführt werden. Die Durchführung dieser Interviews wird durch einen Interviewleitfaden unterstützt. Bei der Erstellung des Leitfadens werden die in der Vorerhebung gewonnenen Erkenntnisse berücksichtigt. Zur Haupterhebung werden Besuche bei Projekten des SFB/TRR77 durchgeführt. Im halbstrukturierten Interview werden zunächst die jeweiligen Prozesse zur Erzeugung der Forschungsdaten erfragt. Bezüglich des Vokabulars interessieren im Interview vor allem vorhandene Forschungsdaten und Prozesse, die zur Erzeugung von solchen Daten führen. Beispiele für solche Prozesse sind biochemische Analysen wie Fluorescence-activated Cell Sorting (FACS) oder Quantitative Polymerase Chain Reaction (qPCR) sowie Datenbankabfragen.

Während des Interviews notiert der Interviewer die Terme, welche von den Wissenschaftlern bei der Beschreibung ihrer Arbeit gebraucht werden. Im Rahmen der Nachbereitung eines jeden Interviews werden die Ergebnisse auf zwei Arten dokumentiert: In einer Termliste und in einer Mind-Map. Diese beiden Sichtweisen werden in den folgenden Abschnitten beschrieben.

ROHFASSUNG DES VOKABULARS Nach dem Interview werden die erfassten Bezeichnungen in eine Tabelle überführt. Somit entsteht eine Dokumentation analog zu der in der Vorerhebungsphase erzeugten Kandidatenliste für Terme. Werden im Zuge der Analyse der projektspezifischen Begriffe auch Synonyme, Akronyme oder Abkürzungen erkannt, so werden diese neben den entsprechenden Vorzugsbezeichnungen in separaten Spalten notiert. Ähnlich verhält es sich mit der Sprache: Da viele Forschungsprojekte mit internationalen Teams besetzt sind, werden oft englische Bezeichnungen den entsprechenden deutschen Bezeichnungen vorgezogen. Daher werden als Vorzugsbezeichnungen für das gemeinsame Vokabular englische Bezeichnungen verwendet, und gegebenenfalls zusätzlich die deutsche Entsprechung in einer separaten Spalte mitgeführt.

Tabelle 7: Mustertabelle für die projektweise Dokumentation der Begriffe

VORZUGSBEZEICHNUNG	AKRONYM	DEUTSCH
Die präferierte Bezeichnung für den Begriff (englisch)	Gegebenenfalls ein Akronym oder eine Abkürzung	Deutsche Übersetzung der Vorzugsbezeichnung

Tabelle 7 zeigt eine Mustertabelle nach der die Bezeichnungen eines Projektes dokumentiert werden. Die Listen werden auf Fehler, die beispielsweise durch Missverständnisse oder Schreibfehler entstanden sind, überprüft und gegebenenfalls korrigiert. Nachdem alle projektspezifischen Termlisten erstellt wurden, werden die Tabellen, unter Einbeziehung der in der Vorphase erzeugten Tabellen, zu einer einzigen Tabelle zusammengeführt.

Diese Tabelle stellt eine erste Rohfassung des gewünschten Vokabulars dar. Sie wird anhand der Spalte der *Vorzugsbezeichnungen* alphabetisch sortiert und auf doppelte Einträge hin überprüft und bereinigt. Die Liste der Vorzugsbezeichnungen dient als Grundlage für den Abgleich mit dem UMLS.

MIND-MAPS Ein Ziel bei der Erstellung des kontrollierten Vokabulars ist es, die projektspezifische Terminologie soweit zu erfassen, dass die im Projekt erzeugten Daten und Ergebnisse semantisch angemessen beschrieben werden. Ein Aspekt ist hierbei die Dokumentation der Beziehungen zwischen den von den Wissenschaftlern gebrauchten Begriffen. Zur Beschreibung solcher Beziehungen eignen sich Mind-Maps [T Buzan, B Buzan 2002]: Sie stellen eine informelle Möglichkeit dar, die Konzepte und Begriffe einer Domäne auch ohne tiefergehendes Wissen über Taxonomien oder Ontologien miteinander in Beziehung zu setzen und zu visualisieren [Braun et al. 2008]. Abbildung 4 zeigt die in einem Projekt erfassten Begriffe in einer beispielhaften Mind-Map. Der Wurzelknoten stellt das Projekt selbst dar. Er wird umgeben von den Bezeichnungen, welche die Forschungsprozesse des Projektes beschreiben. Die Struktur der Mind-Map ist die Grundlage für die hierarchische Anordnung der Begriffe in einem späteren Schritt der Vokabularentwicklung.

Automatisierter Abgleich mit UMLS-Vokabularen

Um sicherzustellen, dass das neu zu erstellende kontrollierte Vokabular einfach mit bestehenden, öffentlich zugänglichen Vokabularen integriert werden kann, werden die erfassten Bezeichnungen mit den Vokabularen des UMLS abgeglichen. Zu diesem Zweck wird ein Software-Werkzeug namens

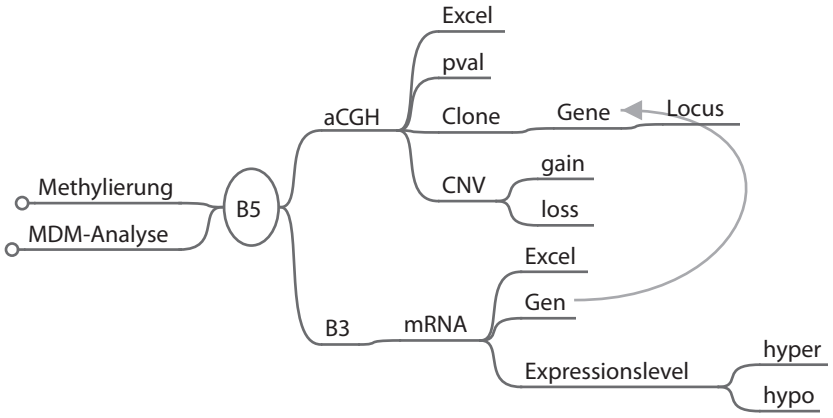


Abbildung 4: Muster für die Darstellung von Begriffen in einer Mind-Map am Beispiel eines Projektes (B5). Im Projekt werden englische und deutsche Bezeichnungen gemischt verwendet. Weiterhin arbeitet das Projekt mit einem anderen Projekt (B3) zusammen. Auch nicht-hierarchische Beziehungen können in Mind-Maps dargestellt werden, wie hier zwischen Gen und Gene gezeigt.

Vocabulary Alignment (VOCALIGN) entwickelt, das den Abgleichprozess in Teilen automatisiert.

Bei VOCALIGN handelt es sich um eine Java-Anwendung, welche die Webservice-Schnittstelle der UTS ansteuert. Das Programm erwartet eine Liste mit Bezeichnungskandidaten als Eingabe. Diese Bezeichnungen werden nacheinander als Suchworte an den UTS geschickt, wobei die Suche auf die beiden biomedizinisch bedeutsamen Vokabulare MeSH und NCIt beschränkt wird.

Auswertung

Im Zuge der Auswertung werden die erhobenen Terminologiefragmente mit den Taxonomien aus caBIG verglichen. Dadurch wird untersucht, ob die caBIG-Taxonomien hinreichend sind, um die durch den Forschungsverbund aufgespannte Domäne zu beschreiben. Umgekehrt wird auch diejenige Untermenge der Taxonomien definiert, die notwendig ist, um die Domäne abzubilden. Durch eine solche Einschränkung entsteht eine besser handhabbare, projektspezifische Taxonomie.

3.3.2 *Ontologie für Zelllinien*

Die Entwicklung der Zelllinien-Ontologie orientiert sich an den Empfehlungen von Noy und McGuinness [Noy, McGuinness 2001]. Dementsprechend ist es für die umfassende Beschreibung von Zelllinien zunächst

Tabelle 8: Die Katalog-Datenbanken von fünf gängigen Zellbanken wurden in den Vergleich der Datenfelder eingeschlossen. (Stand: 08.01.2014, 8:43)

NAME DER ZELLBANK	INTERNETADRESSE
American Type Culture Collection (ATCC)	*http://www.atcc.org*
European Collection of Cell Cultures (ECACC)	*https://www.phe-culturecollections.org.uk*
Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSMZ)	*http://www.dsmz.de*
Interlab Cell Line Collection (ICLC)	*http://www.iclc.it*
Riken	*http://www.brc.riken.jp*

notwendig, die erforderlichen Metadatenelemente zu analysieren. Als geeignete Grundlage hierfür werden die von den wichtigsten Zellbanken veröffentlichten Kataloge für Zelllinien angenommen. Des Weiteren wird eine Literaturrecherche in PubMed und Google Scholar durchgeführt, um die üblicherweise herangezogenen Zellbanken zu identifizieren. Dieses Vorgehen führt zu den in Tabelle 8 dargestellten Zellbanken.

Aus den über das Internet verfügbaren Katalogen werden die entsprechenden Datenstrukturen extrahiert. Die Datenfelder werden verglichen und zu einem Feld konsolidiert, das als vorläufige Bezeichnung dient. In einigen Quelldatenbanken werden mehrere Konzepte gemeinsam in einem Freitextfeld abgelegt. In diesem Fall werden die Konzepte in mehrere vorläufige Benennungen aufgespalten. Die konsolidierten Benennungen werden zu Gruppen verschiedener Bereiche der Beschreibung von Zelllinien zusammengefasst. Diese Gruppen sind hilfreich, um Kandidaten für Ontologien zu identifizieren, da diese Ontologien idealerweise einen definierten Spezialbereich abdecken.

Entwicklung der Ontologie

Unter Berücksichtigung von Publikationen zur Entwicklung von Ontologien [Cimino 1998; Noy, McGuinness 2001] wird ein aus vier Schritten bestehendes Verfahren definiert:

1. Definition des Begriffsbereichs
2. Durchsicht bestehender Ontologien
3. Erstellen eines konzeptuellen Modells
4. Formale Darstellung der Ontologie

DEFINITION DES BEGRIFFSBEREICHS Aus der Analyse der Zellbank-Kataloge folgt die Definition des Begriffsbereichs und des Umfangs der neuen Ontologie. Weiterhin werden die aus diesen Katalogen extrahierten Feldnamen direkt in Elemente des neuen Begriffsbereichs überführt. Die geplante Ontologie soll in möglichst großem Umfang die Begriffe aus dem Bereich Zelllinien umfassen, entweder durch die Wiederverwendung bestehender Ontologien oder durch die Definition neuer Ontologie-Klassen.

ANALYSE BESTEHENDER ONTOLOGIEN Im nächsten Schritt werden bereits existierende Ontologien identifiziert, welche die Gruppen der Beschreibungselemente für Zelllinien abdecken. Dies geschieht durch Recherchen in Internetdatenbanken für biomedizinische Ontologien. Die bekannteste Datenbank ist BioPortal [Noy, Shah et al. 2009] sein, welche unter der Internetadresse *http://bioportal.bioontology.org* (Stand: 25.09.2013, 14:12) erreichbar ist. Im März 2012 enthielt BioPortal 302 Ontologien mit zusammen über fünfeinhalb Millionen Begriffen. Als weitere Quelle für biomedizinische Ontologien wird der Ontology Lookup Service (OLS) herangezogen [Côté et al. 2006]. Dieser Dienst ist unter der Adresse *http://www.ebi.ac.uk/ontology-lookup* (Stand: 25.09.2013, 14:18) im Internet verfügbar. Vervollständigt wird die Suche nach geeigneten Ontologien durch eine Literaturrecherche in PubMed.

Ontologien, die als Kandidaten für die Verwendung in der Zielontologie betrachtet werden, sollten möglichst viele der in den vorangegangenen Schritten identifizierten konsolidierten Felder abdecken. Da die bestehenden Ontologien jedoch im Allgemeinen für einen bestimmten Zweck, jedoch nicht notwendigerweise für die Beschreibung von Zelllinien erstellt wurden, ist nicht zu erwarten, dass exakt passende Ontologien gefunden werden. Vielmehr können nur Teile bestehender Ontologien sinnvoll zur Abdeckung einiger Bereiche herangezogen werden. Im Idealfall kann der komplette Begriffsbereich für Zelllinien durch die Kombination geeigneter Elemente aus bestehenden Ontologien abgedeckt werden.

ERSTELLEN EINES KONZEPTUELLEN MODELLS Die aus bestehenden Ontologien importierten Klassen werden ebenso wie die neu definierten Klassen in eine hierarchische Struktur gebracht. Dieses konzeptuelle Modell ist das Grundgerüst der neuen Ontologie. Es sorgt für die Integrierbarkeit und Wiederverwendbarkeit der Ontologie, indem beispielsweise eine übergeordnete Ontologie integriert wird.

FORMALE DARSTELLUNG DER ONTOLOGIE Im letzten Schritt wird die Ontologie in Form einer OWL-Datei (siehe Abschnitt 2.6.2) beschrieben. Diese Datei bildet auch die Grundlage für die Evaluation der Ontologie. Weiterhin wird die Datei für den Zugriff über das Internet veröffentlicht.

Evaluierung der Ontologie

Die neu erstellte Ontologie bedarf der Evaluierung bezüglich ihrer Korrektheit und Benutzbarkeit. In der Literatur werden verschiedene Ansätze zur Evaluierung beschrieben [Obrst et al. 2007; Rogers 2006; Vrandečić 2009]. Für die Evaluierung der Zelllinien-Ontologie werden die von Obrst et al. vorgestellten Methoden angewandt, soweit dies für die Zelllinien-Ontologie sinnvoll ist. Ein Überblick über diese Techniken ist in Tabelle 9 dargestellt. Als Teil der Evaluation werden neben neun humanen und vier Maus-Zelllinien auch je ein primärer humaner und ein primärer Maus-Zelltyp als Testfälle herangezogen.

Weiterhin wird die Konsistenz der Ontologie wie von Gómez-Pérez vorgeschlagen überprüft [Gómez-Pérez 2001]: Um sicherzustellen, dass die deklarierten Axiome frei von Widersprüchen sind, wird die Ontologie mittels zweier Reasoner-Programme klassifiziert.

Tabelle 9: Techniken zur Evaluierung von Ontologien [Obrst et al. 2007]

TECHNIK	BESCHREIBUNG
Evaluierung der Nutzbarkeit in der Anwendung	Die anwendungsorientierte Evaluation wird eingesetzt um die praktischen Aspekte der Ontologie zu bewerten. Mögliche Techniken um die Anforderungen an die Ontologie zu charakterisieren, sind beispielsweise Use-Cases oder Szenarios.
Vergleich der Ontologie mit dem Anwendungsbereich	Die Ontologie wird mit anderen Ontologien oder Datenbanken der entsprechenden Wissensdomäne verglichen.
Bewertung durch Anwender nach einem Kriterienkatalog	Die Ontologie wird auf die Einhaltung eines spezifischen Kriterienkatalogs hin untersucht. Die Kriterien ergeben sich aus der entsprechenden Domäne.
Sprachverarbeitungs-techniken	Die Ontologie wird hinsichtlich ihrer Eignung für Aufgaben der Sprachverarbeitung (zum Beispiel Wissensextraktion) bewertet.
Bewertung gegenüber dem Stand der Technik	Für verschiedene Versionen der Ontologie wird eine Metrik berechnet, die beschreibt, wie gut sie die Realität der jeweiligen Domäne abbilden. Diese Technik misst, wie stark die Ontologie im Lauf der Entwicklungszyklen verbessert wird.
Akkreditierung und Zertifizierung	Die Ontologie kann formal akkreditiert oder zertifiziert werden, sofern standardisierte Kriterien festgelegt und entsprechende Organisationen etabliert sind.

4 ERGEBNISSE

Im nächsten Abschnitt wird das Referenzmodell beschrieben, welches Anforderungen an ein IT-System für Forschungsverbünde enthält, die einen allgemeinen Charakter besitzen. Anforderungen, die spezifisch für den SFB/TRR77 sind, werden ebenso wie ihre Umsetzung in eine IT-Architektur in Abschnitt 4.2 behandelt. In Abschnitt 4.3 werden schließlich die verbundspezifischen Metadatenspezifikationen in Form des kontrollierten Vokabulars sowie der Ontologie für Zelllinien vorgestellt.

4.1 REFERENZMODELL FÜR ANFORDERUNGEN

Im nächsten Abschnitt werden die erfassten Anforderungen anhand ihrer Quellen dargestellt. In den darauf folgenden Abschnitten werden die Ziele und Anforderungen konsolidiert und abstrahiert und somit das eigentliche Referenzmodell beschrieben.

4.1.1 *Quellenanalyse*

Eine mögliche Darstellung der Anforderungen ist, diese nach ihrer Herkunft zu gruppieren. In den nächsten Abschnitten werden daher die Anforderungen, welche aus den Projektunterlagen, von Projektleitern sowie aus den Erfahrungen mit einem Prototyp abgeleitet wurden, aufgezeigt.

Projektunterlagen

Die übergeordneten Ziele des Forschungsverbundes wurden aus den Projektbeschreibungen der Antragsunterlagen entnommen [Universität Heidelberg, Medizinische Hochschule Hannover 2009].

Oberstes Ziel eines Forschungsverbundes ist demnach die Durchführung des Verbundprojektes. Konkreter setzt sich dieses Ziel aus zwei Unterzielen zusammen: Zum einen müssen forschungsbezogene Daten erzeugt, zusammengeführt und verwaltet werden. Diese Daten werden oftmals nicht im Projekt erzeugt, sondern für dessen Durchführung genutzt. Umgekehrt werden im Projekt Daten erzeugt, die über den konkreten Bedarf hinaus Bedeutung haben. Das zweite Unterziel ist die Beantwortung konkreter Forschungsfragen der einzelnen Projekte. Hieraus ergibt sich wiederum das Ziel, die gesammelten Forschungsdaten im Bezug auf die Fragestellung zu analysieren.

Im Zusammenhang mit dem Ziel, die Daten zu verwalten, besteht weiterhin das Ziel, den Zugriff auf die Daten sowie deren Nutzung zu kontrollieren und zu steuern. Dies bezieht sich zum einen auf die Unterbindung unberechtigter Zugriffe auf die Daten, zum anderen auf die korrekte Verwendung der Daten durch zugriffsberechtigte Projektpartner.

Projektleiter

FRAGEBOGEN Aus den Rückläufen der Umfrage unter den Projektleitern (siehe Tabelle 2) konnten weitere Anforderungen abgeleitet werden. Die Ergebnisse sind in Abbildung 5 zusammengefasst. Insgesamt werden 18 ausgefüllte Fragebogen in die Auswertung einbezogen. Für drei Projekte werden mehrere Fragebogen zurückgegeben, da diese Projekte mehrere Projektleiter besitzen. Damit werden 14 der 22 Projekte durch die Fragebogen abgedeckt.

Die Frage nach den in den Projekten verwendeten Datenarten ergibt, dass Microarray-, Bild- und Textdaten zu fast gleichen Teilen verwendet werden. Daraus lässt sich ableiten, dass eine IT-Anwendung geeignet sein muss, flexibel mit diesen und weiteren Datenarten umzugehen. Weiterhin ergibt sich die Anforderung, Wege zur Integration dieser heterogenen Daten bereitzustellen.

Aus den unterschiedlichen Formaten, in denen die Daten vorliegen, ergeben sich weitere Anforderungen: Die Daten müssen in geeigneter Weise dargestellt werden, sodass ein verbundweiter Zugriff auf die Daten möglich ist. Dabei müssen die Daten gegebenenfalls in ein anderes Format überführt werden, oder es müssen Mechanismen bereitgestellt werden, die einen einheitlichen Zugriff erlauben.

Die zu verarbeitenden Daten liegen zu Beginn der Projektlaufzeit noch nicht vollständig vor, wie aus den Antworten auf die Frage „Wann werden die Daten erzeugt?“ hervorgeht. Da die Daten zumindest teilweise während der Laufzeit des Forschungsverbundes generiert werden, ergibt sich die Anforderung an das IT-System, mit neu erzeugten Daten umzugehen. Die meisten Projektleiter sind der Ansicht, dass die eigenen Daten für einige oder alle anderen Projekte innerhalb des Verbundes von Interesse sind. Daraus lässt sich die Anforderung ableiten, Daten verschiedenen Ursprungs bereitzustellen und integrieren zu können.

Die Mehrheit der Projektleiter stuft die eigenen Daten als vertraulich ein. Bis auf wenige Ausnahmen sind die Projektleiter jedoch bereit, die Daten mit anderen Projekten des Verbundes zu teilen, wenn auch zum Teil mit Auflagen. Somit ergibt sich die Anforderung an das IT-System, den Datenzugriff und die Datennutzung zu kontrollieren. Daraus ergeben sich wiederum folgende Anforderungen: Zum einen sind die Daten vor unberechtigtem Zugriff zu schützen und der Zugang zu kontrollieren. Zum anderen besteht aber auch das Erfordernis, auch bei *berechtigtem* Zugriff

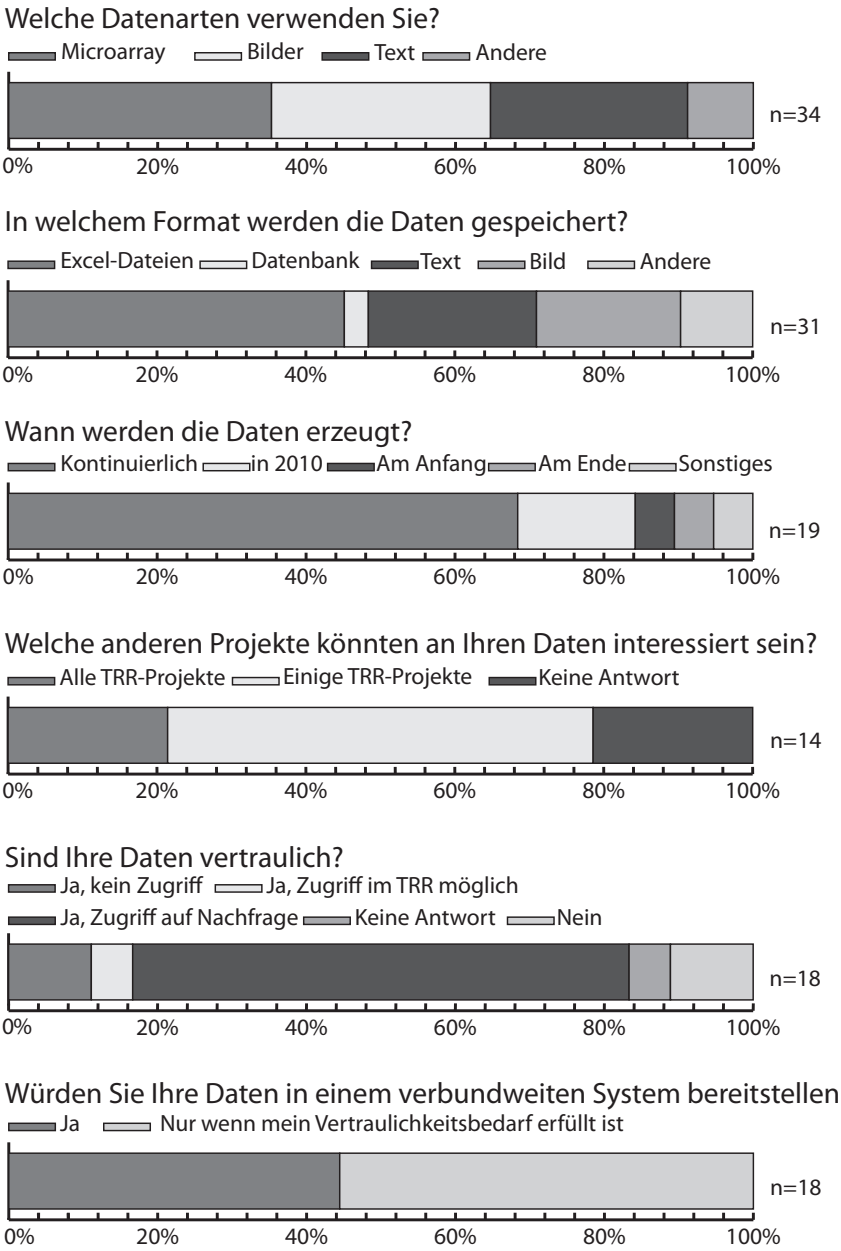


Abbildung 5: Auswertung einer Auswahl der Fragen des in Anhang A aufgeführten Fragebogens. Da mehrere Antworten zu einer Frage ausgewählt werden können, ergeben sich unterschiedliche Gesamtzahlen zu den Antworten.

Mechanismen zu etablieren, um die Nutzung der Daten nachvollziehbar zu gestalten und so Möglichkeiten zum Schutz des geistigen Eigentums der Urheber der Daten bereitzustellen.

INTERVIEWS Im Rahmen der Projektbesuche werden weitere Anforderungen erfasst. Wesentlicher Bedarf besteht darin, die Projektdaten zu analysieren. Hierzu bedienen sich die Projekte vielfältiger, meist statistischer, Methoden. Somit ergibt sich die Anforderung, geeignete Analysemethoden im IT-System definieren zu können. In der Regel ist eine einzelne Methode nicht ausreichend, um die vollständige Analyse durchzuführen, sodass es notwendig ist, mehrere Methoden zu kombinieren. Es ist somit erforderlich, einen zusammengesetzten *Analyseprozess* zu definieren. Die Steuerung dieser Prozesskette erfolgt in der Benutzerschnittstelle, ebenso wie die Anzeige des Analyseergebnisses. Zur Analyse der Daten sind weiterhin externe Datenquellen, wie beispielsweise Genomdatenbanken, einzubinden.

Prototyp

Die Erstellung und Evaluierung eines auf den zuvor erfassten Anforderungen basierenden Prototypen führt zu weiteren Anforderungen im Referenzmodell [Ganzinger, Noack, Diederichs et al. 2011]. Es stellt sich heraus, dass es sinnvoll ist, die Anforderung bezüglich der Darstellung der Daten in weitere Unteranforderungen zu unterteilen. Dies führt zu den folgenden Anforderungen: Zur automatisierten Verknüpfung und Auswertung der Daten ist die Spezifikation der Syntax der Daten sowie des zu Grunde liegenden Datenmodells erforderlich. Weiterhin sind Mechanismen zum automatisierten Identifizieren und Auffinden der Daten im Verbund erforderlich, sowie die Spezifikation der Semantik der einzelnen Datenquellen.

4.1.2 *Referenzziele*

Abbildung 6 zeigt die verallgemeinerten Ziele eines Forschungsverbundes in einem UML-Diagramm. Die Dokumentation der Ziele in Form von Tabellen befindet sich im Anhang C.1.

Die Ziele sind dem übergeordneten Ziel RZ1, *Forschungsprojekt durchführen*, nachgeordnet. Direkten Bezug zum Ziel RZ1 haben RZ2 und RZ3:

RZ2 Ziel eines Forschungsverbundes ist es, Forschungsfragestellungen zu beantworten. Diese Fragestellungen stellen den Kern dar, der zur Einrichtung des Verbundes geführt hat.

RZ3 Ein Forschungsverbund hat in der Regel das Ziel, Daten, die er zur Durchführung benötigt, zu erzeugen, zu speichern und abzurufen.

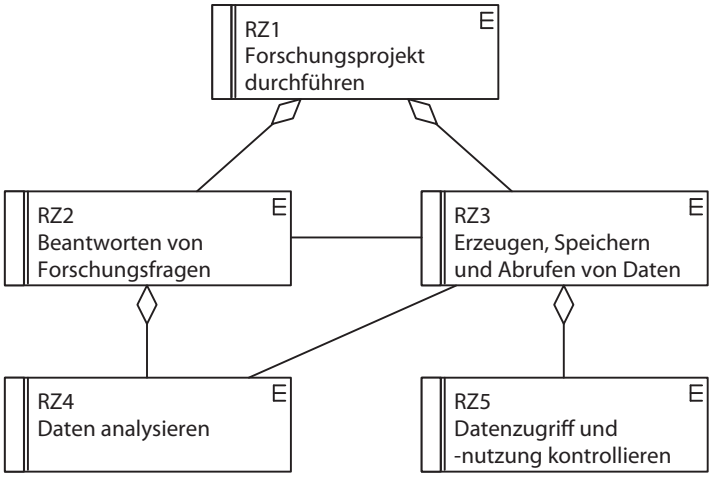


Abbildung 6: Referenzmodell für die Ziele eines Forschungsverbundes

RZ2 und RZ3 stehen miteinander in Beziehung: Ohne die Existenz entsprechender Daten ist die fundierte Bearbeitung der Fragestellungen nicht möglich. Im Sinne der Nachvollziehbarkeit der Arbeit des Verbundes müssen die Daten auch in geeigneter Weise für den längerfristigen Zugriff gespeichert und in passender Form abrufbar gehalten werden.

Die rohen Daten an sich sind jedoch nicht geeignet, die Forschungsfragestellungen zu beantworten. Vielmehr ist die Analyse der Daten und die Aufbereitung zu Wissen erforderlich. Somit ist RZ4 *Daten analysieren* dem Ziel, Forschungsfragen zu beantworten (RZ2) nachgeordnet, wobei es in Beziehung zu RZ3 steht, da die Daten des Projektes unmittelbar betroffen sind.

Weiterhin ist es das Ziel eines Forschungsverbundes, die Datennutzung zu kontrollieren und unberechtigten Zugriff zu unterbinden (RZ5). Dies umfasst rechtliche Aspekte des Datenschutzes, wenn die Forschungsarbeit beispielsweise patientenbezogene Daten umfasst. Weiterhin stellen die Daten eines Verbundes wertvolles geistiges Eigentum dar, das bis zur Veröffentlichung der Daten geschützt werden soll.

4.1.3 *Referenzanforderungen*

Die Anforderungen von Forschungsverbünden an die IT-Infrastruktur sind in Abbildung 7 zusammen mit den Zielen in einem UML-Diagramm dargestellt. Die Anforderungen lassen sich zwei Gruppen zuordnen: Der Teil *Datenrepräsentation* umfasst die Ziele und Anforderungen, welche die Erzeugung und Verarbeitung der Daten selbst betreffen. Daneben werden im Teil *Datenanalyse* diejenigen Aspekte zusammengefasst, welche die

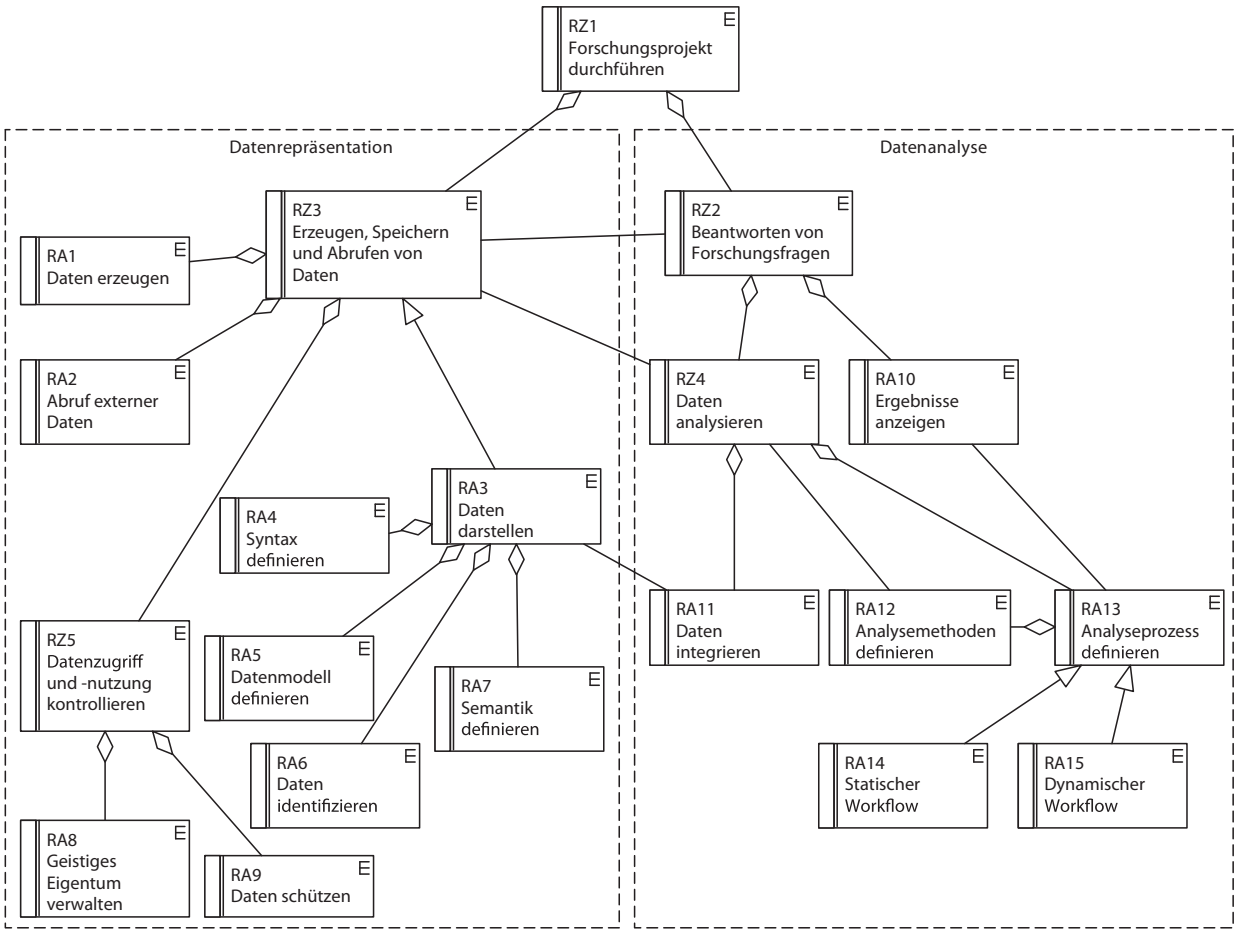


Abbildung 7: Referenzmodell für die Ziele und Anforderungen eines Forschungsverbundes

Auswertung der Daten zum Inhalt haben. Die tabellarische Beschreibung der Anforderungen ist im Anhang C.2 aufgeführt.

Die im vorherigen Abschnitt beschriebenen Ziele eines Forschungverbundes werden durch die Anforderungen verfeinert. RZ3 *Erzeugen, Speichern und Abrufen von Daten* enthält die Anforderungen *Daten erzeugen* (RA1) und *Abruf externer Daten* (RA2). Diese beiden Anforderungen berücksichtigen die möglichen Herkunftsarten für die vom Verbund benötigten Daten. Weiterhin besteht die spezialisierte Anforderung *Daten darstellen* (RA3), die sich auf die Betrachtung sämtlicher Daten des Verbundes bezieht. RA3 wird wiederum in den Anforderungen bezüglich der Syntax (RA4), des Datenmodells (RA5), der Identifikation (RA6) und der Semantik (RA7) präzisiert.

Das Ziel RZ5 *Datenzugriff und -nutzung kontrollieren* hat zwei Aspekte, die sich in den Anforderungen RA8 und RA9 widerspiegeln. RA8 stellt die Anforderung dar, dass die geistige Eigentümerschaft an den im Verbund

bereitgestellten Daten bei deren Verwendung durch andere berücksichtigt wird. Dies bedeutet, dass auch für Anwender, die berechtigterweise auf die Daten zugreifen, Regeln für die Nutzung bestehen, die, soweit möglich, durch das IT-System geprüft werden. Im Gegensatz dazu bezieht sich RA9 auf die Anforderung, die Daten des Verbundes vor dem Zugriff durch Unberechtigte zu schützen.

Die zweite Gruppe von Anforderungen deckt den Bereich *Datenanalyse* ab. Auf oberster Ebene befinden sich die beiden Ziele RZ2 *Beantworten von Forschungsfragen* und das nachgeordnete Ziel RZ4 *Daten analysieren*. Ziel RZ4 wird in die beiden Anforderungen RA11 *Daten integrieren* und RA13 *Analyseprozess definieren* zerlegt. Die Anforderung RA11 ist mit der Anforderung RA3 *Daten darstellen* assoziiert, da die technische Bereitstellung der Daten im Verbund von großer Relevanz für ihre Integration ist. RA13 besitzt die Unteranforderung RA12 *Analysemethoden definieren*, welche zum methodischen Hintergrund des Analyseprozesses führt. Die Anforderungen RA14 und RA15 beschreiben zwei verschiedene Ausprägungen der Anforderung RA13 *Analyseprozess definieren:* RA14 betrachtet den *statischen Workflow*, bei dem die Prozessschritte fest vorgegeben sind. Die Abfolge der Analyseschritte und die verwendeten Datenquellen können hier durch den Anwender nicht verändert werden. Im Gegensatz dazu behandelt RA15 die Anforderung des *dynamischen Workflows*. In diesem Fall werden die einzelnen Datenquellen und Analyseschritte durch den Benutzer zur Laufzeit zusammengestellt. Da vorab nicht bekannt ist, welche Datenarten miteinander verknüpft werden, impliziert diese Anforderung auch einen höheren Anspruch an die semantische Interoperabilität. Die Daten müssen für diesen Fall so annotiert sein, dass eine automatische Transformation zum Angleich der Datenfelder möglich ist. Anforderung RA13 ist weiterhin mit der Anforderung RA10 *Ergebnisse anzeigen* assoziiert. Die Anforderung RA10 *Ergebnisse anzeigen* stellt die Anforderung, die Analyseergebnisse geeignet darzustellen und ist damit eine Teilanforderung zu Ziel RZ2 *Beantworten von Forschungsfragen*.

4.2 IT-ARCHITEKTUR FÜR DEN FORSCHUNGSVERBUND

Für den SFB/TRR77 wird die Datenintegrationsplattform PELICAN entwickelt. Die IT-Architektur für dieses System wird aus den verbundspezifischen Anforderungen, welche im nächsten Abschnitt aufgezeigt werden, abgeleitet. Darauf aufbauend wird dann die IT-Architektur in Form eines Komponenten- sowie eines Verteilungsmodells vorgestellt.

4.2.1 *Verbundspezifische Anforderungen*

Im nächsten Abschnitt werden die Ziele des SFB/TRR77 betrachtet. In den daran anschließenden Abschnitten folgen, thematisch zusammengefasst, die entsprechenden verbundspezifischen Anforderungen an die IT-Architektur. Die Ziele und Anforderungen sind in Anhang D in Tabellenform nach dem beim Referenzmodell eingeführten Muster (siehe Tabelle 4) vollständig aufgeführt.

Ziele des Verbundes

Zunächst ergeben sich die spezifischen Ziele des SFB/TRR77: Das Ziel Z1 spiegelt den Forschungsauftrag des Verbundes wieder, ein tiefes Verständnis der molekularen Basis der Leberkrebsentstehung zu erhalten. Dies umfasst zu Beginn die Betrachtung der chronischen Lebererkrankung und führt über die Progression zum metastasierenden Krebs. Weiterhin ist die Identifizierung neuer vorbeugender, diagnostischer und therapeutischer Ansätze Ziel des Verbundes.

Aus diesem Hauptziel ergibt sich ein konkretes Ziel bezüglich der erforderlichen Daten (Ziel Z2). Da molekulare Prozesse eine wesentliche Rolle im Verbund spielen, sind genomische Daten aus Microarrays von zentraler Bedeutung für den Verbund. Sie werden ergänzt durch Bilddaten wie beispielsweise TMA-Daten und klinische Daten.

Ziel Z3 *Beantwortung von Forschungsfragen* wird durch folgende exemplarische Fragestellungen zum Leberkrebs konkretisiert:

- Welche allgemeinen oder spezifischen Mechanismen der chronischen Leberkrankheiten, insbesondere der chronischen Virusinfektion und der entzündungsvermittelten Prozesse prädisponieren oder initiieren Leberkrebs?
- Welche molekularen Schlüsselereignisse, die Leberkrebs fördern oder aufrecht erhalten, können als Tumormarker dienen oder stellen aussichtsreiche Angriffsziele für zukünftige therapeutische Interventionen dar?

Ein weiteres Ziel des Verbundes ist es, seine Daten so zu schützen, dass der Zugriff, je nach Datenart, nur für bestimmte Mitarbeiter des Verbundes, alle Mitarbeiter des Verbundes oder die Öffentlichkeit möglich ist. Schließlich besteht noch das Ziel Z5, laut dem die Daten der Projekte des Verbundes für die projektübergreifende Auswertung bereitgestellt werden müssen.

Erhebung der Datenarten

Abbildung 5 enthält einen Überblick über die Antworten der Projektleiter auf den Fragebogen. An projektspezifischen Datenarten werden im SFB/

TRR77 vorrangig Microarraydaten (n=12), gefolgt von Bilddaten (n=10) genannt. Bezüglich der verwendeten Dateiformate dominieren Excel-Dateien (n=14) die Antworten, gefolgt von Textdateien (n=7) und Bilddateien (n=6). Die Diskrepanz zwischen der hohen Anzahl der Nennungen von Bilddaten als Datenart und der deutlich niedrigeren Anzahl der Nennungen von Bilddateien als Dateiformat lässt sich folgendermaßen erklären: Viele der in der biomedizinischen Forschung angewandten Analysemethoden führen zur Erzeugung eines Bildes. Relevant ist letztlich jedoch nur die Interpretation des Bildes, die im Text- oder Excel-Format gespeichert werden kann. Beispielsweise werden histologische Schnittbilder vom Pathologen begutachtet und diagnostiziert. Die Ergebnisse dieses Vorgangs werden dokumentiert und den Projekten im Verbund bereitgestellt. Das Schnittbild selbst ist nach der Begutachtung nur noch von untergeordneter Bedeutung.

Datenschutz und Vertraulichkeit

Im Rahmen der Anforderungserhebung mithilfe von Fragebogen (siehe Abschnitt 4.1.1) geben die Projektleiter der SFB/TRR77-Projekte an, nur dann Daten in eine verbundweite Plattform abgeben zu wollen, wenn die Sicherheit der Daten gewährleistet ist. Mit n=12 Projekten stellt ein großer Teil der Projekte die Anforderung, mitbestimmen zu können, wer auf die eigenen bereitgestellten Daten zugreifen darf. Insgesamt sind alle Projekte bereit, Daten verfügbar zu machen, wenn auch bisweilen nur in einem geschützten Umfeld. Somit ist eine wesentliche Anforderung im SFB/TRR77, dass die Kontrolle über die Daten bei den Projekten verbleiben kann.

Analyse der Daten

Bezüglich der eingesetzten Analysemethoden werden hauptsächlich statistische Verfahren wie t-Test, Wilcoxon-Mann-Whitney-Test oder Analysis of Variance (ANOVA) genannt. Auch Programme für die Durchführung von statistischen Berechnungen wie Excel und SPSS (*http://www.ibm.com/software/de/analytics/spss/*, Stand 10.12.2013, 15:34) werden hier genannt. Für die Analyse der Microarray-Daten werden spezielle Analyseprogramme wie GeneSpring GX (*http://genespring-support.com/*, Stand: 10.12.2013, 15:32) angeführt.

4.2.2 *Technologieauswahl*

Die verbundspezifischen Anforderungen haben direkten Einfluss auf die Auswahl der Komponenten, die zur Umsetzung von PELICAN eingesetzt werden können. Für dieses System grundlegend ist die Entscheidung, ob die Daten der Teilprojekte in einem gemeinsamen Datenbestand zentral zusammengeführt werden, oder dezentral bei den Projekten verbleiben.

Aus dieser Auswahl wiederum ergibt sich unmittelbar, welches Bioinformatikframework (i2b2 oder caBIG, siehe Abschnitt 2.5.4) die geeignete Grundlage für das zu erstellende System darstellt.

Von den Projekten des Verbundes wird die Anforderung, die Kontrolle über die eigenen Daten zu behalten, als besonders bedeutsam herausgestellt. Dies ist als wichtiger Indikator für die Erstellung einer Architektur mit zumindest konzeptionell verteilten Datenbeständen zu werten. Daher wird für das System die grundlegende Entscheidung getroffen, eine dezentrale serviceorientierte Architektur zu erstellen. Folglich wird caBIG als Bioinformatikframework zu Grunde gelegt.

4.2.3 *Abbildung der Anforderungen auf Systemeigenschaften*

Die im Abschnitt 4.2.1 identifizierten Systemanforderungen werden auf Systemeigenschaften abgebildet. Da das System aus verschiedenen Komponenten besteht, werden die Eigenschaften wiederum den Komponenten zugeordnet. Die Darstellung der Zuordnungen ist nach den Zielen (siehe Abbildung 6) gegliedert, wobei das Ziel Z1 ausgelassen wird, da es keine direkt zugeordneten Anforderungen besitzt.

Tabelle 10 fasst zusammen, wie die Anforderungen zu Ziel Z2 auf die entsprechenden Komponenten und Eigenschaften des Systems abgebildet werden. Zur Einbindung neu erzeugter Daten wird die Systemeigenschaft *Automatisierte Erstellung von Datendiensten* benötigt. Diese Eigenschaft ist nicht als Eigenschaft einer Komponente verfügbar, vielmehr bedarf es einer eigenen Entwicklung auf der Basis eines *Datendienst-Frameworks*. Ein solches Framework unterstützt die Entwicklung von Diensten durch die Bereitstellung von Werkzeugen, beispielsweise zur automatisierten Erzeugung von Webservices auf der Basis von UML-Modellen.

Die Anforderung A2, *Abruf externer Daten*, wird zum Teil auf die Komponente caCORE SDK abgebildet. Einige Aspekte der Einbindung externer Daten werden jedoch von der Portal-Komponente bereitgestellt. Die Anforderungen A3, A4 und A5 werden jeweils vollständig auf Eigenschaften der Komponente *Datendienst-Framework* abgebildet. Anforderung A6 wird auf die Systemeigenschaft *Metadatenverzeichnis* abgebildet. Die für Anforderung A7 erforderliche Systemeigenschaft *Definition von kontrolliertem Vokabular und Ontologien* bedarf der eigenen Entwicklung, da diese Vorgaben zur Erstellung von Metadaten darstellen, die für den Forschungsverbund spezifisch sind. Bereitgestellt werden Vokabular und Ontologien über die Komponente *Terminologieserver*.

Die zu Ziel Z4 *Datenzugriff und -nutzung kontrollieren* gehörigen Abbildungen sind in Tabelle 11 dargestellt. Anforderung A8 wird auf die Systemeigenschaft *Protokollierung der Datennutzung* abgebildet. Dies geschieht zum einen in der Komponente Portal, zum anderen beim *Datendienst*. Im Bezug auf den Schutz vor unberechtigtem Zugriff auf die Daten wird die

Tabelle 10: Abbildung der Anforderungen zu Z2 *Erzeugen, Speichern und Abrufen von Daten* auf Systemeigenschaften und Komponenten

ANFORDERUNG	EIGENSCHAFT	KOMPONENTE
A1 Daten erzeugen	Automatisierte Erstellung von Datendiensten	Datendienst-Framework
A2 Abruf externer Daten	Einbindung externer Datendienste	Datendienst-Framework, Portal
A3 Daten darstellen	Datendienst, Dokumentationsdienst	Datendienst-Framework, Dokumentenmanagementsystem
A4 Syntax definieren	Dienstbeschreibung	Datendienst-Framework
A5 Datenmodell definieren	Erstellung des Datenmodells	Datendienst-Framework
A6 Daten identifizieren	Bereitstellung der Lokalisierungsinformationen	Metadatenverzeichnis
A7 Semantik definieren	Definition von kontrolliertem Vokabular und Ontologien	eigene Entwicklung, Terminologieserver

Tabelle 11: Abbildung der Anforderungen zu Z4 *Datenzugriff und -nutzung kontrollieren* auf Systemeigenschaften und Komponenten

ANFORDERUNG	EIGENSCHAFT	KOMPONENTE
A8 Geistiges Eigentum verwalten	Protokollierung der Datennutzung	Portal/Datendienst
A9 Daten schützen	Benutzerauthentisierung	Portal/Sicherheitsdienst
	Benutzerautorisierung	Portal/Datendienst/ Sicherheitsdienst

Anforderung A9 *Daten schützen* auf zwei Systemeigenschaften abgebildet: Die *Benutzerauthentisierung* ist der Komponente *Portal* zugeordnet, während die *Benutzerautorisierung* sowohl in der Komponente *Portal* als auch beim *Datendienst* umgesetzt wird.

Die Anforderungen zu den Zielen Z3 *Beantworten von Forschungsfragen* und Z5 *Daten analysieren* sind schließlich in Tabelle 12 zusammengefasst. Die Anforderung A10 wird auf die von der Komponente *Portal* bereitgestellte Eigenschaft, datenspezifische Portlets darzustellen, abgebildet.

Anforderung A11 *Daten integrieren* wird auf die Eigenschaft *Datendienst* des Systems abgebildet. Diese Systemeigenschaft kann jedoch nicht einer einzelnen Komponente zugeordnet werden. Je nach Art und Komplexität der Analyse wird der Analysedienst von einer der Komponenten *Portlet, Statistiksystem* oder dem *Datendienst-Framework* bereitgestellt. Die An-

Tabelle 12: Abbildung der Anforderungen zu Z3 *Beantworten von Forschungsfragen* und Z5 *Daten analysieren* auf Systemeigenschaften und Komponenten

ANFORDERUNG	EIGENSCHAFT	KOMPONENTE
A10 Ergebnisse anzeigen	Datenspezifische Portlets	Portal
A11 Daten integrieren	Analysedienste	Portlet, Statistiksystem, Datendienst-Framework
A12 Analysemethoden definieren	Statistikmethoden	Statistiksystem
A13 Analyseprozess definieren	Dokumentationsdienst	Dokumentenmanagement-system
A14 Statischer Workflow	Workflow in Portletanwendung	Portal
A15 Dynamischer Workflow	Flexible Pipeline	Pipeline Management

forderung A12 hingegen wird auf die Eigenschaft *Statistikmethoden* des Statistiksystems abgebildet.

Auf die Systemeigenschaft *Dokumentationsdienst* wird Anforderung A13 abgebildet. Die für die Analyseprozesse erforderlichen Schritte werden zunächst festgelegt und in der Komponente *Dokumentenmanagementsystem* dokumentiert. Anforderung A14 *Statischer Workflow* wird als Unteranforderung von A12 auf die Eigenschaft der Komponente Portal, statische Prozesse durchführen zu können, abgebildet. Ebenso wird A15 auf die Eigenschaft *Flexible Pipeline* der Komponente *Pipeline Management* abgebildet.

4.2.4 *Komponentenmodell*

Die im vorangegangenen Abschnitt identifizierten Komponenten werden in einem Komponentenmodell dargestellt. In diesem Modell sind die Komponenten zunächst mit generischen Bezeichnungen versehen. Abbildung 8 stellt eine vereinfachte Sicht auf das Komponentenmodell in Form eines Komponentendiagramms in UML-Notation dar. In diesem Modell nimmt das Portal einen zentralen Platz ein, da es den Anwendern die Benutzerschnittstelle bereitstellt. Weiterhin agiert das Portal den Diensten gegenüber als Servicekonsument, der die Daten- und Analysedienste nutzt, um im Auftrag des Benutzers Abfragen durchzuführen und Ergebnisse darzustellen. Tabelle 13 enthält eine Übersicht über die Abbildung der generischen Komponenten auf konkrete Umsetzungskomponenten.

Zur vollständigen Durchführung aller erforderlichen Aufgaben innerhalb des SOA-Systems sind weitere Dienste erforderlich, die der Infrastruktur zuzuordnen sind. In Abbildung 8 sind exemplarisch die Dienste *Vokabulardienst* und *Metadatenverzeichnis* dargestellt. Weitere Infrastrukturdienste werden im Zuge der Betrachtung der Umsetzungskomponenten in den folgenden Abschnitten vorgestellt.

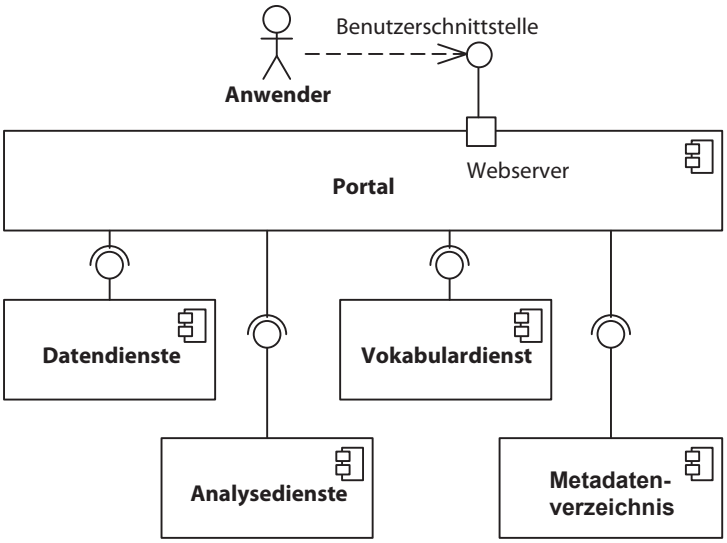


Abbildung 8: Komponentendiagramm für PELICAN in UML-Notation.

Portal

Das Portal ist diejenige Komponente, welche die Benutzerschnittstelle bereitstellt. Es wird im SFB/TRR77-Verbund durch die Implementierung der Open-Source-Software *Liferay* (*http://www.liferay.com*, Stand: 27.11.2013, 14:15) umgesetzt. Liferay stellt eine Reihe von Funktionen zur Verfügung, die verschiedene Komponenten des Komponentenmodells betreffen. Daher wird die Komponente *Portal* in Untereinheiten zerlegt, wie in Abbildung 9 dargestellt.

Die einzelnen Unterkomponenten sind *Portlets*, *Dokumentenmanagement* und *Benutzermanagement*. Sie werden in den folgenden Abschnitten erläutert.

DATENSPEZIFISCHE PORTLETS Liferay unterstützt die Erstellung von Portlets nach der *Java Portlet Specification* [JSR 286, 2008]. Diese Eigenschaft wird dazu verwendet, um für die spezifischen Datentypen des Verbundes Anzeigemodule zu schaffen. In PELICAN werden zu sämtlichen Typen genomischer Datendienste (aCGH, Genexpression, Methylierung) spezifische Portlets zur Anzeige bereitgestellt. Um aus allen Datendiensten parallel Daten zum gleichen Gen abzurufen, erfolgt die Eingabe des Gensymbols in einem separaten Portlet, welches mit den Anzeigeportlets über die Inter-Portlet-Communication nach [JSR 286, 2008] kommuniziert.

Die Interaktion mit den Datendiensten erfolgt auf logischer Ebene unter Verwendung der CQL: Die jeweilige Dienstabfrage wird als XML-Dokument gemäß der CQL-Spezifikation an den Dienst gesandt. Als Ergebnis

Tabelle 13: Konkretisierung der Komponenten des Komponentenmodells durch Zuordnung zu Umsetzungskomponenten

GENERISCHE KOMPONENTE	UMSETZUNGSKOMPONENTE
Portal	Liferay
Datendienst-Framework	caCORE SDK
Metadatenverzeichnis	Eigenentwicklung (caCORE SDK), Nationales Metadatenverzeichnis (geplant)
Terminologieserver	TemaTres
Sicherheitsdienst	caCORE SDK, LDAP
Statistiksystem	R
Dokumentenmanagementsystem	Alfresco
Pipeline Management	Galaxy

wird ein XML-Dokument mit einer Liste derjenigen Datensätze zurückgeliefert, die den Kriterien der Abfrage entsprechen. Das XML-Dokument wird in Java-Objekte umgewandelt, sodass die Ergebnisse schließlich im Portlet zur Aufbereitung für die Darstellung bereitstehen.

Für statische Workflows (Anforderung A14) erfolgt die Implementierung des Prozesses in einem Workflow-spezifischen Portlet. Somit steht ein statischer Workflow zur Verfügung, der über das im Portalserver angezeigte Portlet durch den Anwender parametrisiert werden kann, sofern dies vom jeweiligen Workflow unterstützt wird. Exemplarisch wird dies in einem Portlet zur Analyse differenziell exprimierter Gene in Microarray-Daten umgesetzt [Ganzinger, Kesselmeier et al. 2012].

DOKUMENTENMANAGEMENTSYSTEM Das Dokumentenmanagement wird durch das Produkt *Alfresco* (*http://www.alfresco.com*, Stand: 27.11.2013, 23:10) umgesetzt [Berman et al. 2012]. Diese Funktion wird also nicht durch die Komponente *Liferay* erbracht, sondern an Alfresco delegiert. Die Benutzerschnittstelle kann dabei wahlweise in das Liferay-Portal integriert werden, oder über einen eigenen URI angesprochen werden. In diesem Fall benutzt Alfresco denselben Sicherheitsdienst wie der Liferay-Portalserver, sodass der Benutzer lediglich eine Kennung für beide Systeme benötigt.

Das Dokumentenmanagementsystem stellt Funktionen bereit, die es den Benutzern erlauben, eigene Dokumente oder Dateien im Verbund verfügbar zu machen. Dies ist besonders relevant für die gemeinsame Nutzung von vorläufigen Daten, Zwischenergebnissen und Analyseprotokollen. Als weitergehende Funktionen bietet Alfresco die Möglichkeit, Diskussionsforen, Wikis [Ebersbach, Glaser 2005] oder Blogs [Nardi et al. 2004] zu

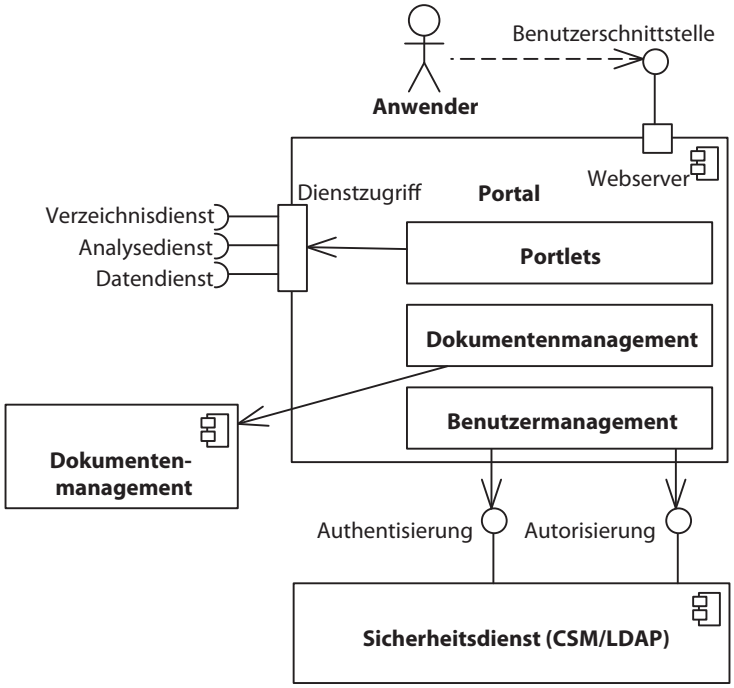


Abbildung 9: Struktur der Komponente *Portal*

erstellen, sodass auf einfache Weise Erkenntnisse und Erfahrungen auch zu einem frühen Zeitpunkt innerhalb des Netzwerks geteilt werden können. Dem Autor eines Wiki- oder Blog-Eintrags steht es dabei frei, die Zugriffsberechtigung auf eine Teilmenge der Angehörigen des SFB/TRR77 einzuschränken.

BENUTZERMANAGEMENT Der Portalserver besitzt ein integriertes Benutzermanagement, durch das die Authentisierung der Anwender sowie deren Autorisierung für den Zugriff auf die einzelnen Portlets vollzogen wird. Weiterhin umfasst das Benutzermanagement des Liferay-Portals Funktionen zur Verwaltung der Benutzereinstellungen. Da die Identität der Benutzer auch bei den übrigen Komponenten der SOA-Umgebung bekannt sein muss, wird die Verwaltung der eigentlichen Identitätsinformationen, insbesondere von Benutzername, Passwort und Gruppenzugehörigkeit, an die LDAP-Komponente des Sicherheitsdienstes delegiert. Die Verwaltung der Zugriffsberechtigungen für die Portalseiten und Portlets verbleibt beim Portal.

Datendienste

Für alle im SFB/TRR77 anfallenden Microarray-Datensätze (siehe Abschnitt 2.3.1) werden mit dem caCORE SDK die Datendienste zur Einbindung in PELICAN erzeugt. Da sich die Datenmodelle je nach Arraytyp unterscheiden, wird für jeden Typ ein eigenes UML-Modell erstellt. Abbildung 10 zeigt exemplarisch die Darstellung des Modells als UML-Klassendiagramm. Das Modell ist gemäß den Vorgaben der caCORE SDK-Dokumentation annotiert, indem den Elementen des Modells sogenannte *Custom Tag Values* hinzugefügt werden, die im folgenden Schritt vom caCORE-Generator interpretiert werden.

Im Rahmen der Generierung einer Reihe von Datendiensten hat sich herausgestellt, dass die Datenmodelle für die genomischen Datensätze sich zwar im Bezug auf Typ und Anzahl der Datenelemente unterscheiden, aber von der Struktur her sehr ähnlich sind: In der Regel enthalten die verwendeten Excel-Tabellen Einträge, die entweder *Metadaten* oder *Messdaten* enthalten. Um von dieser Ähnlichkeit zu profitieren habe ich einen *Modellgenerator* entworfen, der für diese Klasse von Datendiensten die Erzeugung des erforderlichen UML-Modells teilweise automatisiert. Die Umsetzung des Modellgenerators erfolgte im Rahmen einer von mir betreuten Diplomarbeit [Ganzinger, Senghas et al. 2013].

Bei der Nutzung des Modellgenerators wird die ursprüngliche Excel-Datei analysiert und visualisiert. Der Benutzer muss nun lediglich die Zeilen und Spalten einer der beiden Klassen *Metadaten* und *Messdaten* zuordnen. Aus diesen Informationen erzeugt der Modellgenerator eine XMI-Datei, die den Anforderungen des caCORE SDK-Generators entspricht. Insbesondere enthält diese Datei bereits die vollständige Annotation mit *Custom Tag Values*, sodass die aufwändige und fehlerträchtige manuelle Annotation entfällt. Diese Datei kann dann, analog zu manuell erstellten Modellen, durch den caCORE SDK-Generator verarbeitet werden. Weiterhin wird eine Data Definition Language (DDL)-Datei mit einem zum Datenmodell passenden Datenbankschema erzeugt. Mit diesem Schema wird eine MySQL-Datenbank erzeugt, in der die Daten des Datendienstes abgelegt werden.

Metadatenverzeichnis und Terminologieserver

Die naheliegende Wahl für das Metadatenverzeichnis ist das zum caBIG gehörige Metadatenverzeichnis caDSR. Daher wird caDSR auf die Verwendbarkeit in PELICAN hin untersucht. Dabei werden die folgenden Probleme identifiziert:

1. Struktur: caDSR wird als zentraler Metadatendienst konzipiert, von dem es nur eine Instanz gibt, die von den National Institutes of Health (NIH) betrieben wird. Eigene Elemente müssen über einen

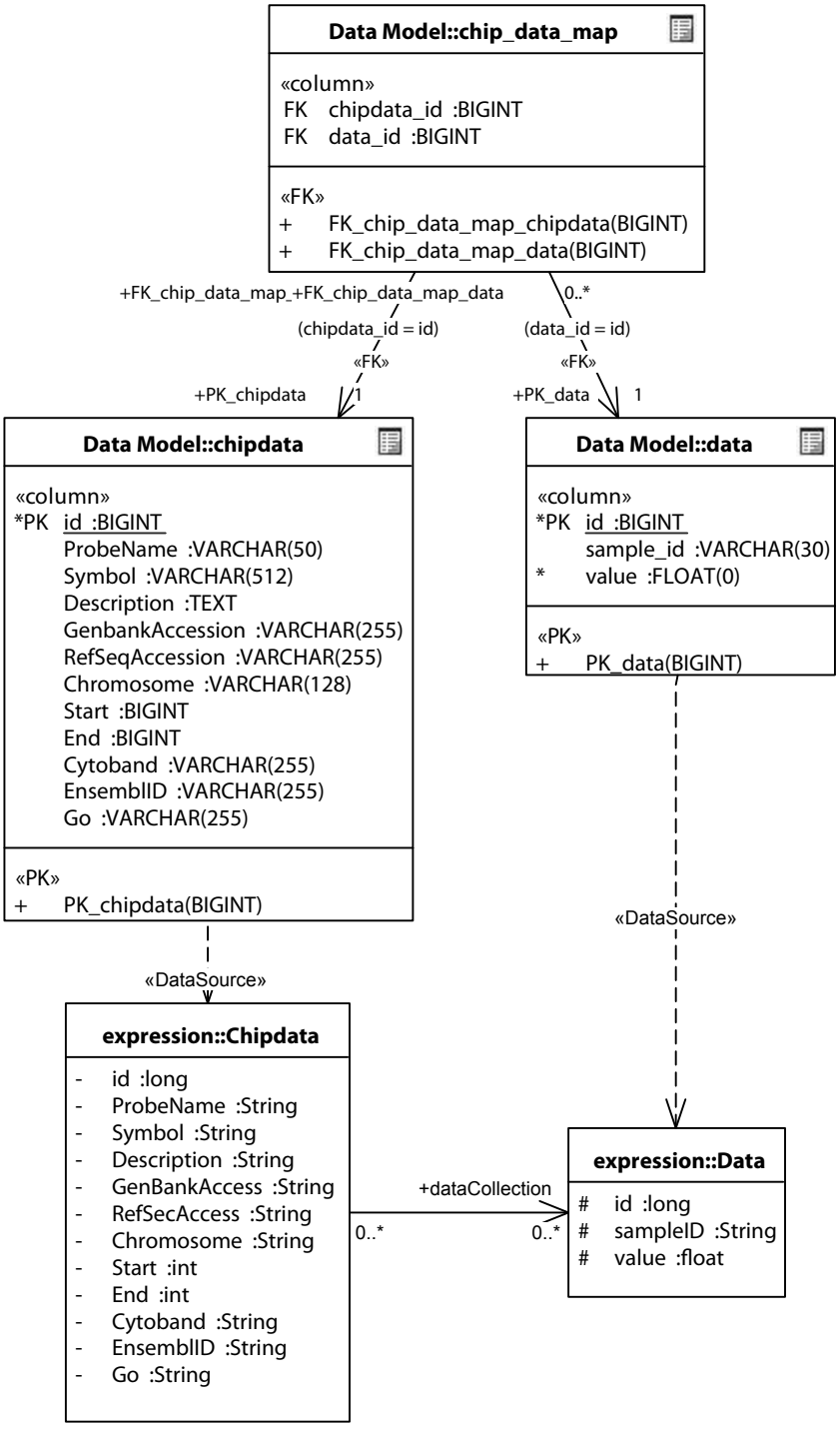


Abbildung 10: caCORE SDK konformes UML-Diagramm des Datendienstes für die Bereitstellung von Microarray-Daten zur Genexpression

Kuratierungsprozess vom NCI eingepflegt werden. Der Betrieb einer eigenen caDSR-Installation ist zwar grundsätzlich möglich, aber sehr aufwändig.

2. Verfügbarkeit: Die Software für caDSR ist, wie die übrigen caBIG-Komponenten auch, als Open-Source-Software frei verfügbar. Allerdings ist zum Betrieb von caDSR eine kommerzielle Lizenz des Oracle Application Servers erforderlich.

Aufgrund dieser Aspekte wird die Verwendung von caDSR für PELICAN verworfen. Zukünftig könnte erwogen werden, zumindest für einen Teil der Aufgaben des Metadatenverzeichnisses das *nationale Metadatenverzeichnis*, welches sich derzeit im Aufbau befindet, zu nutzen [Ngouongo et al. 2013; Stausberg et al. 2009].

Für PELICAN wird daher ein Metadatenverzeichnis mit eingeschränktem Funktionsumfang entwickelt. Es wird durch einen caBIG-Datendienst, der mit dem caCORE SDK erstellt wird, umgesetzt. Dieser Datendienst gibt auf Anfrage den zu einer Dienstkennung gehörigen URI des entsprechenden Webservice-Endpunktes zurück. Somit müssen die Endpunkte der von einer Applikation verwendeten Dienste nicht statisch in der Anwendung hinterlegt sein, sondern es erfolgt erst zur Laufzeit die Verbindung zu einem spezifischen Service-Endpunkt.

Die von caDSR ebenfalls abgedeckte Funktion der Verwaltung von kontrollierten Vokabularen wird als Vokabulardienst in PELICAN von einer separaten Komponente bereitgestellt. Hierzu wird die Vokabularserver-Software *TemaTres* implementiert.

Orchestrierung

Im Kernsystem von PELICAN werden vordefinierte Workflows verwendet, die keine komplexen Abhängigkeiten zwischen den einzelnen Prozessschritten besitzen [Ganzinger, Kesselmeier et al. 2012]. Aus diesem Grund ist an dieser Stelle eine dynamische Orchestrierungskomponente zur Steuerung des Workflows nicht zwingend erforderlich. Es ist hier vielmehr ausreichend, mit Hilfe des Metadatenverzeichnisses geeignete Dienste auszuwählen und dynamisch einzubinden.

Für bestimmte Problemstellungen sind die Anforderungen an die Prozesssteuerung komplexer. Beispielsweise gibt es bei der Analyse von Next Generation Sequencing (NGS)-Daten viele Möglichkeiten die Analysepipeline zusammenzustellen. Dabei variieren nicht nur die vorgesehenen Teilschritte selbst, es existieren auch mehrere alternative Komponenten, die einen Teilschritt umsetzen können. Je nach Konfiguration der Pipeline ergeben sich Zwischenergebnisse in unterschiedlichen Datenformaten, die vom Pipelinemanagementsystem entsprechend aufbereitet werden müssen, sodass sie als semantisch und syntaktisch korrekte Eingabedaten für den folgenden Prozessschritt dienen können.

Für den Forschungsverbund SFB/TRR77 wird das Pipelinemanagement durch die Komponente Galaxy durchgeführt [Goecks et al. 2010]. Diese Software wurde im Rahmen einer von mir betreuten Bachelorarbeit anhand verbundspezifischer Kriterien im Vergleich mit alternativen Workflowmanagementprogrammen ausgewählt.

Statistiksystem

Als Statistiksystem wird die Open-Source-Software *R* in PELICAN eingebunden [R Development Core Team 2011]. Somit stehen umfangreiche Statistikmethoden zur Verfügung, die teilweise speziell für den Bereich der Biomedizin entwickelt wurden [Gentleman et al. 2004]. R stellt seine Analysedienste über die R-spezifische Schnittstelle *Rserve* für den Zugriff über das Netzwerk bereit [Urbanek 2003]. Somit können die Daten aus den Datendiensten extrahiert und über den R-Dienst weiterverarbeitet werden [Ganzinger, Kesselmeier et al. 2012].

Die Ergebnisse der Analysen werden an ein entsprechendes Portlet innerhalb des Portalservers zurückgegeben und dort dargestellt. Je nach Art der Analyse wird das Ergebnis in Form einer Wertematrix zur Darstellung in Tabellenform übertragen (beispielsweise für aCGH-Daten), oder als Diagramm. Zur Erzeugung der Grafiken stehen dabei die Grafikpakete von R zur Verfügung. Genexpressions- und Methylierungsdaten werden beispielsweise in Form von Boxplots ausgegeben.

4.2.5 *Verteilungsmodell*

Die Softwarekomponenten des PELICAN-Systems werden den virtualisierten Hardwareressourcen zugeordnet. Abbildung 11 gibt einen Überblick über diese Zuordnung in Form eines UML-Verteilungsdiagramms für das PELICAN-System. Die im Diagramm dargestellten Assoziationen stellen die wichtigsten Kommunikationsverbindungen über das Netzwerk dar. Dabei spielt die Virtualisierung der Hardware aus Sicht der Verteilungsmodellierung keine Rolle. Alle Rechnerknoten können wahlweise als virtuelle oder physische Server umgesetzt werden. Für PELICAN werden die gezeigten Knoten als virtuelle Maschinen auf den beiden physischen Servern umgesetzt.

Auf Grund der größeren Ressourcenanforderungen werden die Komponenten *Portal* und *Dokumentenmanagement* jeweils auf eigenen Knoten installiert. Generell werden alle Hilfskomponenten, wie Datenbanksysteme, die ausschließlich bestimmten Komponenten zuzuordnen sind, ebenfalls auf den entsprechenden Knoten installiert. Ein eigener Knoten steht für die Sicherheitsinfrastruktur, bestehend aus dem LDAP-Server sowie dem PDP/PAP-Server bereit. Dadurch wird zum einen die gemeinsame Nutzung dieser Dienste vereinfacht, zum anderen ist aber auch eine bessere

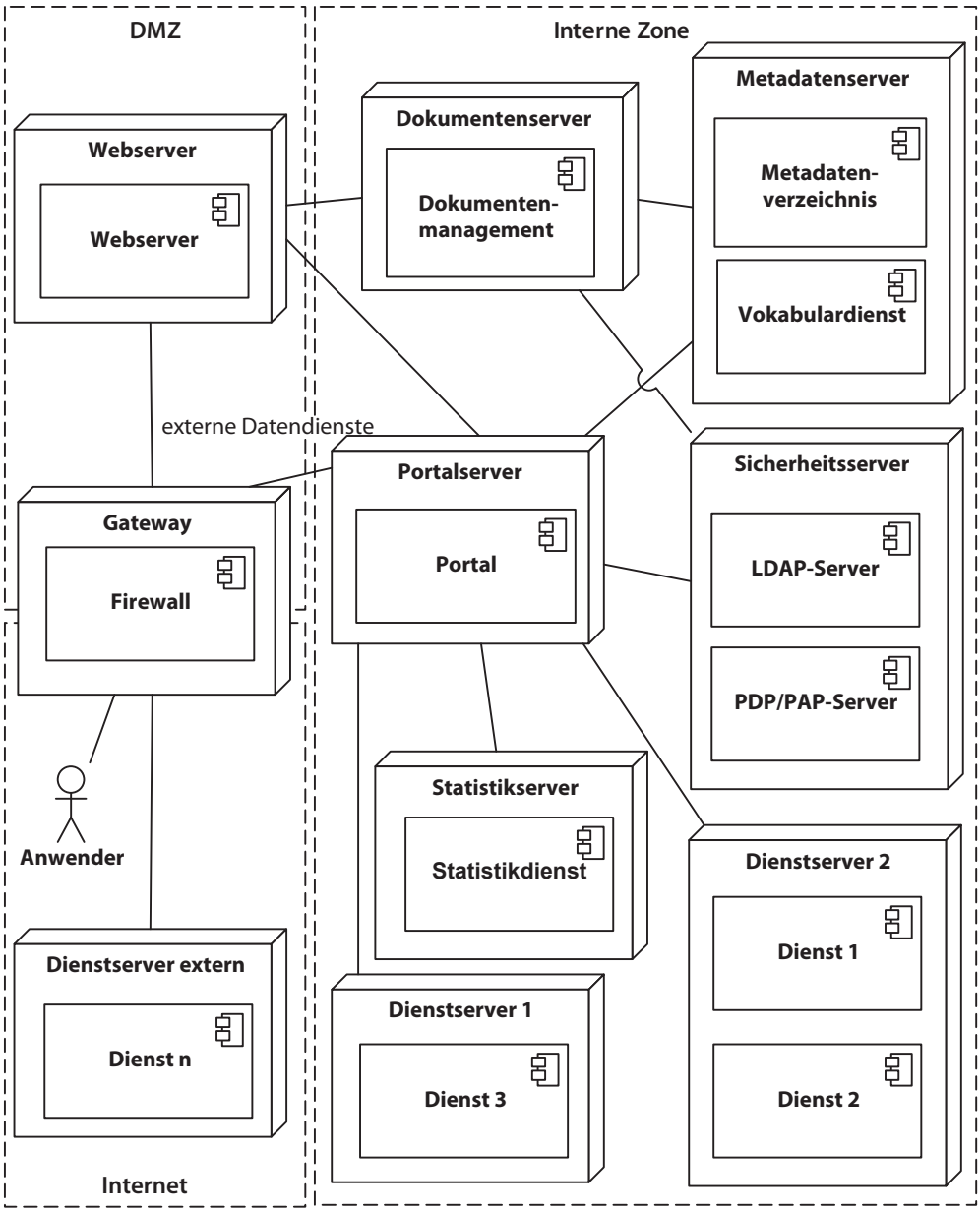


Abbildung 11: Verteilungsdiagramm der Komponenten von PELICAN in UML-Notation. Die Knoten sind auf die drei Zonen Internet, DMZ und interne Zone auf geteilt. Die Kommunikation zwischen den Zonen ist nur über ein Gateway möglich.

Abschottung des Knotens gegenüber den anderen Knoten möglich. Aus Sicht der IT-Sicherheit ist diese Abgrenzung erforderlich.

Des Weiteren werden die Metadatendienste auf einem eigenen Knoten betrieben. Auch für diese beiden Dienste, das Metadatenverzeichnis und den Vokabulardienst, erleichtert die Trennung von den übrigen Knoten die gemeinsame Nutzbarkeit und die Verwaltung der Daten.

Die in Abbildung 11 dargestellten Dienste *Dienst 1* bis *Dienst n* nehmen eine Sonderstellung im Diagramm ein, da sie lediglich exemplarisch in dieser Verteilung dargestellt sind. Tatsächlich können, entsprechende Systemressourcen vorausgesetzt, mehrere Dienste auf einem Knoten installiert werden (Dienst 1 und Dienst 2). Alternativ kann ein Dienst einem eigenen Knoten zugeordnet werden, wie für *Dienst 1* und *Dienst n* dargestellt. *Dienst n* stellt einen externen Dienst dar, der außerhalb der PELICAN-Umgebung betrieben wird.

Somit ergeben sich für die Dienste mehrere Verteilungsmöglichkeiten, die je nach den Anforderungen der beitragenden Projekte genutzt werden können: Liegen nur geringe Anforderungen bezüglich der Kontrolle der eigenen Daten durch die Projekte vor, so können die Dienste zusammen mit Diensten anderer Projekte in geteilten Knoten auf der PELICAN-Hardware installiert werden. Dazu ist es erforderlich, dass der komplette Datensatz physikalisch auf die PELICAN-Server gebracht wird, damit die Daten über die Dienstschnittstellen angeboten werden können. Bei höheren Anforderungen bezüglich der Kontrolle wird ein separater Knoten vorgesehen, der durch die Bereitstellung eines eigenen Betriebssystems einen hohen Grad an Autonomie gegenüber den übrigen Diensten bietet. Den höchsten Grad an Kontrolle über seine Daten kann ein Projekt ausüben, wenn es den Dienst auch selbst auf seiner eigenen Hardware betreibt. In diesem Fall verbleibt der vollständige Datensatz auch unter der physikalischen Kontrolle des Projektes. Lediglich Teilmengen der Daten können von anderen Nutzern des Verbundes kontrolliert über die Serviceschnittstelle abgerufen werden.

Zur Erhöhung der IT-Sicherheit wird das Verteilungsmodell in die drei Netzwerkzonen *Internet*, *Demilitarized Zone (DMZ)* und *Interne Zone* unterteilt (siehe auch Abbildung 11). Die Kommunikation ist nur zwischen den Knoten derselben Zone uneingeschränkt möglich. Die zonenübergreifende Kommunikation ist nur zwischen bestimmten Knoten und mit bestimmten Protokollen zulässig. So befindet sich der Webserver des PELICAN-Systems in der DMZ. In Richtung der internen Zone sind nur zwei Verbindungen zulässig: Jeweils zum Portalserver und zum Dokumentenserver um von dort die Inhalte zum Abruf aus dem Internet bereitzustellen. Die Komponente *Gateway* stellt das Bindeglied zwischen den Zonen dar. Sie schottet das PELICAN-System derart ab, dass aus dem Internet ausschließlich Verbindungen auf den Webserver unter Verwendung des Hypertext Transfer Protocol Secure (HTTPS)-Protokolls möglich sind. Umgekehrt ist

es allerdings für den Portalserver möglich, Verbindungen zu Diensten im Internet über das Gateway aufzubauen. Voraussetzung ist dabei, dass die Netzwerkverbindung vom Portal initiiert wird.

4.2.6 *Sicherheitsarchitektur*

Die Berücksichtigung der IT-Sicherheit ist integraler Bestandteil des SOA-Systems für PELICAN, daher werden verschiedene Sicherheitsaspekte bereits in den vorangegangenen Abschnitten dieses Kapitels betrachtet. Wesentliche Funktionen von PELICAN beruhen auf dem caCORE SDK, daher werden auch die verfügbaren Sicherheitsmechanismen durch dieses Framework bestimmt. Da für PELICAN die Implementierung von caGRID nicht erforderlich ist, können bestimmte Mechanismen, wie die der Security Assertion Markup Language (SAML)-Sicherheitstoken, nicht verwendet werden. Für die Authentisierung und Autorisierung bei den caBIG-Diensten stehen als Identifikationsmerkmale Benutzername und Passwort zur Verfügung.

Die Anmeldeinformationen der Benutzer werden in einem gemeinsamen LDAP-Verzeichnis abgelegt, sodass die Anmeldung bei allen Komponenten mit einem identischen Paar bestehend aus Benutzername und Passwort möglich ist. Weiterhin werden keine Gruppenzugänge, sondern nur individuelle Zugänge vergeben, sodass protokolliert werden kann, wer die Forschungsdaten des Verbundes abruft.

Sofern die Dienste auf der PELICAN-Hardwareumgebung betrieben werden, steht eine an etablierten Empfehlungen orientierte Sicherheitsarchitektur zur Verfügung, die ein entsprechendes Zonenkonzept umsetzt [Bundesamt für Sicherheit in der Informationstechnik 2009; VMware Inc. 2013].

4.3 VERBUNDSPEZIFISCHE METADATEN

Zur eindeutigen Annotation aller Daten in PELICAN wird ein kontrolliertes Vokabular erstellt, welches auf den SFB/TRR77 -Verbund zugeschnitten ist. Weiterhin wird eine Ontologie entwickelt, die geeignet ist, Zelllinien auch über den SFB/TRR77 hinaus umfassend zu beschreiben.

4.3.1 *Kontrolliertes Vokabular*

Zunächst werden die Ergebnisse der beiden Erhebungsstufen zur Erfassung der Terme für das Vokabular dargestellt. Des Weiteren wird der Abgleich des neu erstellten Vokabulars mit bestehenden Vokabularen des UMLS sowie die Veröffentlichung des Vokabulars beschrieben.

Vorerhebung

Durch die Analyse der Projektbeschreibungen des Forschungsverbundes mit Hilfe von qualitativer Inhaltsanalyse können 142 unterschiedliche Bezeichnungen zusammengetragen werden. Weitere Bezeichnungen (rund 30) kommen durch die Auswertung der Fragebogen an die Projektleiter (siehe Abschnitt 3.3.1) hinzu.

Weiterhin tragen die Fragebogen zu einem besseren Verständnis der von den einzelnen Projekten durchgeführten Forschung bei. Diese Erkenntnisse fließen in die Entwicklung des Interviewleitfadens für die anschließende Haupterhebung ein.

Haupterhebung

Unter Verwendung des Interviewleitfadens (siehe Anhang B) werden Interviews in Projekten des Forschungsverbundes durchgeführt. Dabei waren stets die jeweiligen Projektleiter anwesend, oft jedoch auch weitere Mitarbeiter des Projektes. Durch die Interviews werden 219 Bezeichnungen erfasst. Da manche Projekte Englisch und manche Deutsch als Arbeitssprache benutzen, wird ein zweisprachiges Vokabular erstellt. Weiterhin werden für ungefähr 10% der Begriffe Akronyme erhoben. Beispielsweise wird neben der Vorzugsbezeichnung *Tissue Microarray* auch die Abkürzung *TMA* verwendet. Zu manchen Begriffen sind auch mehrere Akronyme gebräuchlich, zum Beispiel DNA und DNS für Desoxyribonukleinsäure.

Tabelle 14 zeigt einen Auszug der Terme, die im Rahmen des Besuchs bei einem Projekt des Forschungsverbundes gesammelt wurden. Die erste Spalte der Tabelle enthält die Vorzugsbezeichnung in englischer Sprache, die zweite Spalte das zugehörige Akronym, die dritte Spalte die deutsche Benennung für den Begriff. Für einige Begriffe gibt es kein Akronym oder keine deutsche Entsprechung. Wird für einen Begriff in den Projekten ausschließlich ein Akronym gebraucht, so wird der vollständige Ausdruck recherchiert und als Vorzugsbezeichnung aufgenommen.

TERMLISTE Im nächsten Schritt werden die Tabellen der einzelnen Projekte zu einer zentralen Tabelle zusammengeführt. Nach der Bereinigung von doppelten Einträgen enthält das Vokabular Bezeichnungen zu 245 Begriffen. Diese Liste stellt eine erste Version des kontrollierten Vokabulars für den Forschungsverbund dar. Bereits in diesem Stadium wäre eine Veröffentlichung und Nutzung des Vokabulars möglich, sofern der hierarchischen Anordnung der Begriffe im Verbund keine Bedeutung beigemessen wird.

MIND-MAPS Für jedes besuchte Projekt wird eine Mind-Map erstellt. Abbildung 12 zeigt beispielsweise die aus einem Projektbesuch hervorgegangene Mind-Map. In dieser Abbildung werden die Bezeichnungen gemäß ihrer Beziehungen um den Wurzelknoten (B7) angeordnet. Da andere Pro-

Tabelle 14: Auszug aus der Tabelle mit Begriffen eines Projektes.

VORZUGSBEZEICHNUNG	AKRONYM	DEUTSCH
cell line		Zelllinie
cell migration		Zellmigration
evaluation		Auswertung
expression profile		Expressionsprofil
membrane		Membran
messenger RNA	mRNA	
microarray		
microRNA	miRNA	
mouse model		Mausmodell
proliferation		Proliferation
protein		Protein
pro-tumorigenic		protumorigen
quantification		Quantifizierung
staining		Färbung
structure		Struktur
Tissue microarray	TMA	
western blot		Western Blot

jekte eine von diesem Projekt abweichende Zielsetzung haben, können die Begriffe aus verschiedenen Projekten unterschiedlich miteinander verknüpft sein.

VOCALIGN

Die VOCALIGN-Software wurde von mir entworfen und im Rahmen eines von mir betreuten Praktikums programmiert. Zunächst wird durch das Programm in UTS abgefragt, ob ein zur Bezeichnung passendes *concept* im UMLS vorhanden ist. Dabei erlaubt die Schnittstelle auch die unscharfe Suche, sodass Schreibfehler und Variationen in gewissem Umfang kompensiert werden können. Unter ungünstigen Umständen kann diese Funktion aber auch ein Ergebnis liefern, das in keiner Beziehung zum Suchwort steht. Aus diesem Grund ist bei Nutzung der aktuellen Version von UTS eine manuelle Überwachung der Ergebnisse angezeigt. Wird ein Konzept von UTS zurückgeliefert, so wird der entsprechende präferierte Bezeichner sowie die hierarchische Anordnung des Konzeptes extrahiert.

Gemäß Definition müssen Begriffe in einem kontrollierten Vokabular nicht notwendigerweise in einer Hierarchie geordnet sein. Allerdings ist

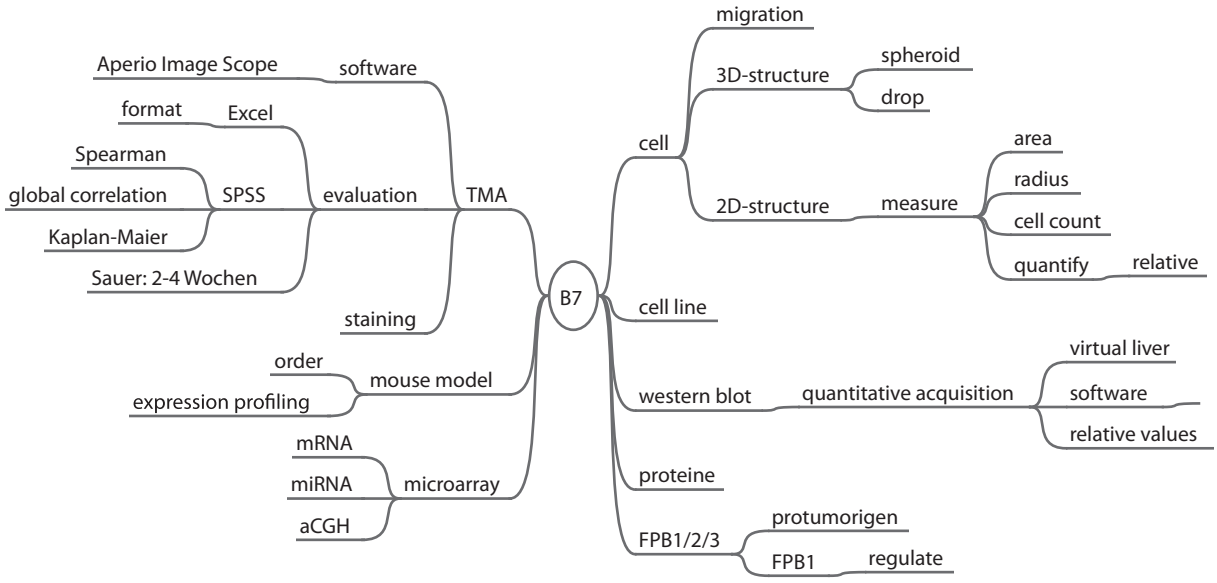


Abbildung 12: Beispiel für eine Mind-Map mit den Bezeichnungen des SFB/TRR77-Projektes B7.

es für viele Anwendungen hilfreich, eine Klassifikation durchzuführen und die Begriffe des Vokabulars in einer Taxonomie anzuordnen (siehe Abschnitt 2.6.1). VOCALIGN unterstützt den Kurator des Vokabulars bei der Erstellung einer Taxonomie indem es versucht, die übergeordneten Konzepte eines Suchwortes bis zum Wurzelbegriff aus dem UMLS zu extrahieren. Sollten mehrere übergeordnete Pfade existieren, so wird nur ein Pfad übernommen. In diesem Fall setzt sich der Pfad aus dem ersten von UTS zurückgegebenen Element der jeweiligen Hierarchieebene zusammen. Alternative Pfade werden nicht übernommen, um Zyklen bei der Abfrage zu vermeiden.

Auf der Grundlage dieser Vorschlagsliste kann der Kurator entscheiden, welche Teile des Referenzvokabulars bezüglich eines gesuchten Begriffes in das lokale Vokabular übernommen werden sollen. Die Entscheidung wird durch die Mind-Maps unterstützt, da in diesen die Information über projektrelevante Beziehungen zwischen Begriffen dokumentiert sind.

In einem weiteren Aufruf von UTS werden zum Konzept gehörige sogenannte Atome (engl. *atoms*) abgefragt und dem Konzept zugeordnet. Diese Atome sind nicht Bestandteil des UMLS-Metathesaurus, sondern verschiedener anderer Quellen des UMLS wie MeSH. Weiterhin enthalten diese Quellen auch Definitionen zu den Suchworten, die im abschließenden dritten Schritt extrahiert werden. Diese Schritte werden wiederholt für alle Terme auf der Kandidatenliste durchgeführt. Die Ergebnisse werden in Textdateien ausgegeben, je nach Grad der Vollständigkeit des Suchergebnisses.

Tabelle 15: Ausgabedateien des VOCALIGN-Programms.

DATEINAME	INHALT
hits	Liste der Vorzugsbezeichnungen aus UMLS zusammen mit Definitionen aus NCIt, MeSH oder beidem
noHit	Alle Bezeichnungen, die Bestandteil von noConcept, noAtoms oder noDefinitions sind. Dies ist eine Zusammenfassung aller Begriffe mit unvollständigem Datensatz.
noConcept	Bezeichnungen ohne passendes UMLS-Konzept
noAtoms	Bezeichnungen mit passendem UMLS-Konzept, aber ohne Atom aus NCIt oder MeSH
noDefinitions	Bezeichnungen mit passendem UMLS-Konzept und Atom aus NCIt oder MeSH, aber ohne Definition
relations	Bezeichnungen mit hierarchischer Information, sofern diese vorhanden ist.
remainingTerms	Nicht prozessierte Bezeichnungen der Eingabeliste. Bei Programmabbruch kann die Bearbeitung der verbleibenden Bezeichnungen fortgesetzt werden

Tabelle 15 gibt einen Überblick über die unterschiedlichen Ergebnislisten. Die Bezeichnungen und Definitionen in der Datei *hit-list.txt* sind dabei so formatiert, dass sie direkt in den Vokabularserver TemaTres (siehe Abschnitt 2.6.4) eingelesen werden können. Listing 1 zeigt einen Ausschnitt der Datei als Beispiel.

UMLS-Abgleich

Mit Hilfe des VOCALIGN-Programms werden alle Bezeichnungen aus der konsolidierten Liste mit den Bezeichnungen des UMLS verglichen. Zu den meisten Bezeichnungen wird über die UTS-Schnittstelle eine Entsprechung zurückgeliefert. Wie in Tabelle 16 ersichtlich ist, werden insgesamt 143 Definitionen aus NCIt oder MeSH übernommen. Für 71 Begriffe werden Definitionen in beiden Vokabularen gefunden, die zusammen mit den jeweiligen Identifikationsnummern gespeichert werden (siehe auch Seite 74). In der Tabelle werden jedoch nur die Begriffe berücksichtigt, die auch in UMLS gefunden werden. Für weitere 90 Begriffe, die nicht gefunden werden, kann auch keine Definition aus UMLS übernommen werden. Diese Begriffe werden, ebenso wie elf Begriffe, die zwar als solche in UMLS gefunden werden, jedoch nicht mit einer Definition versehen sind, manuell behandelt. Zu diesen Begriffen wird, unter Berücksichtigung des Projekt-

Listing 1: Auszug aus der Liste von VOCALIGN gefundener Bezeichnungen (Beispiel). In diesem Fall ist die erste Bezeichnung *Cultured Cell Line*, der so in UMLS gefunden wurde. Da in den Projekten jedoch die Bezeichnung *cell line* erfasst wurde, wird *cell line* als Synonym im Feld UF (use for) geführt. In den DEF-Feldern sind die aus UMLS extrahierten Definitionen aufgeführt: Die erste Definition stammt aus dem NCIt und besitzt die Identifikationsnummer C16403 aus diesem Thesaurus. Die zweite Definition stammt aus MeSH und trägt die MeSH-Nummer M0003757. Die zweite Bezeichnung in der Datei, *Purity*, wird genauso als Suchwort an VOCALIGN übergeben und besitzt daher zunächst keinen UF-Eintrag. Eine Definition wird lediglich im NCIt gefunden, nicht jedoch in MeSH.

```
Cultured Cell Line
UF: cell line
DEF: NCI - ID C16403 - A permanently established
   cell culture that will proliferate indefinitely
    given appropriate fresh medium and space. (On-
   line Medical Dictionary)
DEF: MSH - ID M0003757 - Established cell cultures
    that have the potential to propagate
   indefinitely.


Purity
DEF: NCI - ID C62352 - A quantitative assessment
   of the homogeneity or uniformity of a mixture.
   Alternatively, purity refers to the degree of
   being free of contaminants or heterogeneous
   components.
```

Tabelle 16: Anzahl der Begriffe, die im UMLS gefunden werden und keine, eine oder zwei Definitionen aus MeSH oder NCIt besitzen.

	MeSH	KEIN MeSH	SUMME
NCIt	71	58	129
KEIN NCIt	14	11	25
SUMME	85	69	154

kontextes, eine Definition in der einschlägigen Literatur recherchiert und ins Vokabular übernommen.

Für den Fall, dass die Vorzugsbezeichnung aus dem UMLS von der lokal erfassten Bezeichnung abweicht, wird die UMLS-Bezeichnung beibehalten und die lokale Bezeichnung als Synonym eingetragen. Somit wird die Interoperabilität mit anderen UMLS-annotierten Systemen und Datensätzen verbessert.

Hierarchie des Vokabulars

Die aus dem UMLS entnommene Hierarchieinformation wird, sofern sie für den entsprechenden Begriff vorliegt, in der Datei *relations.txt* gespeichert. Die Analyse dieser hierarchischen Beziehung ergab, dass zu den meisten Termen mehrere übergeordnete Begriffe im UMLS vorhanden sind. Um das lokale Vokabular so kompakt wie möglich zu halten, werden keine übergeordneten Begriffe des UMLS übernommen, es sei denn, der übergeordnete Begriff ist seinerseits Teil des projektspezifischen Vokabulars. Bei der hierarchischen Anordnung der Begriffe kommen die in den Mind-Maps dokumentierten Beziehungen zum Tragen. Widersprechen sich die Mind-Maps verschiedener Projekte, so wird eine Variante zur Einordnung ausgewählt. Begriffe, für die keine geeigneten Hierarchieinformationen erfasst wurden, werden als unverbundene Begriffe im Vokabular geführt.

Veröffentlichung des Vokabulars

Das für den SFB/TRR77 erstellte kontrollierte Vokabular ist auf dem Vokabularserver des Verbundes veröffentlicht. Es ist auch aus dem Internet mit einem Webbrowser erreichbar. Da der Server auf TemaTres basiert, kann das Vokabular in verschiedenen Formaten, einschließlich SKOS, heruntergeladen werden. Weiterhin ist das Vokabular über die Webservices-Schnittstelle nach dem REST-Protokoll abrufbar. Die Adresse für den Zugriff mit einem Webbrowser lautet *https://livercancer.imbi.uni-heidelberg.de/vocabulary*.

4.3.2 *Ontologie für Zelllinien*

In den folgenden Abschnitten werden die Ergebnisse der durchgeführten Analysen zur Zelllinienontologie sowie die Cell Culture Ontology (CCONT) selbst vorgestellt.

Definition des Begriffsbereichs

Die Analyse der in Tabelle 8 aufgeführten Zellbank-Kataloge zeigt, dass diese verschiedene Ansätze zur Beschreibung der Zelllinien verfolgen. Viele Informationsfelder, wie der Ursprung der Zellen, sind zwar in allen Datenbanken vorhanden, jedoch nicht immer in einer strukturierten Form. So führt die Zellbank ATCC beispielsweise die Felder *Age*, *Gender* und *Ethnicity*, die Zellbank European Collection of Cell Cultures (ECACC) besitzt jedoch nur ein Feld mit dem Titel *cell line description*, das die gleichen Informationen in einem einzigen Freitextfeld zusammenfasst.

Weiterhin ergeben sich Ähnlichkeiten zwischen den Zellbankkatalogen bezüglich der Identifikation der Zelllinien: Alle Kataloge verfügen über einen Namen für die Zelllinien sowie eine interne Identifikationsnummer. Der Zelllinienname kann über verschiedene Zellbanken hinweg einheitlich

Tabelle 17: Gegenüberstellung der Datenfelder *medium* und *organ* bezüglich ihrer Behandlung in verschiedenen Zellbanken. Die mit * markierten Felder sind Freitextfelder, die auch für andere Einträge genutzt werden können.

ZELLBANK	MEDIUM	ORGAN
ATCC	propagation*	source/organ
DSMZ	medium*	—
ECACC	culture medium*	tissue
ICLC	culture conditions*	description*
Riken	medium	tissue

sein, jedoch sind Inkonsistenzen möglich [Sarntivijai et al. 2008]. Da die Identifikationsnummern zellbankspezifisch sind, eignen sie sich nicht, um Informationen katalogübergreifend zusammenzuführen.

In Tabelle 17 werden die unterschiedlichen Ansätze der Zellbanken anhand der Elemente *medium* und *organ* exemplarisch aufgezeigt. Für *medium* stellt nur eine Zellbank ein spezifisches Datenfeld bereit, während die übrigen Kataloge Mehrzweckfelder mit Freitext verwenden. Die meisten Kataloge führen hingegen ein spezifisches Feld für die Beschreibung des Ursprungsorgans der Zelllinie. Lediglich die DSMZ stellt kein eigenes Datenfeld für diesen Zweck bereit. Die vollständige Analyse aller Datenfelder wurde als ergänzende Information S1 zu einem Zeitschriftenartikel veröffentlicht [Ganzinger, He et al. 2012]. Insgesamt werden rund 30 Datenfelder identifiziert, die zur Beschreibung der Zelllinien verwendet werden. Die Felder lassen sich den folgenden fünf Gruppen zuordnen:

1. Identifikation: Der Name der Zelllinie
2. Ursprung: Beschreiben der Herkunft des Gewebes, aus dem die Zelllinie entwickelt wurde
3. Eigenschaften: Felder, die charakteristische Eigenschaften der Zelllinie beinhalten
4. Kultivierung: Daten, welche die Wachstumsbedingungen beschreiben, werden in dieser Gruppe zusammengefasst
5. Tests: Daten für Kontaminationstests von Zelllinien

Die Datenfelder, die auf diese Gruppen abgebildet werden, sind in den ersten beiden Spalten von Tabelle 18 dargestellt. Diese Tabelle fasst auch die essentiellen Datenfelder zur Beschreibung von Zelllinien zusammen.

Tabelle 18: Die Tabelle zeigt die Datenfelder, die aus den Katalogen der Zellbanken abgeleitet werden. Die letzten drei Spalten zeigen, welche Ontologien mit Bezug zu Zelllinien Klassen besitzen, welche die entsprechenden Datenfelder abdecken. Dabei bedeutet vollständige Abdeckung, ○ teilweise Abdeckung und × keine Abdeckung.

DATENGRUPPE	FELDNAME	MCCL	CLO	EFO
Identifikation	cell line name			
Ursprung	ethnicity	×	×	
	age	×	×	
	sex	×	×	
	species		×	○
	strain	×	×	○
	organ		×	
	disease		×	
Eigenschaften	growth mode			×
	morphology			
	cellular products	×	×	
	cytogenetics	×	×	×
	biosafety level	×	×	×
	risk group	×	×	×
Kultivierung	medium	×	×	×
	supplements	○	×	○
	temperature		×	
	atmosphere	×	×	
	confluence rate	×	×	×
	seed density	×	×	×
	split ratio	×	×	×
	detachment aid	×	×	×
Tests	STR fingerprint	×	×	×
	viruses	×	×	×
	mycoplasma	×	×	×

Analyse bestehender Ontologien

Im Zuge der Untersuchung existierender Ontologien, die zur Beschreibung von Zelllinien herangezogen werden können, wird keine Ontologie identifiziert, welche den vollständigen Umfang der im vorigen Abschnitt genannten Beschreibungsmerkmale abdeckt. Aus diesem Grund werden in diesem Abschnitt Ontologien betrachtet, die für die Beschreibung einer oder mehrerer Gruppen aus Tabelle 18 geeignet sind.

IDENTIFIKATION VON ZELLLINIEN Die Suche in BioPortal nach der Bezeichnung *cell line* ergibt 17 Ontologien (Stand 1.3.2012, 11:54). In sechs von diesen Ontologien besitzt die jeweilige *cell line*-Klasse keine Unterklassen. Dies bedeutet, dass diese Ontologien keine Klassen für bekannte Zelllinien besitzen, sondern lediglich einen Platzhalter, an dem diese Informationen eingefügt werden können. Die Gesamtstruktur der Ontologie könnte dennoch für die umfassende Erfassung von Zellliniendaten nützlich sein. Da im Rahmen dieser Arbeit jedoch der Ansatz verfolgt wird, einen möglichst großen Anteil bestehender Ontologien wiederzuverwenden, werden im Folgenden nur Ontologien mit einem umfassenden Bestand an Klassen für etablierte Zelllinien berücksichtigt. Als solche Ontologien werden diejenigen angenommen, in denen die *cell line*-Klasse mehr als 100 Unterklassen besitzt.

Es existieren zwei Ontologien mit dem Namen *cell line ontology*. Die erste namens Molecular Connections Cell Line Ontology (MCCL), ist über BioPortal mit der Identifikationsnummer 1245 verfügbar [Molecular Connections 2011]. Diese Ontologie beinhaltet Zelllinniennamen, die aus Zellbankkatalogen wie zum Beispiel von ATCC exportiert wurden. Weiterhin enthält sie anatomische Informationen aus der Ontologie *Foundational Model of Anatomy (FMA)* [Rosse, Mejino 2003] sowie Verweise auf Krankheiten aus der Human Disease Ontology (DO) [Osborne et al. 2009].

Die zweite *cell line ontology* wird mit CLO abgekürzt. Entwickler der CLO konzentrieren sich in ihrer Arbeit auf die automatisierte Extraktion und Harmonisierung von Zelllinniennamen aus Zellbanken und anderen Datenbanken [Sarntivijai et al. 2008]. Die zum Zeitpunkt dieser Arbeit auf BioPortal veröffentlichte Version 2.0.27 verfügt über 8728 Namen von Zelllinien, die alle als Unterklassen der Klasse *permanent cell line* modelliert sind. Im Gegensatz zu MCCL verfügt CLO nicht über Referenzen zu weiteren Zelllinieneigenschaften wie der Anatomie des Ursprungsgewebes.

Weiterhin enthält die Experimental Factors Ontology (EFO) rund 700 Namen von Zelllinien [Malone et al. 2010]. Die entsprechenden Zelllinien-Klassen sind mit Verweisen auf die Anatomie sowie auf Krankheiten versehen. Da diese Ontologie mit dem Ziel entwickelt wurde, biomedizinische Experimente zu beschreiben, bringt sie viele Klassen mit, die zu einer umfassenden Beschreibung von Zelllinien mit Metadaten erforderlich sind.

URSPRUNG DER ZELLEN Diese Metadatengruppe beschreibt den Ursprung der Gewebeprobe, aus der die Zelllinie entwickelt wurde. In dieser Gruppe sind Informationen über die Art des Ursprungs (zum Beispiel *Homo sapiens* oder *Mus musculus*), aber auch detailliertere Informationen wie die Ethnizität des Spenders oder der Stamm eines Tiermodells zusammengefasst. Weiterhin wird das Ursprungsorgan der Zellen sowie eine mit diesen assoziierte Erkrankung betrachtet.

Von den im vorigen Abschnitt beschriebenen Ontologien besitzt lediglich EFO Klassen für die Beschreibung der Ethnizität des Zellspenders. Alternativ zur Verwendung von EFO kann die Systematized Nomenclature of Medicine (SNOMED) Ethnic Groups (*http://bioportal.bioontology.org/ontologies/SNOMED-Ethnic-Group*, Stand: 02.10.2013, 13:01) verwendet werden. Diese Ontologie enthält 261 Klassen zur Beschreibung der Ethnizität.

Alle beschriebenen Zelllinienontologien stellen Klassen zur Beschreibung des Geschlechts bereit, jedoch enthält nur EFO eine Klasse für das Alter. Weiterhin enthält EFO Klassen mit zelllinienrelevanten Arten. Die Unterstützung der Arten ist in MCCL größer als in EFO, da dort auch Klassen für Spezies enthalten sind, die im Labor nur von geringer Bedeutung sind. Zuchtstämme einiger Modellorganismen, vor allem von *Mus musculus*, sind wiederum in EFO enthalten. CLO und MCCL enthalten zwar Klassen von Zelllinien, die von solchen Modellorganismen abgeleitet wurden, jedoch keine Klassen zu den Arten selbst.

Alle im vorigen Abschnitt identifizierten Zelllinien enthalten Klassen zur Beschreibung der Ursprungsorgane, indem sie spezifische Anatomie-Ontologien wie FMA einbinden. Sämtliche Ontologien stellen auch krankheitsbezogene Informationen durch Einbindung der DO bereit.

EIGENSCHAFTEN VON ZELLLINIEN Diese Gruppe der Datenelemente enthält grundlegende Informationen zur Identifikation der Zelllinien sowie zum Umgang mit ihnen:

- *Modus* beinhaltet Informationen darüber, ob die Zellen adhärent oder in Suspension wachsen
- *Morphologie* beschreibt die Gewebeart, von der die Zellen abstammen, zum Beispiel Epithelzellen oder neuronale Zellen
- *Zellprodukte* fasst die Substanzen zusammen, welche die Zellen produzieren
- *Zytogenetik* enthält Informationen über die chromosomalen Eigenschaften der Zelllinie
- *Short Tandem Repeat (STR)-Profil* enthält Daten um die Authentizität humaner Zelllinien zu bestätigen

- *Sicherheitsstufe (Biosafety level) und Risikokategorie* werden benutzt um gefährliche Stoffe zu kennzeichnen
- *Viren und Mykoplasmen* beschreibt, auf welche Kontaminationen hin die Zelllinien untersucht wurden

Während Klassen für die Morphologie in allen drei Zelllinienontologien vorhanden sind, enthält EFO keine Möglichkeit, den Modus zu beschreiben. Zellprodukte sind im Wesentlichen chemische Verbindungen und können als solche unterhalb der Klasse *chemical compounds* in EFO dargestellt werden, nicht jedoch in den beiden übrigen Ontologien.

Der Begriff *cytogenetics* kommt in keiner der drei Zelllinienontologien vor, allerdings in 13 anderen auf BioPortal verzeichneten Ontologien (Stand: 2.3.2012, 13:16). Sämtliche dieser Ontologien wurden für klinische Anwendungen entwickelt. Daher sind ihre zytogenetischen Klassen auf die Beschreibung von Krankheiten, die durch chromosomale Aberrationen entstehen, ausgerichtet. Um eine allgemeinere Beschreibung zytogenetischer Befunde zu ermöglichen, wäre es sinnvoll, das International System for Human Cytogenetic Nomenclature (ISCN) heranzuziehen [Shaffer et al. 2009]. Die Entwickler der Clinical BioInformatics Ontology (CBO) geben an, dass mit Hilfe dieser Ontologie chromosomale Variationen in einem semantischen Netzwerk beschrieben werden können [Hoffman et al. 2005].

Die Ergebnisse zur Suche nach *STR profile* sind ähnlich: Der Begriff kommt zwar in drei Ontologien von BioPortal vor, jedoch besitzt keine von ihnen Unterklassen zur STR-Klasse, um ein konkretes Profil zu beschreiben. Damit ein STR-Profil abgelegt werden kann, müssen Klassen für die gegenwärtig beim Menschen verwendeten Marker vorgesehen werden. Bis ein verabschiedeter Standard [Barallon et al. 2010] verfügbar ist, sollten zumindest die von den gängigen forensischen Testkits verwendeten Marker in der Ontologie berücksichtigt werden.

Zwei Ontologien enthalten Begriffe zu den Sicherheitsstufen 1-4 (*biosafety levels 1-4*, Stand: 2.3.2012, 14:48): Die BioAssay Ontology (BAO) enthält diese Klassen als Unterklassen zum *assay biosafety level* [Visser et al. 2011]. Da im Rahmen des Forschungsverbundes jedoch die Eigenschaften der Zelllinien und nicht die mit ihnen durchgeführten Versuche betrachtet werden, liegt in BAO offenbar eine andere Bedeutung zu Grunde, sodass diese Ontologie nicht herangezogen wird. Die zweite Ontologie, die Klassen zur Sicherheit enthält, ist SNOMED Clinical Terms (*http://bioportal.bioontology.org/ontologies/SNOMEDCT* Stand: 2.10.2013, 15:30). Dort sind die vier Klassen unabhängig von einer bestimmten Verwendung geführt. Sie sind unterhalb des Begriffs *Levels*, der wiederum Unterklasse von *Ranked categories* ist, einsortiert.

KULTIVIERUNG In diese Gruppe fallen Bedingungen und Prozesse, die Voraussetzung für eine erfolgreiche Kultivierung der Zelllinien sind: Es ist

von großer Bedeutung, welche Wachstumsmedien, Zusätze und sonstigen Umgebungsbedingungen die optimalen Wachstumsbedingungen für die jeweilige Zelllinie darstellen.

Alle Zellkulturen benötigen ein abgestimmtes Wachstumsmedium, welches den Zellen individuell optimierte Wachstumsbedingungen bereitstellt. Diese Medien werden häufig nach standardisierten Rezepten hergestellt, oder können fertig von Biochemie-Firmen gekauft werden. Keine der Zelllinienontologien stellt Klassen für Standardmedien bereit. Diese Lücke kann mit der INOH Event Ontology (IEV) des Integrating Network Objects with Hierarchies (INOH) geschlossen werden, da IEV diese Zellkulturmedien umfassend berücksichtigt [Kushida et al. 2006].

Zusätze sind Substanzen wie Salze oder Seren die für das Zellwachstum erforderlich sind. Diese chemischen Verbindungen müssen der Zellkultur in der Regel zusätzlich zum Medium beigefügt werden. In CLO sind solche Substanzen nicht modelliert, wohingegen MCCL einen kleinen Teilaspekt der Zusätze abdeckt. EFO stellt den Unterbaum *chemical compound* bereit, der einen großen Teil der üblichen Zusätze enthält.

Erwartungsgemäß spielt der Begriff *temperature* laut BioPortal in 104 Ontologien eine Rolle. Zu diesen Ontologien gehören EFO, MCCL und die Ontology for Biomedical Investigations (OBI) [Brinkman et al. 2010]. *Atmosphere* kommt in 15 Ontologien vor, wobei EFO die einzige unter diesen ist, die einen Bezug zu Zelllinien aufweist.

Die restlichen Begriffe dieser Gruppe beziehen sich auf den Prozess der Subkultivierung. Die Begriffe *confluence rate*, *seed density*, *split ratio* und *detachment aid* werden über BioPortal in keiner Ontologie gefunden.

Erstellen des konzeptuellen Modells

In diesem Abschnitt wird das konzeptuelle Modell, welches der im Rahmen dieser Arbeit erstellten Ontologie CCONT zu Grunde gelegt wird, vorgestellt. Dabei wird versucht, möglichst viele der in Tabelle 18 aufgeführten Kategorien durch Wiederverwendung von bestehenden Ontologien oder Teilen davon abzudecken. Dennoch ist die Definition von 98 neuen Klassen innerhalb von CCONT erforderlich. Zusammen mit den aus anderen Ontologien importierten Klassen ergeben sich insgesamt 245 Klassen, die in der OWL-Datei von CCONT enthalten sind. Abbildung 13 zeigt die Struktur von CCONT in komprimierter Form.

Zur Entwicklung der Ontologie werden die OWL-Funktionen des Ontologie-Editors Protégé [Noy, Crubezy et al. 2003] verwendet. Um die Interoperabilität mit anderen Ontologien zu ermöglichen, wird CCONT als Präfix allen in dieser Ontologie definierten Klassen vorangestellt. Gemäß einer BioPortal-Abfrage am 1. März 2012 wird dieses Präfix für keine andere Ontologie verwendet. Jede Klasse besitzt eine eindeutige Identifikation, wie zum Beispiel *http://livercancer.imbi.uni-heidelberg.de/CCONT_0000001*.

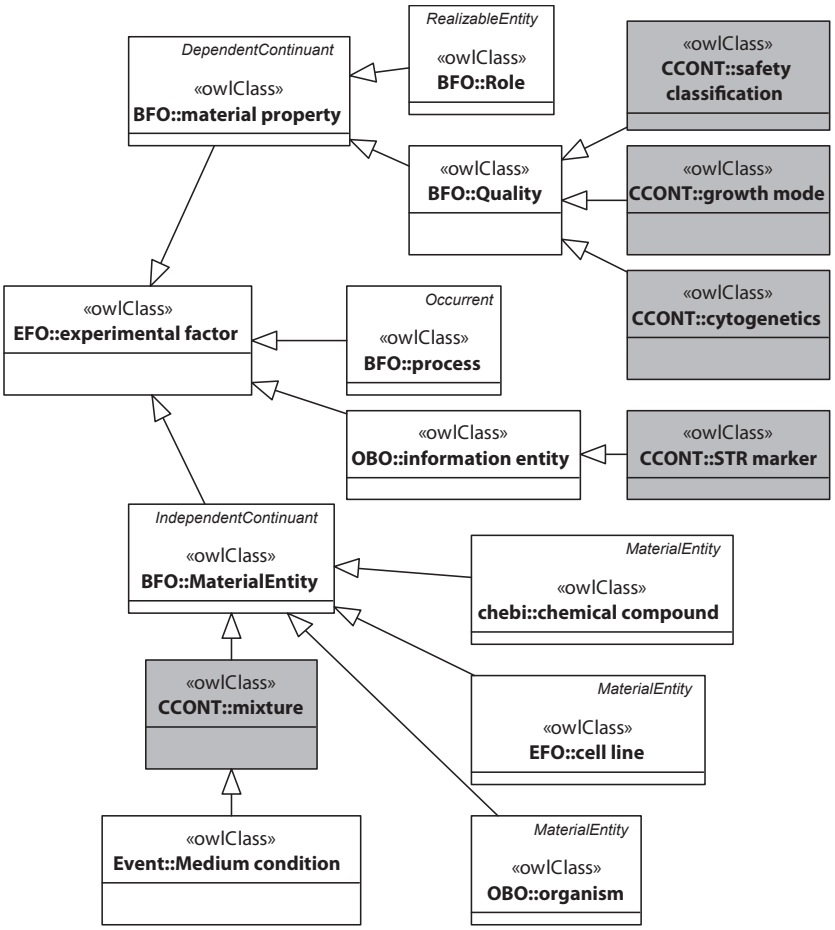


Abbildung 13: Die Abbildung zeigt die Klassen der obersten Ebene von CCONT in UML-Notation. Die grau hinterlegten Klassen werden in CCONT neu definiert. Die Klasse *Medium condition* wird aus der IEV-Ontologie importiert. Alle übrigen Klassen stammen aus EFO. Da EFO ihrerseits Ontologien wie OBO oder BFO referenziert, werden diese als UML-Paketnamen dargestellt. Zur Vereinfachung der Darstellung werden nur *subClassOf* -Relationen dargestellt.

Listing 2: Import von IEV. Mit den hier dargestellten RDF-Befehlen wird die Klasse *medium conditions* aus der IEV-Ontologie mit ihren Unterklassen nach CCONT importiert.

```
<rdf:Description rdf:about="http://purl.org/
    obo/owl/IEV#IEV_0000293">
     <rdfs:subClassOf
      rdf:resource="http://livercancer.imbi.uni
         -heidelberg.de/ccont#CCONT_0000047"/>
     <rdfs:subClassOf>
         <owl:Restriction>
             <owl:onProperty rdf:resource="
                 http://purl.obolibrary.org/obo/
                 OBI_0000316"/>
             <owl:someValuesFrom rdf:resource="
                 http://www.ebi.ac.uk/efo/
                 EFO_0000579"/>
         </owl:Restriction>
     </rdfs:subClassOf>
</rdf:Description>
```

Als Ausgangspunkt für die neue Ontologie wird eine Ontologie aus der vorangegangen Recherche gewählt, die möglichst viele Anforderungen an die Repräsentation von Zellliniendaten erfüllt. In den letzten drei Spalten von Tabelle 18 werden die Beiträge der in Abschnitt 4.3.2 beschriebenen Zelllinienontologien zur vollständigen Abdeckung der Zelllinieneigenschaften dargestellt. Da EFO die meisten Datengruppen abdeckt, wird diese Ontologie als Basis für CCONT gewählt. Technisch geschieht dies durch den Import des URI *http://www.ebi.ac.uk/efo/*. Alle Klassen aus EFO tragen das Präfix EFO, zum Beispiel EFO_0000001. CCONT wurde mit EFO Version 2.14, abgerufen am 27.7.2011, 13:04 entwickelt.

Weiterhin wird der Unterbaum *medium conditions* (Klasse IEV_0000293) aus der IEV-Ontologie nach CCONT importiert, um Zellkulturmedien zu beschreiben. *Medium conditions* wird unterhalb der neu erstellten Klasse *mixture* eingereiht. *Mixture* selbst ist wiederum Unterklasse der EFO-Klasse *material entity*.

Listing 3 in Anhang E illustriert die technische Umsetzung von CCONT und die Integration von EFO und IEV. EFO stellt mit *experimental factor* (EFO_0000001) das Wurzelelement der Ontologie bereit. Da aus IEV lediglich der Teil *medium conditions* in CCONT zur Anwendung kommt, wird diese Klasse durch das in Listing 2 dargestellte RDF-Fragment der Klasse *mixture* (CCONT_0000047) untergeordnet.

Für die in Tabelle 18 dargestellten Felder, die weder von EFO noch von IEV abgedeckt werden, wird keine Ontologie identifiziert, die in geeigneter Wei-

se durch Referenzierung in CCONT eingebunden werden kann. In diesen Fällen werden die Klassen direkt in CCONT definiert. Was die inhaltliche Definition der Klassen anbelangt, so wird versucht, die Begriffsdefinitionen aus MeSH oder der Chemical Entities of Biological Interest (ChEBI)-Datenbank heranzuziehen [Matos et al. 2009].

Das erste Element, welches EFO oder IEV nicht abdecken, ist *growth mode*. Dazu wird eine entsprechende Klasse unterhalb der EFO-Klasse *Quality* (EFO_0001436, siehe auch Abbildung 14) im *material property* Unterbaum definiert. Die Klasse *Quality* stammt ursprünglich aus der BFO und wurde von den EFO-Entwicklern importiert. *Growth mode* besitzt die beiden Unterklassen *adherent* and *suspended*.

Die Klasse *cellular product* wird parallel zur Klasse *role* (EFO_0001440) aus der EFO eingeordnet. Diese Klasse wird dazu verwendet, Substanzen, die Unterklassen von *chemical compounds* sind, als Zellprodukte zu markieren, indem sie zusätzlich als Unterklassen von *cellular product* geführt werden.

Durch die Vielzahl möglicher chromosomaler Aberrationen ist die Darstellung der Zytogenetik in einer Ontologie eine besondere Herausforderung. Die einzige Ontologie, die zytogenetische Informationen abdeckt (CBO), kann auf Grund von Einschränkungen in ihren Lizenzbedingungen nicht in CCONT eingebunden werden. Da nicht alle Zellbanken (einschließlich ATCC) Informationen zur Zytogenetik in einer ISCN-konformen Weise bereitstellen, wird in der ersten Version von CCONT lediglich eine Klasse mit der Bezeichnung *cytogenetics* als Platzhalter definiert. Um die automatisierte Inferenz zytogenetischer Informationen zu ermöglichen, wird angestrebt, in einer zukünftigen Version von CCONT eine ISCN-konforme Klassenstruktur bereitzustellen. Vorläufig können zytogenetische Eigenschaften in Textform im Datenfeld *cytogeneticsProperty* hinterlegt werden.

Zur Beschreibung der STR-Signatur wird die Klasse *STR marker* unterhalb der Klasse *information entity* (EFO_0001435, siehe auch Abbildung 13) definiert. Die Loci *Amelogenin, CSF1PO, D13S317, D16S539, D18S51, D19S433, D21S11, D2S1338, D3S1358, D5S818, D7S820, D8S1179, F13A01, F13B, FESFPS, FGA, THO1, TPOX* und *vWA* werden als Unterklassen zur Klasse *STR marker* eingefügt. Diese Marker werden auch von Cell Line Integrated Molecular Authentication (CLIMA) [Romano et al. 2009] und ATCC verwendet.

Sicherheitsstufe (*engl. biosafety level*) und Risikogruppe (*engl. risk group*) sind zwei unterschiedliche Konzepte um den Grad der Gefährdung zu beschreiben, die von einer Zelllinie ausgeht. In CCONT wird daher die Klasse *safety classification* unterhalb der Klasse *quality* (EFO_0001436) im *material property*-Unterbaum definiert (siehe auch Abbildung 14). *Safety classification* enthält die beiden Unterklassen *biosafety level* und *risk group*. Jede dieser Klassen besitzt wiederum vier Unterklassen: *biosafety level 1* bis *biosafety level 4* stammen aus SNOMED CT und sind mit der entsprechen-

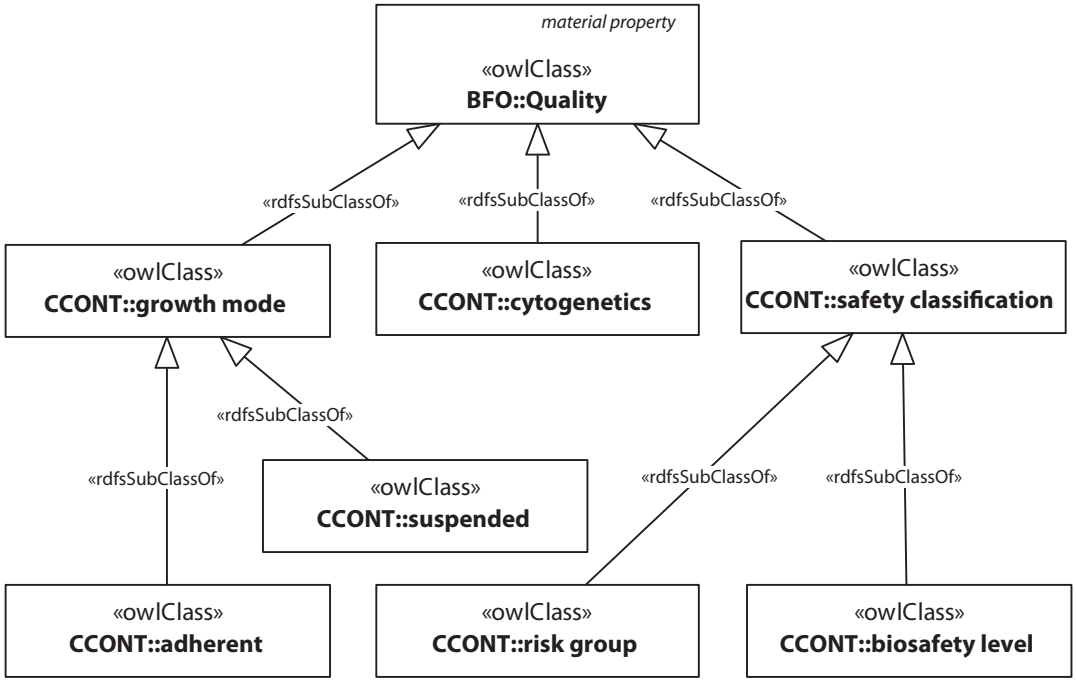


Abbildung 14: Der *Quality*-Unterbaum von CCONT. Nicht enthalten in dieser Darstellung sind die Klassen *risk group 1-4* unterhalb der Klasse *risk group* sowie die Klassen *biosafety level 1-4* unterhalb der Klasse *biosafety level*.

den SNOMED-Kennung versehen. Weiterhin werden *risk group 1* bis *risk group 4* gemäß den Richtlinien der NIH definiert [Department of Health and Human Services, National Institutes of Health 2011].

Zur Abbildung von Testergebnissen zum Nachweis von Viren in Zellkulturen werden sechs weit verbreitete Virustypen wie HBV, HCV und HIV als Unterklassen der in EFO präsenten Klasse *Virus* (NCBITaxon_10239) hinzugefügt. Informationen zu Mykoplasma werden zunächst als Freitextfeld beibehalten, da für eine formale Darstellung zunächst eine Taxonomie für Pilzkontaminationen definiert werden muss.

Zusätzlich zu den Kulturmedien, die aus IEV stammen, werden Klassen für Seren als Unterklassen zu *mixture* definiert. Die Klasse *growth mode* wird unterhalb der Klasse *quality* eingeordnet. Die meisten Klassen, die hier definiert werden, dienen der Beschreibung von Zusätzen, die für die Zellkultur benötigt werden. Als Behälterklasse wird die Klasse *supplement* unterhalb von *growth condition* (EFO_0000523) im *role*-Unterbaum von EFO definiert. Die Struktur der Klasse *supplement* und ihrer Unterklassen ist in Abbildung 15 dargestellt.

Es ist zu beachten, dass einige Klassen, die in EFO definiert werden, zusätzlich als Unterklassen zu *supplement* eingebunden werden, um deren

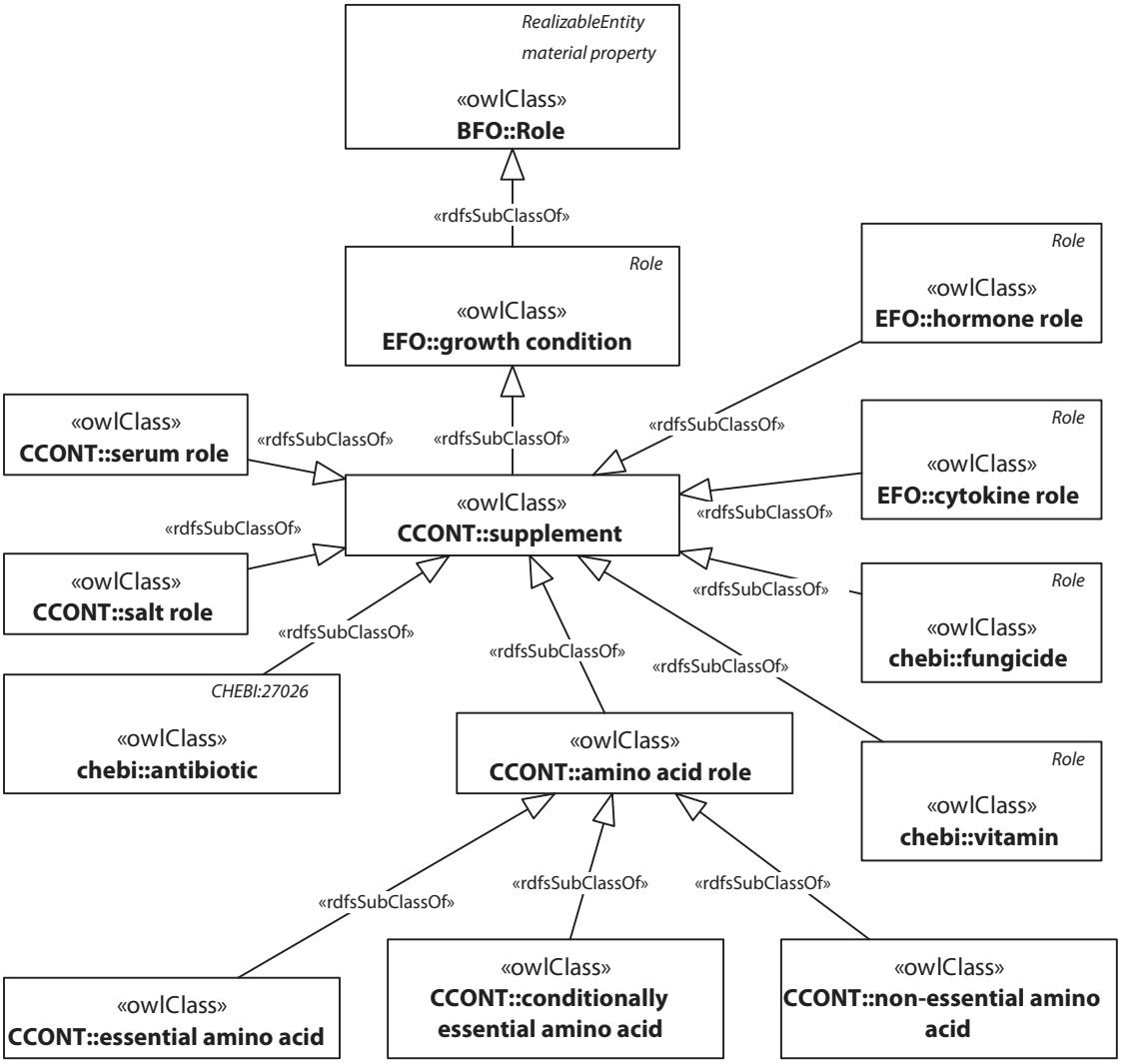


Abbildung 15: Der *Supplement*-Unterbaum von CCONT.

Tabelle 19: Chemische Verbindungen spielen verschiedene Rollen beim Wachstum von Zelllinien. In der zweiten Spalte wird zusammengefasst, wie viele Verbindungen die jeweilige Rolle innehaben. Die dritte Spalte zeigt die entsprechenden Zahlen für Verbindungen, die mit EFO importiert wurden.

ROLLE	ANZAHL CHEMISCHER VERBINDUNGEN	
	GESAMTZAHL	IMPORTIERT AUS EFO
amino acid	20	0
antibiotic	19	1
cytokine	7	7
fungicide	3	1
hormone	13	5
salt	7	0
vitamin	1	1

Rolle bei der Zellkultur abzubilden. Die Unterklassen von *supplement* sind folgendermaßen strukturiert:

- *amino acid role*, mit den Unterklassen *essential*, *conditionally essential* und *non-essential amino acid*
- *antibiotic* (from EFO_0001485)
- *cytokine role* (from EFO_0003787)
- *fungicide* (from EFO_0001823)
- *hormone role* (from EFO_0001824)
- *salt role*
- *serum role*
- *vitamin* (from EFO_0001831)

Die Substanzen selbst sind, mit Ausnahme der Seren, unterhalb der Klasse *chemical compound* eingeordnet. Sie werden ihrer spezifischen Rolle durch die *has_role*-Relation zugeordnet. Tabelle 19 zeigt die Anzahl der chemischen Verbindungen, die in CCONT im Bezug auf die entsprechende Rolle definiert sind oder aus EFO importiert werden. Seren werden unterhalb der bereits beschriebenen Klasse *mixture* eingeordnet. Außerdem referenzieren sie die Klasse *serum_role* mittels der *has_role*-Relation.

Zur Aufrechterhaltung der optimalen Zellkonzentration in der Zellkultur sind die Maße Konfluenzrate, Aussaatdichte und Teilungsverhältnis

von Bedeutung. Entsprechende Klassen werden unterhalb der Klasse *measurement* (EFO_0001444) im *information entity*-Unterbaum von CCONT eingefügt.

Neben der Definition von Klassen werden in der Ontologie auch Relationen zur Darstellung der Beziehungen zwischen den Klassen benötigt. In CCONT werden die bereits in EFO definierten Relationen (siehe Tabelle 21) weiterverwendet. Von besonderer Bedeutung für die Anwendung in CCONT sind die EFO-Relationen *has_role*, *has_quality*, *is_executed_in* und *participates_in*. Zusätzlich wird in CCONT *has_growth_condition* als Relation definiert, um die Beziehung von Zellkulturen zu Wachstumsfaktoren abzubilden.

Bereitstellung der formalen Darstellung der Ontologie

CCONT ist im OWL-Format auf BioPortal mit der Identifikationsnummer 3108 veröffentlicht. Die Ontologie kann unter dem URI *http://bioportal.bioontology.org/ontologies/CCONT* über das Internet abgerufen werden. Zukünftige Versionen von CCONT werden ebenfalls unter dieser Adresse abrufbar sein.

Evaluierung von CCONT

Zunächst wird CCONT *verifiziert*. Verifizierung meint dabei den (technischen) Vorgang, der die Korrektheit der Ontologie sicherstellt [Gómez-Pérez 2001]. Daher wird die Ontologie mit zwei unterschiedlichen Reasoner-Programmen getestet, die als Module für Protégé zur Verfügung stehen. Das erste Programm, HermiT Version 1.3.6 [Motik et al. 2009], berichtet keine Inkonsistenzen in CCONT. Weiterhin wird der FaCT++-Reasoner in der Version 1.5.3 [Tsarkov, Horrocks 2006] herangezogen. Auch bei diesem Test wird die Klassifikation von CCONT ohne Probleme durchgeführt.

Gómez-Pérez definiert *Validierung* als Prozess, der sicherstellt, dass die Ontologie dem System entspricht, das sie repräsentieren soll [Gómez-Pérez 2001]. Um dies zu erreichen, werden die Techniken aus Tabelle 9 möglichst vollständig angewandt:

- *Evaluierung der Nutzbarkeit in der Anwendung* Um die Korrektheit von CCONT sicherzustellen, wird die Ontologie zur Annotation von 15 Zelllinien im Forschungsverbund verwendet. Die Namen dieser Zelllinien werden in Tabelle 20 aufgeführt.

  Als Beispiel wird das Ergebnis der Annotation für die Zelllinie Hep3B in Tabelle 22 gezeigt. Dort werden die Klassen und Attributwerte für ein bestimmtes Individuum, in diesem Fall der Standard Hep3B-Zellkultur von ATCC, dokumentiert. Die Annotationen der übrigen Zelllinien erfolgt analog zu der von Hep3B und ist in [Ganzinger, He et al. 2012], Supporting Information S2 enthalten.

Tabelle 20: Zelllinien, die im Forschungsverbund SFB/TRR77 verwendet werden.

ZELLLINIENTYP	NAME
Humane Zelllinien	Hep3B
	HLE
	HLF
	HuH7
	PLC/PRF-5
	HepaRG
	THLE2
	THLE3
	HepG2
Maus-Zelllinien	Hep1-6
	Hep55.1C
	Hep56.1D
	BNL
Primärzellen	PHH
	PMH

- *Vergleich der Ontologie mit dem Anwendungsbereich* Durch die bei der Entwicklung gewählte Methode ist sichergestellt, dass CCONT sich im Einklang mit bereits etablierten zelllinienbezogenen Ontologien befindet. Weiterhin ist das Wissen aus zelllinienspezifischen Datenbanken in die Entwicklung eingeflossen, da die Klassen aus den Katalogen der in Tabelle 8 aufgeführten Zellbanken abgeleitet sind. Daraus lässt sich schließen, dass CCONT den vorgesehenen Anwendungsbereich in angemessener Weise abdeckt.
- *Bewertung durch Anwender nach einem Kriterienkatalog* CCONT wird im Rahmen des Entwicklungsprozesses nach dem in Tabelle 18 aufgeführten Feldkatalog bewertet. Dabei stellt die vollständige Abdeckung der Felder durch die Ontologie das Zielkriterium dar.
- *Sprachverarbeitungstechniken* Eine Validierung von CCONT durch die Anwendung von Programmen zur automatisierten Verarbeitung natürlicher Sprache ist bisher nicht erfolgt. Dieser Ansatz könnte allerdings zukünftig bedeutsam werden, wenn beispielsweise konventionelle Laborprotokolle oder Versuchsbeschreibungen automatisch ausgewertet werden sollen.
- *Bewertung gegenüber dem Stand der Technik*

Da CCONT eine neu entwickelte Ontologie ist, kann sie nicht mit früheren Versionen verglichen werden. Dieser Ansatz wird jedoch bei der Fertigstellung zukünftiger Versionen von CCONT relevant.

- *Akkreditierung und Zertifizierung* Bis jetzt wurde CCONT keinem formalen Akkreditierungs- oder Zertifizierungsprozess unterzogen. Dies könnte sich in Zukunft ändern, zum Beispiel, wenn Daten an ein öffentliches Metadatenverzeichnis gegeben werden. Ein solches Metadatenverzeichnis könnte die Zertifizierung der beteiligten Ontologien vorschreiben.

5 DISKUSSION

5.1 DISKUSSION DER ERGEBNISSE

Im Rahmen dieser Arbeit wurde eine Datenintegrationsplattform für einen biomedizinischen Forschungsverbund auf Basis einer serviceorientierten Architektur entwickelt. Dabei entstanden folgende Ergebnisse, die von zukünftigen Forschungsverbünden wiederverwendet oder als Grundlage für eigene Entwicklungen benutzt werden können:

- Referenzmodell für Anforderungen
- IT-Architektur für den Forschungsverbund
- Metadatendefinitionen bestehend aus einem kontrollierten Vokabular und einer Ontologie für Zelllinien

Neben diesen konzeptionellen Ergebnissen entstanden zwei Software-Werkzeuge:

- Ein Werkzeug zur automatisierten Erstellung des UML-Modells für die Erzeugung von caBIG-Datendiensten
- Das VOCALIGN-Programm zum Abgleich von Vokabularen mit dem UMLS

Diese Werkzeuge können ebenfalls über den Einsatz im SFB/TRR77 hinaus wiederverwendet werden. In den folgenden Abschnitten werden die genannten Ergebnisse im Detail diskutiert.

Referenzmodell für Anforderungen

Der SFB/TRR77 ist ein Forschungsverbund nach dem SFB/Transregio-Förderprogramm der DFG. Aufgrund der Förderrichtlinien sind alle derartigen Verbünde auf mehrere Standorte verteilt und in einzelne Projekte untergliedert. Durch diese strukturelle Ähnlichkeit kann angenommen werden, dass sich, trotz der unterschiedlichen inhaltlichen Forschungsausrichtung der Verbünde, gemeinsame Anforderungen an die zur Projektdurchführung erforderliche IT-Infrastruktur formulieren lassen. In dieser Arbeit wird zu diesem Zweck ein Referenzmodell für Anforderungen erarbeitet, das als Grundlage für andere SFB/Transregio-Projekte oder auch Forschungsverbünde im Allgemeinen dienen kann. Das Referenzmodell ermöglicht

den Verbünden, sich auf die Erfassung der jeweiligen verbundspezifischen Anforderungen zu konzentrieren, da das Referenzmodell grundlegende Anforderungen bereits abdeckt.

Dabei obliegt es den Nutzern des Referenzmodells, die Anwendbarkeit der einzelnen Referenzanforderungen auf den jeweiligen Fall kritisch zu hinterfragen. Das Referenzmodell basiert auf den Daten eines konkreten Verbundes, die abstrahiert wurden. Somit ergeben sich, wie bei jedem Modell, Grenzen in der Übertragbarkeit in einen anderen Kontext. Zukünftige Forschungsverbünde, die das Referenzmodell für sich nutzen wollen, sollten es daher in eine verbundspezifische Instanz überführen. Dazu kann das in Abschnitt 3.2.1 gezeigte Vorgehen angewandt werden. Bei dieser Adaption des Referenzmodells kann das spezifische Anforderungsmodell dann um zusätzliche Anforderungen ergänzt werden und es können nicht anwendbare Referenzanforderungen entfernt werden. Das Referenzmodell ist ein Werkzeug, das seinen Nutzern hilft, ein konkretes Anforderungsmodell möglichst vollständig zu erstellen. Da es den Anwendern dann möglich ist, auf diese Grundlagen zurückzugreifen, ist zu erwarten, dass das Modell zu einer Reduzierung der bei der Modellierung einzusetzenden Ressourcen beiträgt.

IT-Architektur für den Forschungsverbund

Für die Umsetzung der Datenintegrationsplattform des SFB/TRR77 wird die IT-Architektur als SOA entworfen. Diese Architektur wird unter Verwendung von Komponenten aus dem caBIG-Projekt erfolgreich implementiert werden. Damit ist PELICAN eines der wenigen Systeme in Deutschland, in denen caBIG-Komponenten erfolgreich eingesetzt werden. Durch die Etablierung dieser SOA-Infrastruktur wurde ein flexibles System geschaffen, welches die Anforderungen der Projekte des SFB/TRR77 in hohem Maße erfüllt.

Allerdings ist das SOA-Paradigma nicht ohne Kritik: So führt [Schlimm 2010] verschiedene Probleme auf, die bei der Durchführung von Projekten zur Einführung der SOA beobachtet werden. Die meisten der von ihm beschriebenen Probleme haben ihre Ursache darin, dass bestehende Systeme und Strukturen in eine SOA transformiert werden sollen. Bei diesem Prozess entstehen Interessenskonflikte und andere Reibungspunkte innerhalb der jeweiligen Organisation, die unter Umständen zu einem ungünstigen Verhältnis zwischen Projektkosten und -nutzen führen. Diese Probleme sind jedoch bei einem neu einzuführenden System, wie es in dieser Arbeit beschrieben wird, nicht zu erwarten. Damit überwiegen also die positiven Aspekte einer SOA, die dazu führen, dass ein flexibles und zukunftssicheres System entsteht.

Auch nach der Entscheidung, eine SOA als Grundlage für die IT eines Forschungsverbundes einzusetzen, besteht noch eine Vielzahl an möglichen

Alternativen für die konkrete Umsetzung der SOA. Dies liegt vor allem an der Modularität, die ein zentrales Element des SOA-Paradigmas darstellt. Die Erstellung eines verbundspezifischen Anforderungsmodells gibt dabei entscheidende Hinweise für die Auswahl der geeigneten Komponenten.

Die Architektur wird durch das verbundspezifische Anforderungsmodell, das Komponentenmodell und das Verteilungsmodell immer spezifischer beschrieben und damit auch immer stärker an die konkreten Anforderungen des SFB/TRR77 gekoppelt. Daher muss bei diesen Aspekten der IT-Architektur davon ausgegangen werden, dass bei einer Übertragung auf andere Forschungsverbünde größerer Adaptionsaufwand betrieben werden muss, als dies noch beim Referenzmodell für Anforderungen der Fall ist.

Für PELICAN wird die Anforderung der Projektleiter des SFB/TRR77, die Kontrolle über die Daten behalten zu wollen, als besonders wichtig eingestuft und beeinflusst daher das Architekturmodell in hohem Maße. Als Konsequenz werden die strukturiert vorliegenden Daten in PELICAN als eigene caBIG-Datendienste bereitgestellt. Durch diese Architekturentscheidung kann auch Nutzern mit geringeren IT-Kenntnissen plausibel vermittelt werden, dass sie Kontrolle über die Daten ausüben können, auch wenn sie diese im Verbund bereitstellen.

Die Sicherheitsanforderungen schlagen sich auch an weiteren Stellen der Architektur nieder. So werden die einzelnen Komponenten von PELICAN bezüglich des Netzwerks in verschiedene Zonen verteilt, sodass sensitive Komponenten vom Zugriff aus dem Internet entkoppelt sind. Auch auf Dienstebene werden entsprechende Sicherheitsdienste von caBIG eingesetzt. Eine Umstellung auf ein System, welches auf der Basis kryptographischer Token funktioniert, würde hier zu einer weiteren Erhöhung des Sicherheitsstandards führen.

Durch die Entscheidung für caBIG werden Dienste im Sinne einer SOA auch intern für den Zugriff auf die Daten genutzt. Da die Datendienste projektspezifisch mit dem caCORE SDK-Generator erstellt werden, ist eine Anpassung von caBIG an dieser Stelle nicht erforderlich. Allerdings hat sich die Implementierung eines eigenen caDSR-Dienstes als ungeeignet herausgestellt, zum einen auf Grund der hohen Lizenzgebühren für die notwendige kommerzielle Anwendungsserver-Software als auch auf Grund des komplexen Kuratierungsprozesses, der zur Verwaltung der Metadaten erforderlich ist. Eine Nutzung der zentralen caDSR-Instanz von caBIG scheidet ebenfalls aus, da dort neue Daten nur über die Einbindung der NIH hinzugefügt werden können.

Ein Alternative zu caBIG wäre es, die Daten zentral in einem biomedizinischen Datawarehouse wie i2b2 zu halten und den ganzen Datenbestand über einen Dienst für die SOA-Infrastruktur bereitzustellen. So wurde bereits gezeigt, dass es möglich ist, i2b2 und caBIG zu integrieren und so eine übergreifende SOA zu konzipieren [Matney et al. 2011].

Bei einem zentralen Datawarehouse würde sich allerdings der Aufwand der Datenintegration verlagern: Während die Daten beim föderierten Ansatz von caBIG je nach Datenquelle in einem eigenen Datenmodell bereitgestellt werden und erst zum Zeitpunkt der Abfrage mit Hilfe der Metadaten integriert werden, müssen die Daten bei einem Datawarehouse zum Zeitpunkt der Bereitstellung im Rahmen des ETL-Prozesses integriert werden. Allerdings ist es in i2b2 nicht ohne Weiteres möglich, die Kontrolle über einzelne Datensätze dem beitragenden Projekt zu übertragen, sodass dieser Ansatz für PELICAN verworfen wurde.

Das offizielle Ende des caBIG-Projektes hat keine nachteiligen Auswirkungen auf PELICAN, da nur eine geringe Anzahl an Komponenten aus diesem Programm verwendet wird. Die wichtigste Komponente, das caCORE SDK, wird weiterentwickelt, wobei die derzeit letzte Version 4.5 am 15. August 2013 veröffentlicht wurde.

Im Zuge der Analyse der in den Projekten verwendeten Daten hat sich herausgestellt, dass es hilfreich ist, neben Datendiensten auch die Möglichkeit bereitzustellen, Daten in Dokumentenform auszutauschen. Dies geschieht in PELICAN mittels der Dokumentenaustauschplattform, die einen gesicherten, nur den Mitgliedern des SFB/TRR77 zugänglichen Raum zum Austausch und zur Diskussion bietet. Die Einträge in der Dokumentenaustauschplattform können mithilfe des kontrollierten Vokabulars annotiert werden, sodass sie konsistent mit den übrigen Datenquellen des Verbundes zusammengeführt werden können. Auch auf technischer Ebene ist die Dokumentenaustauschplattform in das SOA-System integriert, da der Zugriff auf die Dokumente grundsätzlich auch über Webservices-Schnittstellen erfolgen kann.

Verbundspezifische Metadaten

Zur Etablierung von PELICAN war es erforderlich, Metadaten für den Verbund zu spezifizieren. Dies geschah zum Einen in Form eines verbundspezifischen kontrollierten Vokabulars, welches mit dem UMLS abgeglichen wurde. Die im Rahmen dieser Arbeit entwickelte Methode zur Erstellung des Vokabulars ist auch außerhalb des SFB/TRR77 anwendbar, während das Vokabular selbst wie geplant verbundspezifisch ist und daher nur in Einzelfällen direkt übernommen werden kann. Weiterhin wurde eine Ontologie zur Beschreibung von Zelllinien entwickelt. Dieses Ergebnis kann direkt auf andere Umgebungen, welche die ontologiebasierte Beschreibung von Zelllinien erfordern, übertragen werden.

Kontrolliertes Vokabular

Im Rahmen dieser Arbeit wird eine Methode entwickelt, mit deren Hilfe ein kontrolliertes Vokabular und somit eine Referenzterminologie für For-

schungsverbünde erstellt werden kann. Mit diesem Ansatz wird versucht, einen optimalen Kompromiss bezüglich der Konformität mit etablierten Vokabularen, der Vollständigkeit und der einfachen Handhabbarkeit durch die Berücksichtigung der Gegebenheiten im Forschungsverbund zu finden. Das Sicherstellen der Konformität wird durch die Bereitstellung des Software-Werkzeuges VOCALIGN unterstützt, welches den Abgleich des lokalen Vokabulars mit zwei Vokabularen aus dem UMLS automatisiert durchführt.

Für den SFB/TRR77 hat sich gezeigt, dass die erforderliche Menge an Termen nicht durch das UMLS bereitgestellt werden kann: Obwohl MeSH und NCIt zu den umfangreichsten Terminologien aus dem UMLS gehören, wurden zum Zeitpunkt der Untersuchungen im April 2013 rund 40% der gesuchten Begriffe dort nicht gefunden. Eine mögliche Erklärung dieser hohen Rate liegt im Forschungsthema des Verbundes: Da sich der SFB/TRR77 zu großen Teilen mit Grundlagenforschung beschäftigt, werden dort viele Bezeichnungen verwendet, die im eher klinisch orientierten UMLS nicht geführt werden. Gleichzeitig belegt die hohe Rate an Fehlschlägen jedoch die Relevanz des gewählten Ansatzes, denn die im Forschungsverbund angewandte Terminologie kann offensichtlich nicht als Untermenge eines etablierten Vokabulars dargestellt werden. Insbesondere reichen die über den von caBIG bereitgestellten EVS verfügbaren Vokabulare dazu nicht aus. Zur weiteren Optimierung könnte es lohnend sein, weitere Vokabulare, auch von außerhalb des UMLS, in den Abgleichprozess von VOCALIGN einzubeziehen.

Ontologie für Zelllinien

Die Ontologie für Zelllinien CCONT erlaubt die umfassende Beschreibung von Zelllinien bezüglich ihrer Identität und Kulturbedingungen. Für den SFB/TRR77 eröffnet dies die Möglichkeit, die verwendeten Zelllinien exakt zu referenzieren. Da mit der Ontologie verschiedene Instanzen einer Zelllinienklasse aus CCONT erstellt werden können, ist es nun auch möglich, das Aufteilen des ursprünglich von einer Zellbank gemeinsam erworbenen Zellmaterials auf die einzelnen Projekte zu beschreiben und es somit unterscheidbar zu machen. Durch diese Aufteilung entstehen Unterzelllinien, die zwar den Namen der ursprünglichen Linie tragen, jedoch voneinander abweichende Eigenschaften entwickeln können.

Damit stellt CCONT die Grundlage bereit, um die Zelllinien des Forschungsverbundes und die mit ihnen durchgeführten Untersuchungen in einem größeren Kontext der biomedizinischen Wissensverarbeitung zu betrachten. So plant beispielsweise das OBO-Konsortium die Integration entsprechender Ontologien zu koordinieren [Smith et al. 2007]. EFO, die Ontologie auf der CCONT basiert, ist zwar nicht Teil von OBO, jedoch basiert EFO auf der OBO-kompatiblen Upper-Level-Ontologie BFO [Grenon

et al. 2004; Maojo et al. 2011]. Somit ist CCONT für die Integration mit anderen Ontologien vorbereitet.

Von den beiden Ontologien, die alternativ zu CCONT ebenfalls für die Domäne der Zelllinien entwickelt wurden, besitzt CLO ebenfalls BFO als übergeordnete Ontologie. MCCL hingegen enthält keine Referenz zu einer übergeordneten Ontologie. Für CLO gibt es mit der BAO eine interessante Adaption: Dort werden Zelllinien im Kontext von Hochdurchsatzanalyseverfahren betrachtet, wobei die Analyseprotokolle selbst allerdings nicht Bestandteil der Ontologie sind. Die Protokolle werden vielmehr über einen Verweis auf einen Datenbankeintrag abgebildet und sollen zukünftig eingebettet werden [Vempati et al. 2012].

5.2 DISKUSSION DES VORGEHENS

Referenzmodell für Anforderungen

Für die Erstellung des Referenzmodells werden Anforderungen in einem konkreten Forschungsverbund erhoben und verallgemeinert. Bei der Erhebung werden unterschiedliche Quellen herangezogen, sodass Anforderungen aus verschiedenen Perspektiven in das Modell eingehen. Durch dieses Vorgehen wird auch dem Risiko begegnet, eine einseitige Sichtweise auf die Anforderungen einzunehmen, und so dem Anspruch der Übertragbarkeit des Referenzmodells nicht gerecht zu werden.

Die Dokumentation der Anforderungen sowohl in Form von UML-Diagrammen als auch in Form von Tabellen hat sich als geeignetes Medium für die technische Dokumentation der Referenzanforderungen herausgestellt. Allerdings sind diese Medien für die Kommunikation mit den biomedizinischen Projektleitern und Projektmitarbeitern, die keinen Informatikhintergrund aufweisen, weniger geeignet. Für die detaillierte Diskussion der Anforderungen mit diesem Personenkreis hat sich hingegen die Arbeit am prototypischen System bewährt. Offenbar fällt es den zukünftigen Anwendern leichter, die Stärken und Defizite eines bestehenden Systemvorschlags zu beschreiben, als auf abstrakter Ebene Anforderungen zu formulieren oder zu kommentieren.

IT-Architektur für den Forschungsverbund

Grundlage für die Entwicklung der IT-Architektur ist die Erstellung des verbundspezifischen Anforderungsmodells auf Basis des Referenzmodells. Bei diesem Überführungsprozess wird geprüft, welche der Referenzanforderungen übernommen werden können, und wie diese SFB/TRR77-spezifisch auszuprägen sind. Auch die im Verbund verwendeten Datenarten werden im Rahmen dieses Schrittes erfasst. Schließlich ergibt sich

ein umfassendes Bild bezüglich der Anforderungen des SFB/TRR77 an die Datenintegrationsplattform.

Aufgrund der Anforderungen wird die grundlegende Architekturentscheidung bezüglich der Datenhaltung getroffen. Für den SFB/TRR77 wurde an dieser Stelle die Auswahl zwischen den beiden etablierten Bioinformatik-Frameworks i2b2 und caBIG getroffen. Von beiden Frameworks war von vornherein bekannt, dass sie grundsätzlich geeignet sind, den SFB/TRR77 bei seinem Ziel, biomedizinische Daten im Sinne der translationalen Forschung zu verarbeiten, zu unterstützen. Für andere Verbünde könnte die Vorauswahl dieser beiden Frameworks jedoch weniger geeignet sein.

Die Anforderungen werden im Anschluss auf Systemeigenschaften und generische Komponenten abgebildet. Die generischen Komponenten werden durch konkrete Komponenten, in der Regel Produkte oder Frameworks, umgesetzt. Da es für die Komponenten verschiedene Alternativen gibt, kann deren Funktionsumfang gegebenenfalls abweichen, sodass sich für eine alternative Komponentenkonfiguration auch eine Verschiebung der Zuordnung von Systemeigenschaften ergeben kann.

Das Komponentenmodell schließlich stellt eine logische Sicht auf das PELICAN-System dar. Eine wesentliche Komponentenklasse besteht aus den mit caCORE SDK erstellten Datendiensten. Grundsätzlich hat sich das Vorgehen bewährt, die Dienste mit Hilfe des entsprechenden Generators zu erstellen. Lediglich die Definition des erforderlichen XMI-Modells hat sich auf Grund des großen und fehlerträchtigen Annotationsaufwandes als weniger geeignet herausgestellt. Das von caBIG angebotene Werkzeug zur semantischen Integration *caAdapter* erleichtert zwar die Annotation, aber auch hier sind noch viele manuelle Schritte erforderlich. Daher hat es sich als sinnvoll erwiesen, im Rahmen des SFB/TRR77 eine Anwendung zu entwickeln, welche die Erzeugung der XMI-Datei noch weiter unterstützt.

Im weiteren Verlauf wird für die Komponenten ein Verteilungsmodell erstellt, das schließlich umgesetzt wird. Für PELICAN werden die Komponenten in virtuellen Maschinen auf einem Server implementiert. Dadurch ist eine große Flexibilität der Ressourcennutzung genauso gegeben wie die Isolation der einzelnen virtuellen Maschinen gegeneinander. Die Komponenten können ebenso auf externen Rechnern betrieben werden um dem erhöhten Schutzbedarf der Projekte Rechnung zu tragen.

Verbundspezifische Metadaten

Kontrolliertes Vokabular

Unter Einbeziehung der SFB/TRR77-Projekte wurde ein projektspezifisches kontrolliertes Vokabular erstellt. Durch das beschriebene Vorgehen wird erreicht, dass in diesem Vokabular nur für den Verbund relevante Terme enthalten sind. Weiterhin wird durch den automatisierten Abgleich

mit dem VOCALIGN-Werkzeug die Interoperabilität mit anderen Datenquellen unterstützt. Ein möglicher alternativer Ansatz zur Festlegung der Referenzterminologie eines Forschungsverbundes wäre es, ein bestimmtes, etabliertes Vokabular als verpflichtend für den Verbund vorzuschreiben. Ein solches Vorgehen hätte zwei Vorteile: Zum einen würde die Notwendigkeit, ein kontrolliertes Vokabular zu pflegen, entfallen. Zum anderen wären die lokal verwendeten Benennungen automatisch konform mit dem etablierten Vokabular. Nachteilig wären hingegen die Existenz nicht benötigter Begriffe sowie die fehlende Möglichkeit, das Vokabular um eigene Terme zu ergänzen.

Ontologie für Zelllinien

Informationen über die verfügbaren Metadaten zu Zelllinien wurden aus den Katalogen von Zellbanken gewonnen. Da die Datenfelder aller betrachteten Zellbanken zusammengeführt wurden, wird angenommen, dass durch dieses Vorgehen eine umfassende Grundlage zur Beschreibung von Zelllinien geschaffen wurde. Dies wird durch den Vergleich von CCONT mit anderen Ontologien wie CLO, MCCL oder EFO bestätigt.

Bei der Entwicklung von CCONT wurden bereits etablierte Ontologien, die Informationen zu Zelllinien enthalten, intensiv auf ihre Eignung zur Integration in CCONT untersucht. Dabei wurde nicht nur die Benennung der Zelllinien betrachtet, sondern es wurden auch weitere Informationen über die Kulturbedingungen der betrachteten Zelllinien berücksichtigt. EFO hat sich schließlich als die geeignete Grundlage für die Erstellung der neuen Ontologie erwiesen. Allerdings wird durch die Entscheidung, auf einer existierenden Ontologie aufzubauen, auch die Grundstruktur der neuen Ontologie vorbestimmt. Es besteht in der Folge weniger Gestaltungsspielraum hinsichtlich der verwendeten Upper-Level-Ontologie, der hierarchischen Organisation der Klassen oder der definierten Relationen. Da mit EFO eine sorgfältig entwickelte Ontologie gewählt wurde, sind diese möglichen Nachteile von untergeordneter Bedeutung und es überwiegt der Vorteil, dass CCONT mit einer etablierten Ontologie integriert ist.

5.3 AUSBLICK

Für die zukünftige Weiterentwicklung ist PELICAN durch die zugrunde liegende SOA gut gerüstet. Das System kann ohne großen Aufwand um weitere Dienste ergänzt werden. Eine besondere Herausforderung werden die aus den Sequenzierungsgeräten der neuesten Generation stammenden Genomdaten darstellen: Zum Ende dieser Arbeit wird die Sequenzierung des Tumor- und Kontroll-Exoms von 60 HCC-Patienten durch den SFB/TRR77 in Auftrag gegeben. Das zu erwartende Rohdatenvolumen von rund 3,6 Terabyte übertrifft den bisherigen Datenbestand deutlich.

Bei der Entwicklung von PELICAN wurde großen Wert darauf gelegt, die Datenschutzanforderungen der Projekte zu berücksichtigen und so eine hohe Akzeptanz des Systems zu erreichen. Dennoch besteht weiteres Forschungspotential bezüglich der Verwendung der Daten und der Würdigung ihrer Urheberschaft. Erste Ansätze wurden hierzu bereits im Rahmen unserer Projekttätigkeit untersucht [He, Ganzinger, Hurdle et al. 2013; He, Ganzinger, Knaup 2012].

Durch die Anwendung der Referenzarchitektur in anderen Forschungsverbünden könnte diese validiert und gegebenenfalls weiterentwickelt werden. Auch wäre es denkbar, weitere Elemente der IT-Architektur zu vereinheitlichen und so für weitere Verbünde verfügbar zu machen. Eine solche Entwicklung könnte schließlich zu einem umfassenden IT-Architekturframework für Forschungsverbünde führen.

Was das Metadatenverzeichnis anbelangt, so könnten sich aus der Etablierung des nationalen Metadatenverzeichnisses neue Aspekte ergeben, die für PELICAN relevant sind. Zwar ist das Verzeichnis zunächst auf die Sammlung von Datenelementen aus klinischen Studien ausgerichtet, jedoch könnte es sinnvoll sein, zu prüfen, inwiefern das Metadatenverzeichnis ein geeigneter Ort wäre, weitere Metadaten aus Forschungsverbünden aufzunehmen.

Das für den Forschungsverbund erstellte kontrollierte Vokabular kann nicht als abgeschlossen betrachtet werden, denn mit der Weiterentwicklung des Verbundes und dem damit verbundenen Anwachsen der Datenbestände werden auch neue Terme benötigt. Daher wird es erforderlich sein, die Methode zur Erfassung des Vokabulars wiederholt anzuwenden, um den zukünftigen Erfordernissen gerecht zu werden.

Auch von CCONT kann nicht angenommen werden, dass die Ontologie in der gegenwärtigen Form vollendet ist. Ebenso wie beim Vokabular wird es erforderlich sein, die angewandte Methode wiederholt anzuwenden, um den Umfang der abgedeckten Zelllinienparameter weiter zu vergrößern. Erste Schritte könnten die Einbindung von Kontaminationstests für Mykoplasma, Zytogenetik oder Immunologie sein. Weiterhin könnte die formale Beschreibung von Laborprotokollen eine interessante Erweiterung für CCONT sein. Ein möglicher Anknüpfungspunkt könnte hier die Experiment Actions (EXACT)-Ontologie sein, die zur Darstellung von biomedizinischen Protokollen zum Einsatz in Laborrobotern entwickelt wurde [Soldatova et al. 2008].

6 ZUSAMMENFASSUNG

In biomedizinischen Forschungsverbünden besteht der Bedarf, Forschungsdaten innerhalb des Verbundes und darüber hinaus gemeinsam zu nutzen. Hierzu wird zunächst ein Anforderungsmodell erstellt, das anschließend konsolidiert und abstrahiert wird. Daraus ergibt sich ein Referenzmodell für Anforderungen, welches Forschungsverbünden als Grundlage für die beschleunigte Erstellung eines eigenen SOA-Systems dienen kann.

Zum Referenzmodell wird weiterhin eine konkrete Instanz als Anforderungsmodell für den durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Sonderforschungsbereich/Transregio 77 „Leberkrebs–von der molekularen Pathogenese zur zielgerichteten Therapie“ beschrieben. Aus dem Anforderungsmodell wird ein IT-Architekturmodell für den Verbund abgeleitet, welches aus Komponentenmodell, Verteilungsmodell und der Sicherheitsarchitektur besteht.

Die Architektur wird unter Verwendung des Cancer Biomedical Informatics Grid (caBIG) umgesetzt. Dabei werden die in den Projekten anfallenden Daten in Datendienste umgewandelt und so für den Zugriff in einer SOA bereitgestellt. Durch die Datendienste kann die Anforderung der Projekte, die Kontrolle über die eigenen Daten zu behalten, weitgehend erfüllt werden: Die Dienste können mit individuellen Zugriffsberechtigungen versehen und dezentral betrieben werden, bei Bedarf auch im Verantwortungsbereich der Projekte selbst. Der Zugriff auf das System erfolgt mittels eines Webbrowsers, mit dem sich die Mitarbeiter des Verbundes unter Verwendung einer individuellen Zugangskennung an einem zentralen Portal anmelden. Zum einfachen und sicheren Austausch von Dokumenten innerhalb des Verbundes wird ein Dokumentenmanagementsystem in die SOA eingebunden.

Um die Forschungsdaten aus verschiedenen Quellen auch auf semantischer Ebene integrieren zu können, werden Metadatensysteme entwickelt. Hierzu wird ein kontrolliertes Vokabular erstellt, das mit der hierfür entwickelten Methode aus den von den Projekten verwendeten Terminologien gewonnen wird. Die so gesammelten Begriffe werden mit standardisierten Vokabularen aus dem Unified Medical Language System (UMLS) abgeglichen. Hierfür wird ein Software-Werkzeug erstellt, das diesen Abgleich unterstützt.

Des Weiteren hat sich im Rahmen dieser Arbeit herausgestellt, dass keine Ontologie existiert, um die in der biomedizinischen Forschung häufig verwendeten Zelllinien einschließlich ihrer Wachstumsbedingungen umfassend abzubilden. Daher wird mit der Cell Culture Ontology (CCONT)

eine neue Ontologie für Zelllinien entwickelt. Dabei wird Wert darauf gelegt, bereits etablierte Ontologien dieses Bereichs soweit wie möglich zu integrieren.

Somit wird hier eine vollständige IT-Architektur auf der Basis einer SOA zum Austausch und zur Integration von Forschungsdaten innerhalb von Forschungsverbünden beschrieben. Das Referenzmodell für Anforderungen, die IT-Architektur und die Metadatenspezifikationen stehen für andere Forschungsverbünde und darüber hinaus als Grundlagen für eigene Entwicklungen zur Verfügung. Gleiches gilt für die entwickelten Software-Werkzeuge zum UMLS-Abgleich von Vokabularen und zur automatisierten Modellerstellung für caBIG-Datendienste.

LITERATUR

Abran A (2004)
Guide to the software engineering body of knowledge, 2004 version: SWEBOK ; a project of the IEEE Computer Society Professional Practices Committee.
IEEE Computer Society, Los Alamitos.

Ahmed F, Perz JF, Kwong S, Jamison PM, Friedman C, Bell BP (2008)
National trends and disparities in the incidence of hepatocellular carcinoma, 1998-2003.
Prev Chronic Dis 5.3: A74.

Alberts B (2008)
Molecular biology of the cell.
5. Aufl. Garland Science Taylor & Francis, New York.

Ammenwerth E (2000)
Die Modellierung von Anforderungen an die Informationsverarbeitung im Krankenhaus.
Dissertation. Universität Heidelberg.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000)
Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.
Nat Genet 25.1: 25–29.

Barallon R, Bauer SR, Butler J, Capes-Davis A, Dirks WG, Elmore E, Furtado M, Kline MC, Kohara A, Los GV, MacLeod RAF, Masters JRW, Nardone M, Nardone RM, Nims RW, Price PJ, Reid YA, Shewale J, Sykes G, Steuer AF, Storts DR, Thomson J, Taraporewala Z, Alston-Roberts C, Kerrigan L (2010)
Recommendation of short tandem repeat profiling for authenticating human cell lines, stem cells, and tissues.
In Vitro Cell Dev Biol Anim 46.9: 727–732.

Becker J, Rosemann M, Schütte R (1995)
Grundsätze ordnungsmäßiger Modellierung.
Wirtschaftsinformatik 37.5: 435–445.

Berman AE, Barnett WK, Mooney SD (2012)
Collaborative software for traditional and translational research.
Hum Genomics 6: 21.

Bih J (2006)
Service Oriented Architecture (SOA) – A New Paradigm to

Implement Dynamic E-business Solutions.
Ubiquity 2006.August: 4:1–4:1.

Bosch FX, Ribes J, Borràs J (1999)
Epidemiology of primary liver cancer.
Semin Liver Dis 19.3: 271–285.

Braun S, Busse J, Franz J, Schultz R, Sevilmis N (2008)
WISSENSNETZ Methodik Handbuch.
[Online im Internet:] URL:
http://www.im-wissensnetz.de/Wissensnetz/CMS/Arbeitspakete/AP9/E16%20Handbuch%20final.pdf/view [Stand: 18.2.2011, 14:33].

Brinkman RR, Courtot M, Derom D, Fostel JM, He Y, Lord P, Malone J, Parkinson H, Peters B, Rocca-Serra P, Ruttenberg A, Sansone SA, Soldatova LN, Stoeckert CJ, Turner JA, Zheng J (2010)
Modeling biomedical experimental processes with OBI.
J Biomed Semantics 1 Suppl 1: S7.

Bundesamt für Sicherheit in der Informationstechnik (2009)
SOA-Security-Kompendium: Sicherheit in Service-orientierten Architekturen.
[Online im Internet:] URL:
https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/SOA/SOA-Security-Kompendium_pdf.pdf?__blob=publicationFile [Stand: 20.11.2013, 14:43].

Buzan T, Buzan B (2002)
Das Mind-Map-Buch: Die beste Methode zur Steigerung Ihres geistigen Potenzials.
5. Aufl. mvg, Landberg München.

Caldwell S, Park SH (2009)
The epidemiology of hepatocellular cancer: from the perspectives of public health problem to tumor biology.
J Gastroenterol 44 Suppl 19: 96–101.

Cimino JJ (1998)
Desiderata for controlled medical vocabularies in the twenty-first century.
Methods Inf Med 37.4-5: 394–403.

Coronado Sd, Haber MW, Sioutos N, Tuttle MS, Wright LW (2004)
NCI Thesaurus: using science-based terminology to integrate cancer research results.
Stud Health Technol Inform 107.Pt 1: 33–37.

Côté RG, Jones P, Apweiler R, Hermjakob H (2006)
The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries.
BMC Bioinformatics 7: 97.

Cui Lin, Shiyong Lu, Xubo Fei, Chebotko A, Darshan Pai, Zhaoqiang Lai, Fotouhi F, Jing Hua (2009)

A Reference Architecture for Scientific Workflow Management Systems and the VIEW SOA Solution.
IEEE Transactions on Services Computing 2.1: 79–92.

Department of Health and Human Services, National Institutes of Health (2011)
NIH Guidelines for Research Involving Recombinant DNA Molecules (NIH Guidelines).
[Online im Internet:] URL:
http://oba.od.nih.gov/oba/rac/Guidelines/NIH_Guidelines.pdf
[Stand: 20.3.2012, 14:32].

Deutsche Forschungsgemeinschaft (2013)
Liste der laufenden Sonderforschungsbereiche.
[Online im Internet:] URL:
http://www.dfg.de/foerderung/programme/listen/index.jsp?id=SFB
[Stand: 11.12.2013, 09:47].

Deutsche Forschungsgemeinschaft (o. J.)
Sonderforschungsbereiche.
[Online im Internet:] URL: *http://www.dfg.de/foerderung/Programme/koordinierte_programme/sfb/*
[Stand: 11.12.2013, 08:39].

DIN 1463 (1987)
Teil 1: Erstellung und Weiterentwicklung von Thesauri - Einsprachige Thesauri.
Deutsches Institut für Normung (1987). Deutsche Norm.

DIN 2342 (1992)
Teil 1: Begriffe der Terminologielehre - Grundbegriffe.
Deutsches Institut für Normung (1992). Deutsche Norm.

Dürbeck S, Kolter J, Pernul G, Schillinger R (2011)
Eine verteilte Autorisierungsinfrastruktur unter Berücksichtigung von Datenschutzaspekten.
Informatik Spektrum 34.3: 265–275.

EASL-EORTEC (2012)
EASL–EORTC Clinical Practice Guidelines: Management of hepatocellular carcinoma.
Eur J Cancer 48.5: 599–641.

Ebersbach A, Glaser M (2005)
Wiki.
Informatik Spektrum 28.2: 131–135.

Endrei M, Ang J, Arsanjani A, Chua S, Comte P, Krogdahl P, Luo M, Newling T (2004)
Patterns: Service-Oriented Architecture and Web Services.
1. Aufl. IBM Corp. International Technical Support Organization, Triangle Park.

Erl T (2008)
SOA: Entwurfsprinzipien für serviceorientierte Architektur.
1. Aufl. Addison Wesley, München.

Ethier JF, Dameron O, Curcin V, McGilchrist MM, Verheij RA, Arvanitis TN, Taweel A, Delaney BC, Burgun A (2013)
A unified structural/terminological interoperability framework based on LexEVS: application to TRANSFoRm.
J Am Med Inform Assoc 20.5: 986–994.

Farazi PA, DePinho RA (2006)
Hepatocellular carcinoma pathogenesis: from genes to environment.
Nat Rev Cancer 6.9: 674–687.

Fenstermacher D, Street C, McSherry T, Nayak V, Overby C, Feldman M (2005)
The Cancer Biomedical Informatics Grid (caBIG™).
Conf Proc IEEE Eng Med Biol Soc 1: 743–746.

Fettke P, Loos P (2004)
Referenzmodellierungsforschung.
Wirtschaftsinformatik 46.5: 331–340.

Fielding RT (2000)
Architectural Styles and the Design of Network-based Software Architectures.
Dissertation. Irvine: University of California.

Foster I, Kesselman C, Tuecke S (2001)
The Anatomy of the Grid: Enabling Scalable Virtual Organizations.
Int J High Perform Comput Appl 15.3: 200–222.

Gamma E, Helm R, Johnson R, Vlissides J (1995)
Design patterns: Elements of reusable object-oriented software.
Addison-Wesley, Boston.

Ganslandt T, Mate S, Helbing K, Sax U, Prokosch HU (2011)
Unlocking Data for Clinical Research – The German i2b2 Experience.
Appl Clin Inform 2.1: 116–127.

Ganzinger M, He S, Breuhahn K, Knaup P (2012)
On the Ontology Based Representation of Cell Lines.
PLoS One 7.11: e48584.

Ganzinger M, Kesselmeier M, Fabian J, Knaup P (2012)
An encapsulated R application for the guided analysis of genomic data.
Stud Health Technol Inform 180: 1144–1146.

Ganzinger M, Noack T, Diederichs S, Longerich T, Knaup P (2011)
Service oriented data integration for a biomedical research network.
Stud Health Technol Inform 169: 867–871.

Ganzinger M, Senghas K, Knaup-Gregori P (2013)
Automatisierte Transformation biomedizinischer Daten in einen caBIG-Datendienst, 471.

In: Handels H, Ingenerf J (Hrsg.): *Im Focus das Leben*.
Pro Business, Berlin.
Gentleman RC, Carey VJ, Bates DM (2004)
Bioconductor: Open software development for computational biology and bioinformatics.
Genome Biol 5: R80.
Gluchowski P (1997)
Data Warehouse.
Informatik Spektrum 20.1: 48–49.
Goecks J, Nekrutenko A, Taylor J, Galaxy Team T (2010)
Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences.
Genome Biol 11.8: R86.
Golbeck J, Fragoso G, Hartel F, Hendler J, Oberthaler J, Parsia B (2003)
The National Cancer Institute's Thésaurus and Ontology.
Web Semant 1.1: 75–80.
Gomaa AI, Khan SA, Toledano MB, Waked I, Taylor-Robinson SD (2008)
Hepatocellular carcinoma: epidemiology, risk factors and pathogenesis.
World J Gastroenterol 14.27: 4300–4308.
Gómez-Pérez A (2001)
Evaluation of ontologies.
International Journal of Intelligent Systems 16.3: 391–409.
Goncalves RS, Parsia B, Sattler U (2011)
Analysing the evolution of the NCI Thesaurus, 1–6.
In: Olive M (Hrsg.): *24th International Symposium on Computer-Based Medical Systems (CBMS), 2011*.
IEEE, Piscataway.
Gonzales-Aguilar A, Ramírez-Posada M, Ferreyra D (2012)
TemaTres: software para gestionar tesauros.
El Profesional de la Información 21.3: 319–325.
Grenon P, Smith B, Goldberg L (2004)
Biodynamic ontology: applying BFO in the biomedical domain.
Stud Health Technol Inform 102: 20–38.
Gruber TR (1993)
A translation approach to portable ontology specifications.
Knowledge Acquisition 5.2: 199–220.
Guarino N, Oberle D, Staab S (2009)
What Is an Ontology?, 1–17.
In: Staab S, Studer R (Hrsg.): *Handbook on Ontologies*.
Springer, Berlin Heidelberg.
Haber MW, Kisler BW, Lenzen M, Wright LW (2007)
Controlled terminology for clinical research: a collaboration between

CDISC and NCI enterprise vocabulary services.
Drug Inf J 41.3: 405–412.

Hayflick L, Moorhead P (1961)
The serial cultivation of human diploid cell strains.
Exp Cell Res 25.3: 585–621.

He S, Ganzinger M, Hurdle JF, Knaup P (2013)
Proposal for a data publication and citation framework when sharing biomedical research resources.
Stud Health Technol Inform 192: 1201.

He S, Ganzinger M, Knaup P (2012)
The intellectual property management for data sharing in a german liver cancer research network.
Stud Health Technol Inform 180: 891–895.

Hoffman M, Arnoldi C, Chuang I (2005)
The clinical bioinformatics ontology: a curated semantic network utilizing RefSeq information.
Pacific Symposium on Biocomputing: 139–150.

IEEE 830 (1998)
IEEE Recommended Practice for Software Requirements Specifications.
Institute of Electrical and Electronics Engineers (1998). IEEE Standard.

ISO/IEC 11179-1:2004(E) (2004)
Information technology — Metadata registries (MDR) — Part 1: Framework.
International Organization for Standardization (2004). International Standard.

ISO/IEC/IEEE 24765 (2010)
Systems and software engineering – Vocabulary.
International Organization for Standardization (2010). International Standard.

ISO/IEC 19500-2:2012(E) (2012)
Information technology - Object Management Group Common Object Request Broker Architecture (CORBA), Interoperability.
International Organization for Standardization (2012). International Standard.

JSR 286 (2008)
Java™ Portlet Specification Version 2.0.
Java Community Process (2008). Java Specification Request
[Online im Internet:] URL: *http://www.jcp.org/en/jsr/detail?id=286*
[Stand: 27.11.2013, 13:59].

Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D (2011)
Global cancer statistics.
CA Cancer J Clin 61.2: 69–90.

Kanehisa M, Goto S (2000)
KEGG: kyoto encyclopedia of genes and genomes.
Nucleic Acids Res 28.1: 27–30.

Komatsoulis GA (2011)
Collaboration in cancer reserch community: Cancer Biomedical Informatics Grid (caBIG), 261–280.
In: Williams AJ, Hupcey MAZ, Ekins S (Hrsg.): *Collaborative computational technologies for biomedical research*.
Wiley, Hoboken.

Komatsoulis GA, Warzel DB, Hartel FW, Shanbhag K, Chilukuri R, Fragoso G, Coronado Sd, Reeves DM, Hadfield JB, Ludet C, Covitz PA (2008)
caCORE version 3: Implementation of a model driven, service-oriented architecture for semantic interoperability.
J Biomed Inform 41.1: 106–123.

Kossmann D, Leymann F (2004)
Web Services.
Informatik Spektrum 27.2: 117–128.

Krikov S, Price RC, Matney SA, Allen-Brady K, Facelli JC (2011)
Enabling GeneHunter as a Grid Service.
Methods Inf Med 50.4: 364–371.

Kunz I, Lin MC, Frey L (2009)
Metadata mapping and reuse in caBIG™.
BMC Bioinformatics 10.Suppl 2: S4.

Kushida T, Takagi T, Fukuda KI (2006)
Event ontology: a pathway-centric ontology for biological processes.
Pacific Symposium on Biocomputing: 152–163.

Langella S, Hastings S, Oster S, Pan T, Sharma A, Permar J, Ervin D, Cambazoglu BB, Kurc T, Saltz J (2008)
Sharing Data and Analytical Resources Securely in a Biomedical Research Grid Environment.
J Am Med Inform Assoc 15.3: 363–373.

Langella S, Oster S, Hastings S, Siebenlist F, Phillips J, Ervin D, Permar J, Kurc T, Saltz J (2007)
The Cancer Biomedical Informatics Grid (caBIG) Security Infrastructure.
AMIA Annu Symp Proc: 433–437.

Lindberg DA, Humphreys BL, McCray AT (1993)
The Unified Medical Language System.
Methods Inf Med 32.4: 281–291.

Lowe HJ, Barnett GO (1994)
Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches.
JAMA 271.14: 1103–1108.

Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, Zhukova A, Brazma A, Parkinson H (2010)
Modeling sample variables with an Experimental Factor Ontology.
Bioinformatics 26.8: 1112–1118.

Maojo V, Crespo J, García-Remesal M, La Iglesia Dd, Perez-Rey D, Kulikowski C (2011)
Biomedical Ontologies: Toward Scientific Debate.
Methods Inf Med 50.3: 203–216.

Matney SA, Bradshaw RL, Livne OE, Bray BE, Frey L, Mitchell JA, Narus SP (2011)
Developing a semantic framework for clinical and translational research.
AMIA Summit on Translational Bioinformatics, San Francisco, CA.

Matos Pd, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C (2009)
Chemical Entities of Biological Interest: an update.
Nucleic Acids Res 38.Database: D249–D254.

Mayring P (2010)
Qualitative Inhaltsanalyse: Grundlagen und Techniken.
11. Aufl. Beltz, Weinheim.

McCusker JP, Phillips JA, Beltrán A, Finkelstein A, Krauthammer M (2009)
Semantic web data warehousing for caGrid.
10.Suppl 10: S2.

Milanovic N, Malek M (2004)
Current solutions for Web service composition.
IEEE Internet Comput 8.6: 51–59.

Molecular Connections (2011)
Ontology | Molecular Connections.
[Online im Internet:] URL:
http://www.molecularconnections.com/home/ontology [Stand: 3.8.2011, 15:03].

Motik B, Shearer R, Horrocks I (2009)
Hypertableau Reasoning for Description Logics.
J Artif Intell Res 36: 165–228.

Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I (2010)
Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2).
J Am Med Inform Assoc 17.2: 124–130.

Naedele M (2003)
Standards for XML and web services security.
Computer (Long Beach Calif) 36.4: 96–98.

Nardi BA, Schiano DJ, Gumbrecht M, Swartz L (2004)
Why we blog.
Commun ACM 47.12: 41.

National Institutes of Health (o. J.)
UMLS Terminology Services – Home.
[Online im Internet:] URL: *https://uts.nlm.nih.gov/home.html* [Stand: 30.4.2013, 12:44].

Ngouongo SM, Löbe M, Stausberg J (2013)
The ISO/IEC 11179 norm for metadata registries: does it cover healthcare standards in empirical research?
J Biomed Inform 46.2: 318–327.

NISO (2005)
Guidelines for the construction, format, and management of monolingual controlled vocabulary: ANSI/NISO Z39.19-2005.
NISO Press, Bethesda.

Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, Jonquet C, Rubin DL, Storey MA, Chute CG, Musen MA (2009)
BioPortal: ontologies and integrated data resources at the click of a mouse.
Nucleic Acids Res 37.Web Server: W170–W173.

Noy NF, Crubezy M, Fergerson RW, Knublauch H, Tu SW, Vendetti J, Musen MA (2003)
Protégé-2000: an open-source ontology-development and knowledge-acquisition environment.
AMIA Annu Symp Proc: 953.

Noy NF, McGuinness DL (2001)
Ontology Development 101: A Guide to Creating Your First Ontology.
Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

OASIS SAML (2005)
Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0.
OASIS Organization for the Advancement of Structured Information Standards (2005a). OASIS Standard
[Online im Internet:] URL:
http://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf
[Stand: 20.11.2013, 15:20].

OASIS XACML (2005)
eXtensible Access Control Markup Language (XACML) Version 2.0.
OASIS Organization for the Advancement of Structured Information Standards (2005b). OASIS Standard
[Online im Internet:] URL: *http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf* [Stand: 20.11.2013, 15:25].

OASIS SOA-RM (2006)
Reference Model for Service Oriented Architecture 1.0.
OASIS Organization for the Advancement of Structured Information Standards (2006a). OASIS Standard
[Online im Internet:] URL: *http://docs.oasis-open.org/soa-rm/v1.0/*
[Stand: 21.6.2010, 12:43].

OASIS WSS (2006)
Web Services Security: SOAP Message Security 1.1 (WS-Security 2004).
OASIS Organization for the Advancement of Structured Information Standards (2006b). OASIS Standard
[Online im Internet:] URL: *http://docs.oasis-open.org/wss/v1.1/wss-v1.1-spec-errata-os-SOAPMessageSecurity.htm* [Stand: 20.11.2013, 14:06].

OASIS BPEL (2007)
Web Services Business Process Execution Language Version 2.0.
OASIS Organization for the Advancement of Structured Information Standards (2007). OASIS Standard
[Online im Internet:] URL:
http://docs.oasis-open.org/wsbpel/2.0/OS/wsbpel-v2.0-OS.html
[Stand: 19.11.2013, 12:44].

Obrst L, Ceusters W, Mani I, Ray S, Smith B (2007)
The Evaluation of Ontologies, 139–158.
In: Baker CJO, Cheung KH (Hrsg.): *Semantic Web*.
Springer Science+Business Media LLC, Boston.

Osborne JD, Flatow J, Holko M, Lin SM, Kibbe WA, Zhu L, Danila MI, Feng G, Chisholm RL (2009)
Annotating the human genome with Disease Ontology.
BMC Genomics 10.Suppl 1: S6.

Oster S, Langella S, Hastings S, Ervin D, Madduri R, Phillips J, Kurc T, Siebenlist F, Covitz P, Shanbhag K, Foster I, Saltz J (2007)
caGrid 1.0: an enterprise Grid infrastructure for biomedical research.
J Am Med Inform Assoc 15.2: 138–149.

Ouellette MM (2000)
The establishment of telomerase-immortalized cell lines representing human chromosome instability syndromes.
Hum Mol Genet 9.3: 403–411.

Perz JF, Armstrong GL, Farrington LA, Hutin YJF, Bell BP (2006)
The contributions of hepatitis B virus and hepatitis C virus infections to cirrhosis and primary liver cancer worldwide.
J Hepatol 45.4: 529–538.

Pohl K (2007)
Requirements engineering: Grundlagen, Prinzipien, Techniken.
1. Aufl. dpunkt.verlag, Heidelberg.

Ponniah P (2010)
Data warehousing fundamentals for IT professionals.
2. Aufl. John Wiley & Sons, Hoboken.
Pschyrembel W, Bach M (2010)
Pschyrembel Klinisches Wörterbuch 2011.
262. Aufl. de Gruyter, Berlin.
R Development Core Team (2011)
R: A Language and Environment for Statistical Computing. Vienna Austria.
[Online im Internet:] URL: *http://www.R-project.org/* [Stand: 3.12.2013, 13:03].
Richter JP, Haller H, Schrey P (2005)
Serviceorientierte Architektur.
Informatik Spektrum 28.5: 413–416.
Robbins SL, Cotran RS, Kumar V (2010)
Pathologic basis of disease.
8. Aufl. Saunders/Elsevier, Philadelphia.
Rogers JE (2006)
Quality assurance of medical ontologies.
Methods Inf Med 45.3: 267–274.
Romano P, Manniello A, Aresu O, Armento M, Cesaro M, Parodi B (2009)
Cell Line Data Base: structure and recent improvements towards molecular authentication of human cell lines.
Nucleic Acids Res 37.Database: D925–D932.
Rosse C, Mejino JLV (2003)
A reference ontology for biomedical informatics: the Foundational Model of Anatomy.
J Biomed Inform 36.6: 478–500.
Saltz J, Oster S, Hastings S, Langella S, Kurc T, Sanchez W, Kher M, Manisundaram A, Shanbhag K, Covitz P (2006)
caGrid: design and implementation of the core architecture of the cancer biomedical informatics grid.
Bioinformatics 22.15: 1910–1916.
Saltz J, Oster S, Hastings S, Langella S, Ferreira R, Permar J, Sharma A, Ervin D, Pan T, Catalyurek U, Kurc T (2008)
Translational research design templates, Grid computing, and HPC, 1–15.
In: Translational research design templates, Grid computing, and HPC. *IEEE International Symposium on Parallel and Distributed Processing, 2008*.
IEEE, Piscataway.
Sarntivijai S, Ade AS, Athey BD, States DJ (2008)
A bioinformatics analysis of the cell line nomenclature.
Bioinformatics 24.23: 2760–2766.

Schlimm N (2010)
Serviceorientierte Architektur – eine Standortanalyse.
Informatik Spektrum 33.3: 282–287.

Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P, Virtanen C, Winegarden N, Cheng J, Gingeras T, Petronis A (2006)
Microarray-based DNA methylation profiling: technology and applications.
Nucleic Acids Res 34.2: 528–542.

El-Serag HB, Marrero JA, Rudolph L, Reddy KR (2008)
Diagnosis and treatment of hepatocellular carcinoma.
Gastroenterology 134.6: 1752–1763.

El-Serag HB, Rudolph KL (2007)
Hepatocellular carcinoma: epidemiology and molecular carcinogenesis.
Gastroenterology 132.7: 2557–2576.

Sezov RJ (2012)
Liferay in action: The official guide to Liferay portal development.
Manning, Shelter Island.

Shaffer LG, Slovak ML, Campbell LJ (2009)
ISCN 2009: An international system for human cytogenetic nomenclature (2009).
Karger, Basel.

Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, Goldberg LJ, Eilbeck K, Ireland A, Mungall CJ, Leontis N, Rocca-Serra P, Ruttenberg A, Sansone SA, Scheuermann RH, Shah N, Whetzel PL, Lewis S (2007)
The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.
Nat Biotechnol 25.11: 1251–1255.

Soldatova LN, Aubrey W, King RD, Clare A (2008)
The EXACT description of biomedical protocols.
Bioinformatics 24.13: i295–i303.

Stal M (2006)
Using architectural patterns and blueprints for service-oriented architecture.
IEEE Software 23.2: 54–61.

Stausberg J, Löbe M, Verplancke P, Drepper J, Herre H, Löffler M (2009)
Foundations of a metadata repository for databases of registers and trials.
Stud Health Technol Inform 150: 409–413.

Talia D (2013)
Workflow Systems for Science: Concepts and Tools.
ISRN Software Engineering 2013.1: 1–15.

Thomas O (2006)
Das Referenzmodellverständnis in der Wirtschaftsinformatik: Historie, Literaturanalyse und Begriffsexplikation. Hrsg. von Peter Loos. Saarbrücken.
[Online im Internet:] URL: *http://scidok.sulb.uni-saarland.de/volltexte/2006/636/pdf/IWi-Heft_187.pdf* [Stand: 19.11.2013, 13:39].

Tresch M (1996)
Middleware: Schlüsseltechnologie zur Entwicklung verteilter Informationssysteme.
Informatik Spektrum 19.5: 249–256.

Tsarkov D, Horrocks I (2006)
FaCT++ Description Logic Reasoner: System Description, 292–297.
In: Hutchison D, Kanade T, Kittler J, Kleinberg JM, Mattern F, Mitchell JC, Naor M, Nierstrasz O, Pandu Rangan C, Steffen B, Sudan M, Terzopoulos D, Tygar D, Vardi MY, Weikum G, Furbach U, Shankar N (Hrsg.): *Lecture Notes in Computer Science*.
Springer, Berlin Heidelberg.

Tuttle MS, Blois MS, Erlbaum MS, Nelson SJ, Sherertz DD (1988)
Toward a bio-medical thesaurus: building the foundation of the UMLS, 191.
In: Greenes R (Hrsg.): *Computer applications in medical care*.
IEEE Computer Society Press, New York.

Universität Heidelberg, Medizinische Hochschule Hannover (2009)
Liver Cancer - From Molecular Pathogenesis to Targeted Therapies: Funding Proposal 01.01.2010-31.12.2013.

Urbanek S (2003)
Rserve: A Fast Way to Provide R Functionality to Applications,
in: Hornik K, Leisch F, Zeileis A (Hrsg.): *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*.
Wien.

Varmus H (2012)
About NCIP — CBIIT: Welcome to the NCI Center for Biomedical Informatics and Information Technology.
[Online im Internet:] URL: *http://cbiit.nci.nih.gov/ncip/about-ncip* [Stand: 25.11.2013, 8:24].

Vempati UD, Przydzial MJ, Chung C, Abeyruwan S, Mir A, Sakurai K, Visser U, Lemmon VP, Schürer SC, Cox D (2012)
Formalization, Annotation and Analysis of Diverse Drug and Probe Screening Assay Datasets Using the BioAssay Ontology (BAO).
7.11: e49198.

Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC (2011)
BioAssay Ontology (BAO): a semantic description of bioassays and

high-throughput screening results.
BMC Bioinformatics 12.1: 257.
VMware Inc. (2013)
vSphere-Sicherheit: ESXi 5.5, vCenter Server 5.5. Palo Alto.
Vrandečić D (2009)
Ontology Evaluation, 293–313.
In: Staab S, Studer R (Hrsg.): *Handbook on Ontologies*.
Springer, Berlin Heidelberg.
Weibel S (1997)
The Dublin Core: A Simple Content Description Model for Electronic Resources.
Bull Am Soc Inf Sci 24.1: 9–11.
Weinberg RA (2007)
The biology of cancer.
GS Garland Science, New York.
Wiley A, Gagne B (2012)
caCORE SDK Version 4.3 Object Relational Mapping Guide.
[Online im Internet:] URL: *https://wiki.nci.nih.gov/display/caCORE/caCORE+SDK+Version+4.3+Object+Relational+Mapping+Guide*
[Stand: 9.12.2013, 17:35].
Winter A, Winter Aa, Becker K, Bott O, Brigl B, Gräber S, Hasselbring W, Haux R, Jostes C, Penger OS, Prokosch HU, Ritter J, Schütte R, Terstappen A (1999)
Referenzmodelle für die Unterstützung des Managements von Krankenhausinformationssystemen.
Informatik, Biometrie und Epidemiologie in Medizin und Biologie 30.4: 173–189.
Woll K, Manns M, Schirmacher P (2013)
Sonderforschungsbereich SFB/TRR77: Leberkrebs.
Pathologe 34.S2: 232–234.
W3C RDF (2004)
RDF Primer.
World Wide Web Consortium (2004a). W3C Recommendation
[Online im Internet:] URL: *http://www.w3.org/TR/rdf-primer/* [Stand: 11.12.2013, 22:55].
W3C WSA (2004)
Web Services Architecture.
World Wide Web Consortium (2004b). Working Group Note
[Online im Internet:] URL:
http://www.w3.org/TR/2004/NOTE-ws-arch-20040211/ [Stand: 16.7.2013,].
W3C XML (2006)
Extensible Markup Language (XML) 1.1.
World Wide Web Consortium (2006). W3C Recommendation

[Online im Internet:] URL: *http://www.w3.org/TR/xml11/* [Stand: 18.11.2013,].

W3C SOAP (2007)
SOAP Version 1.2 Part 0: Primer.
World Wide Web Consortium (2007). W3C Recommendation
[Online im Internet:] URL: *http://www.w3.org/TR/soap12-part0/*
[Stand:].

W3C OWL (2009)
OWL 2 Web Ontology Language - Document Overview.
World Wide Web Consortium (2009a). W3C Recommendation
[Online im Internet:] URL:
http://www.w3.org/TR/2009/REC-owl2-overview-20091027/ [Stand: 27.10.2011, 11:14].

W3C SKOS (2009)
SKOS simple knowledge organization system reference.
World Wide Web Consortium (2009b). W3C Recommendation
[Online im Internet:] URL: *http://www.w3.org/TR/skos-reference/*
[Stand: 19.11.2013, 14:36].

EIGENE VERÖFFENTLICHUNGEN

ZEITSCHRIFTENAUFSÄTZE

Capurro D, ***Ganzinger M***, Pérez Lu J, Knaup P (2014)
Effectiveness of eHealth Interventions and Information Needs in Palliative Care – a Systematic Literature Review.
J Med Internet Res (zur Publikation angenommen).

Ganzinger M, He S, Breuhahn K, Knaup P (2012)
On the Ontology Based Representation of Cell Lines.
PLoS One 7.11: e48584.

Spitalewsky K, Rochon J, ***Ganzinger M***, Knaup P (2013)
Potential and requirements of IT for ambient assisted living technologies. Results of a Delphi study.
Methods Inf Med 52.3: 231–8, S1–3.

WEITERE MEDLINE-GEFÜHRTE ORIGINALARBEITEN

Capurro D, ***Ganzinger M***, Pérez-Lu JE (2013)
Palliative care from a medical informatics perspective in Chile, Germany, and Peru.
Stud Health Technol Inform 192: 1013.

Ganzinger M, Kesselmeier M, Fabian J, Knaup P (2012)
An encapsulated R application for the guided analysis of genomic data.
Stud Health Technol Inform 180: 1144–1146.

Ganzinger M, Knaup P (2013)
Semantic prerequisites for data sharing in a biomedical research network.
Stud Health Technol Inform 192: 938.

Ganzinger M, Noack T, Diederichs S, Longerich T, Knaup P (2011)
Service oriented data integration for a biomedical research network.
Stud Health Technol Inform 169: 867–871.

He S, ***Ganzinger M***, Hurdle JF, Knaup P (2013)
Proposal for a data publication and citation framework when sharing biomedical research resources.
Stud Health Technol Inform 192: 1201.

WEITERE KONFERENZBEITRÄGE

Ganzinger M, Müller C, Knaup-Gregori P (2013)
E-Health Anwendungen in der Palliativversorgung, 457.
In: Handels H, Ingenerf J (Hrsg.): *Im Focus das Leben*.
Pro Business, Berlin.

Ganzinger M, Noack T, Knaup P (2011)
Umsetzung eines Intellectual Capital Review Board (ICRB) in einem biomedizinischen Forschungsverbund mit caBIG, 56.
In: Blettner M, Kaatsch P, Kaiser M, Klug S (Hrsg.): *Biometrie, Epidemiologie und Informatik - Gemeinsam forschen für Gesundheit.*
Kirchheim und Co, Mainz.

Ganzinger M, Senghas K, Knaup-Gregori P (2013)
Automatisierte Transformation biomedizinischer Daten in einen caBIG-Datendienst, 471.
In: Handels H, Ingenerf J (Hrsg.): *Im Focus das Leben*.
Pro Business, Berlin.

Knaup P, ***Ganzinger M***, Noack T, Kohl C, Lorenzo-Bermejo J, Dickhaus H, Kieser M (2010)
Konzept der IT-Plattform eines biomedizinischen Forschungsverbundes zu Diagnostik und Therapie von Leberkrebs.
In: Schmücker P (Hrsg.): *Effiziente und wirtschaftliche Gesundheitsversorgung von heute und morgen - nur mit medizinischer Dokumentation, medizinischer Informatik, medizinischer Biometrie und Epidemiologie.*
Antares-Computer-Verl, Dietzenbach.

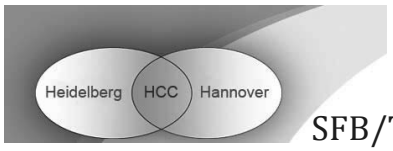
Stephan S, Müller C, ***Ganzinger M***, Knaup-Gregori P (2012)
Entwicklung einer mobilen Anwendung für die Palliativversorgung, in: Goltz U, Ehrich HD (Hrsg.): *Informatik 2012*.
Gesellschaft für Informatik, Bonn.

ANHANG

A

FRAGEBOGEN

Im Folgenden ist der Fragebogen, welcher beim TRR-Retreat verwendet wurde, wiedergegeben:



SFB/TRR77 Livercancer
Project Z2-IT: Questionnaire

You may answer the questionnaire in either English or German.

1. You and your project

Project number: _____ Your name: _____

Function within the project: _____

E-mail: _____ Phone: _____

2. What kind of data do you use?

☐ Microarray data What kind of microarray: _____

☐ Images What kind of images: _____

☐ Text Please specify: _____

☐ Other Please specify: _____

3. In which format are data stored?

☐ Excel files ☐ Database (please specify: _____)

☐ XML files ☐ Character separated value (CSV) files

☐ Text files ☐ Image files

☐ other (please specify:_____)

4. When are data produced?

☐ just once (please specify):

☐ at the beginning ☐ in year_____ ☐ at the end of envisioned period of sponsorship

☐ continuously (please state expected periods:_____)

☐ other (please specify: _____)

5. Are you concerned about the confidentiality of your data?

☐ yes, my data are confidential and can only be accessed by my project

☐ yes, but all TRR partners can have free access

☐ yes, but TRR partners can request access

☐ no, anybody can access my data

6. Would you provide your data in a common TRR77 information system?

☐ yes ☐ only if my confidentiality requirements are met ☐ no

7. How do you analyze your data? (esp. approach, frequency)

8. Are your data related to other projects?

9. Who would be interested in your data?

☐ all TRR projects

☐ a number of TRR projects, please specify if possible:_____

☐ someone else, please specify if possible:_____

10. What data of other TRR projects could be helpful to answer your research questions?

11. What cross TRR questions are you interested in?

12. What could be new research questions to be answered by combining TRR research projects (What is your vision)?

13. Additional Comments and Suggestions

For additional information and support please contact:

Prof. Dr. Petra Knaup-Gregori, Matthias Ganzinger
Universität Heidelberg
Institut für Medizinische Biometrie und Informatik, Sektion Medizinische Informatik
E-mail: matthias.ganzinger@med.uni-heidelberg.de, phone: 06221 56-7368

B

GESPRÄCHSLEITFADEN

B.1 TERMINOLOGIE

- Welche Begriffe / Benennungen spielen eine Rolle?
- Wie stehen die Begriffe zueinander in Beziehung?
- Welche Vorgänge / Prozesse finden im Projekt statt?
- Gibt es Zwischenprodukte?
- Welche Auswertungen finden statt?

B.2 DATENELEMENTE

- Welche Daten liegen vor?
- Welche Struktur haben die Daten?
- In welcher Form liegen die Daten vor (Papier, elektronisch)?
- Gibt es evtl. ein Beispiel?
- Liegen zu den Daten schon Metadaten vor?
- Wie ist der Lebenszyklus der Daten
- Werden die Daten archiviert?

B.3 MODELLE

- Wie sind die Beziehungen zwischen den Datenelementen?

B.4 EXTERNE BEZÜGE

- Sind externe Stellen an der Auswertung der Daten beteiligt?
- Welche externen Datenbanken werden verwendet?
- Wo stehen diese?

B.5 NUTZUNG VON IT

- Werden die Daten elektronisch erzeugt?
- Werden die Daten erfasst?
- Erfolgt ein automatischer Datenbankzugriff?
- Werden Daten mit anderen ausgetauscht?
- Gibt es Medienbrüche?

C

REFERENZANFORDERUNGEN

C.1 ZIELE

In diesem Abschnitt werden die Referenzziele für Forschungsverbünde tabellarisch zusammengestellt. Sie entsprechen der Diagrammdarstellung von Abbildung 6 auf Seite 53.

EIGENSCHAFT	ERKLÄRUNG
Nummer	RZ1
Bezeichnung	Forschungsprojekt durchführen
Beschreibung	Das oberste Ziel eines Forschungsverbundes ist die Durchführung des intendierten Forschungsvorhabens. In der Regel ist dieses Ziel durch finanzierende Fördereinrichtungen vorgegeben.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RZ2
Bezeichnung	Beantworten von Forschungsfragen
Beschreibung	Ein Forschungsverbund besitzt explizit oder implizit Fragestellungen, die erforscht werden sollen. Ein Ziel des Verbundes ist die Beantwortung dieser Forschungsfragen.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RZ3
Bezeichnung	Erzeugen, Speichern und Abrufen von Daten
Beschreibung	Im Rahmen der Arbeit des Verbundes werden Daten erzeugt, welche durch Analyse zur Beantwortung der jeweiligen Forschungsfragen beiträgt. Diese Daten werden gespeichert und zum Abruf bereitgestellt.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RZ4
Bezeichnung	Daten analysieren
Beschreibung	Die rohen Daten eines Forschungsverbundes sind an sich ohne Bedeutung Sie müssen analysiert und in Wissen verwandelt werden, um zur Beantwortung von Forschungsfragen und damit zu den Zielen des Verbundes beizutragen.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RZ5
Bezeichnung	Datenzugriff und -nutzung kontrollieren
Beschreibung	Die im Verbund erzeugten Daten sind schützenswert. Zum Einen kann dies personenbezogene Daten wie Patientendaten betreffen, zum Anderen sind unveröffentlichte Daten das geistige Eigentum ihrer Erzeuger
Gewichtung	mittel

C.2 ANFORDERUNGEN

Zur Verfeinerung der im vorherigen Abschnitt aufgezeigten Ziele eines Forschungsverbundes werden diese Ziele durch Anforderungen konkretisiert. Ein Übersichtsdiagramm zu den Referenzanforderungen ist in Abbildung 7 auf Seite 54 dargestellt.

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA1
Bezeichnung	Daten erzeugen
Beschreibung	Im Forschungsverbund müssen im Rahmen der Durchführung des Forschungsvorhabens neue Daten erzeugt werden.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA2
Bezeichnung	Abruf externer Daten
Beschreibung	Externe Datenquellen müssen erschlossen werden, um die im Forschungverbund erzeugten Daten zu ergänzen. Dies können beispielsweise genomische Annotationen sein.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA3
Bezeichnung	Daten darstellen
Beschreibung	Die im System verfügbaren internen und externen Daten müssen über eine geeignete Darstellung verfügen. Die Darstellung muss präzise definiert sein und die Auffindbarkeit der Daten muss gewährleistet sein.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA4
Bezeichnung	Syntax definieren
Beschreibung	Das Format der Daten muss syntaktisch hinreichend beschrieben sein. Dies ist erforderlich, damit die verbundeigenen Datenquellen korrekt abgerufen und analysiert werden können.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA5
Bezeichnung	Datenmodell definieren
Beschreibung	Zur effizienten internen Verwaltung der Daten ist je ein logisches und physikalisches Datenmodell erforderlich.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA6
Bezeichnung	Daten identifizieren
Beschreibung	Datenquellen müssen verbundweit eindeutig identifizierbar sein. Die Identität und der Ort der Datenquelle müssen im Verbund bekannt sein, beispielsweise durch einen Verzeichnisdienst.
Gewichtung	niedrig

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA7
Bezeichnung	Semantik definieren
Beschreibung	Die Semantik der in den Datenquellen enthaltenen Felder muss genau definiert sein, sodass die Interoperabilität zwischen den Diensten auf semantischer Ebene gewährleistet ist.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA8
Bezeichnung	Geistiges Eigentum verwalten
Beschreibung	Der Zugriff auf die im Verbund bereitgestellten Daten muss bezüglich des Eigentums der Daten geregelt werden. Dies ist eine wesentliche Voraussetzung für die Akzeptanz einer solchen Plattform durch die Anwender.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA9
Bezeichnung	Daten schützen
Beschreibung	Die Daten müssen vor unberechtigtem Zugriff geschützt werden. Dies ist von besonderer Bedeutung wenn personenbezogene Daten wie Patientendaten betroffen sind.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA10
Bezeichnung	Ergebnisse anzeigen
Beschreibung	Zur optimalen Unterstützung der Forscher bei der Bearbeitung der Ziele des Verbundes müssen die Analyseergebnisse in geeigneter Weise aufbereitet und angezeigt werden. Die Wahl der Darstellungsform richtet sich sowohl nach dem Bedarf der Nutzer als auch nach der Art der Ergebnisse.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA11
Bezeichnung	Daten integrieren
Beschreibung	Auswertungen, die auf Daten aus mehreren Projekten des Verbundes zurückgreifen erfordern die Integration dieser Daten. Für die Analyse werden die entsprechend dargestellten Daten zusammengeführt.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA12
Bezeichnung	Analysemethoden definieren
Beschreibung	Zur Bearbeitung der Forschungsfragestellungen müssen Daten mit den jeweils geeigneten Analysemethoden betrachtet werden. Die Methoden müssen ausgewählt oder gegebenenfalls auch neu entwickelt werden.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA13
Bezeichnung	Analyseprozess definieren
Beschreibung	Der Prozess zur Analyse von Daten setzt sich aus einer oder mehreren Datenquellen sowie einer oder mehreren Analysemethoden zusammen. Diese Komponenten können aneinandergereiht und verknüpft werden.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA14
Bezeichnung	Statischer Workflow
Beschreibung	Ein statischer Workflow ist ein festgelegter Analyseprozess. Er bietet dem Anwender keine Möglichkeit, die enthaltenen Komponenten auszuwählen oder in der Reihenfolge zu verändern. Je nach Workflow kann jedoch die Möglichkeit zur Konfiguration von Parametern *innerhalb* der Workflowelemente gegeben sein.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	RA14
Bezeichnung	Dynamischer Workflow
Beschreibung	Der Anwender kann einen dynamischen Workflow selbständig zusammenstellen. Dies bezieht sich nicht nur auf die Parametrierung einzelner Komponenten, sondern auch auf die Auswahl der Elemente selbst sowie deren Reihenfolge.
Gewichtung	niedrig

D

VERBUNDSPEZIFISCHE ANFORDERUNGEN

D.1 ZIELE

In diesem Abschnitt werden die spezifischen Ziele des SFB/TRR77 in tabellarischer Form aufgeführt.

EIGENSCHAFT	ERKLÄRUNG
Nummer	Z1
Bezeichnung	Forschungsprojekt durchführen
Beschreibung	Das Hauptziel des SFB/TRR77 ist es, ein tiefes Verständnis der molekularen Basis der Leberkrebsentstehung zu erhalten. Dies umfasst zu Beginn die chronische Lebererkrankung, und führt über die Progression zum metastasierenden Krebs. Weiterhin ist die Identifizierung neuer vorbeugender, diagnostischer und therapeutischer Ansätze Ziel des Verbundes.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	Z2
Bezeichnung	Erzeugen, Speichern und Abrufen von Daten
Beschreibung	Der Verbund hat das Ziel, Daten die mit Leberkrebs assoziiert sind, bereitzuhalten. Diese Daten umfassen beispielsweise Microarraydaten (aCGH, Mythelierung, Expression usw.), Bilddaten (Tissue Microarrays) und klinische Daten.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	Z3
Bezeichnung	Beantworten von Forschungsfragen
Beschreibung	Der Verbund befasst sich beispielsweise mit folgenden Fragen zum Leberkrebs: Welche allgemeinen oder spezifischen Mechanismen der chronischen Leberkrankheiten, insbesondere der chronischen Virusinfektion und der entzündungsvermittelten Prozesse prädisponieren oder initiieren Leberkrebs? Welche molekularen Schlüsselereignisse, die Leberkrebs fördern oder aufrecht erhalten, können als Tumormarker dienen oder stellen aussichtsreiche Angriffsziele für zukünftige therapeutische Interventionen dar?
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	Z4
Bezeichnung	Datenzugriff und -nutzung kontrollieren
Beschreibung	Die Daten des Verbundes müssen so geschützt werden, dass der Zugriff, je nach Datenart nur für bestimmte Mitarbeiter des Verbundes, alle Mitarbeiter des Verbundes oder die Öffentlichkeit möglich ist.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	Z5
Bezeichnung	Daten analysieren
Beschreibung	Die Daten der Projekte des Verbundes müssen für die projektübergreifende Auswertung bereitgestellt werden.
Gewichtung	hoch

D.2 VERBUNDSPEZIFISCHE ANFORDERUNGEN

In diesem Abschnitt werden die konkreten Anforderungen des SFB/TRR77 dargestellt.

EIGENSCHAFT	ERKLÄRUNG
Nummer	A1
Bezeichnung	Daten erzeugen
Beschreibung	Im Forschungsverbund werden projektübergreifende Microarraydaten erzeugt. Diese Daten werden jedoch nicht innerhalb der IT-Plattform selbst erzeugt, sondern lediglich über diese verwaltet. Die einzelnen Projekte erzeugen ihrerseits eigene Daten, die bei Bedarf dem Verbund zur Verfügung gestellt werden. Somit ergibt sich die spezifische Anforderung, die Daten aufzubereiten.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A2
Bezeichnung	Abruf externer Daten
Beschreibung	Zur Annotation der im Verbund erzeugten Daten müssen diese mit öffentlich verfügbaren Daten aus dem Internet zusammengeführt werden.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A3
Bezeichnung	Daten darstellen
Beschreibung	Die Daten des SFB/TRR77 müssen so vorgehalten werden, dass sie leicht auffindbar sind und miteinander verknüpft werden können. Dazu soll es einen zentralen Zugriffspunkt geben.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A4
Bezeichnung	Syntax definieren
Beschreibung	Die Syntax der zentral erzeugten Daten des Verbundes soll in geeigneter Form (beispielsweise in einer XML-WSDL-Datei) dokumentiert werden.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A5
Bezeichnung	Datenmodell definieren
Beschreibung	Die Datenmodelle der zentral erzeugten Daten sollen nach Möglichkeit harmonisiert werden und in geeigneter Form (zu Beispiel UML-XMI) bereitgestellt werden.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A6
Bezeichnung	Daten identifizieren
Beschreibung	Die Datenquellen des SFB/TRR77 werden mit eindeutigen URI versehen und sind so verbundweit abrufbar.
Gewichtung	niedrig

EIGENSCHAFT	ERKLÄRUNG
Nummer	A7
Bezeichnung	Semantik definieren
Beschreibung	Die Semantik der in den für den SFB/TRR77 bereitgestellten Datenquellen enthaltenen Felder muss genau definiert sein, sodass die projektübergreifende Analyse möglich ist.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A8
Bezeichnung	Geistiges Eigentum verwalten
Beschreibung	Die Projekte des Verbundes sollen die Kontrolle über die von ihnen bereitgestellten Daten behalten können. Weiterhin müssen geeignete Mechanismen etabliert werden, um sicherzustellen, dass die Urheberschaft an den Daten auch bei deren Nutzung durch die Verbundpartner berücksichtigt wird.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A9
Bezeichnung	Daten schützen
Beschreibung	Die Daten des Verbundes müssen vor unberechtigtem Zugriff von außen geschützt werden.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A10
Bezeichnung	Ergebnisse anzeigen
Beschreibung	Die Ergebnisse der Analysen des SFB/TRR77 werden an zentraler Stelle angezeigt. Die Darstellung richtet sich nach der Art der darzustellenden Ergebnisse (zum Beispiel Tabelle oder grafische Darstellung).
Gewichtung	noch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A11
Bezeichnung	Daten integrieren
Beschreibung	Die zentral erzeugten Daten des Verbundes werden projektübergreifend ausgewertet und bereitgestellt.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A12
Bezeichnung	Analysemethoden definieren
Beschreibung	Für die Auswertung der Daten werden geeignete biostatistische Methoden ausgewählt und etabliert.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A13
Bezeichnung	Analyseprozess definieren
Beschreibung	Der Prozess zur Analyse von Daten setzt sich aus einer oder mehreren Datenquellen sowie einer oder mehreren Analysemethoden zusammen. Diese Komponenten können aneinandergereiht und verknüpft werden.
Gewichtung	hoch

EIGENSCHAFT	ERKLÄRUNG
Nummer	A14
Bezeichnung	Statischer Workflow
Beschreibung	Prozesse für die Analyse von Daten werden fest eingerichtet. Anwender können bestimmte Parameter, wie beispielsweise das Gensymbol wählen um ein spezifisches Ergebnis zu erhalten.
Gewichtung	mittel

EIGENSCHAFT	ERKLÄRUNG
Nummer	A14
Bezeichnung	Dynamischer Workflow
Beschreibung	Der Anwender kann den vollständigen Analyseprozess selbst zusammenstellen.
Gewichtung	niedrig

E

CELL CULTURE ONTOLOGY

Listing 3: Kopfzeilen von CCONT

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://livercancer.imbi.uni-heidelberg.de/ccont"
     xml:base="http://livercancer.imbi.uni-heidelberg.de/ccont"
     xmlns:protege="http://protege.stanford.edu/plugins/owl/protege#"
     xmlns:CellCultureOntology="http://livercancer.imbi.uni-heidelberg.de
         /ccont#"
     xmlns:xsp="http://www.owl-ontologies.com/2005/08/07/xsp.owl#"
     xmlns:efo="http://www.ebi.ac.uk/efo/"
     xmlns:snap="http://www.ifomis.org/bfo/1.1/snap#"
     xmlns:NCBITaxon="http://purl.org/obo/owl/NCBITaxon#"
     xmlns:obo="http://purl.obolibrary.org/obo/"
     xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
     xmlns:swrl="http://www.w3.org/2003/11/swrl#"
     xmlns:IEV="http://purl.org/obo/owl/IEV#"
     xmlns:owl="http://www.w3.org/2002/07/owl#"
     xmlns:swrlb="http://www.w3.org/2003/11/swrlb#"
     xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
     xmlns:OBO_REL="http://purl.org/obo/owl/OBO_REL#"
     xmlns:oboInOwl="http://www.geneontology.org/formats/oboInOwl#">
    <owl:Ontology rdf:about="http://livercancer.imbi.uni-heidelberg.de/
        ccont">
         <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string
             ">cell culture ontology (CCONT)</rdfs:label>
         <owl:versionInfo rdf:datatype="http://www.w3.org/2001/XMLSchema#
             string">1.1</owl:versionInfo>
         <efo:creator rdf:datatype="http://www.w3.org/2001/XMLSchema#
             string">Matthias Ganzinger, matthias.ganzinger@med.uni-
             heidelberg.de</efo:creator>
         <owl:imports rdf:resource="http://www.ebi.ac.uk/efo/"/>
    </owl:Ontology>
    ...
</rdf:RDF>
```

Tabelle 21: Liste der Relationen aus EFO, die auch in CCONT zur Verfügung stehen. Die erste Spalte enthält den Namen der Relation. Die zweite Spalte enthält die Original-Beschreibung der Relation in englischer Sprache, sofern diese in der veröffentlichten OWL-Datei von EFO enthalten ist.

RELATIONSHIP	DESCRIPTION
bearer_of	A relation between an entity and a dependent continuant; the reciprocal relation of inheres_in [GOC:cjm] example of usage: red eye bearer_of redness
contained_in	
has_quality	
has_role	A relation between a continuant C and a role R. The reciprocal relation of role_of.
contains	
derived_into	
derives_from	Derivation as a relation between instances.
has_part	
has_participant	Has_participant is a primitive instance-level relation between a process, a continuant, and a time at which the A continuant participates in some way in the process.
has_input	
inheres_in	
role_of	
is_executed_in	
is_realized_by	Relation between a realizable entity and a process. Reciprocal relation of realizes
located_in	
location_of	
part_of	
participates_in	Participates_in is a primitive instance-level relation between a continuant and a process in which it participates. For example a scanner participates in a scanning process at some specific time.
is_input_of	
realizes	Relation between a process and a material fulfilling a role (i.e. realizing a role within the context of the process). For example a human realizing role of teacher within a lesson teching process.
regulates	
relationship	

Tabelle 22: In diesem Besipiel wird gezeigt, wie die Zelllinie Hep3B mit Hilfe von CCONT dargestellt wird. Dazu werden die Datenfelder aus den Datenbanken der Zellbanken extrahiert und auf Klassen von CCONT abgebildet. Die hier dargestellte Instanz der Zelllinie entspricht dem Katalog der ATCC-Datenbank.

GRUPPE	IDENTIFIKATION	KLASSENNAME	WERT
identification	EFO_0002205	Hep3B	
origin	EFO_0003150	African American	
	EFO_0000246	age	8
	EFO_0001266	male	
	NCBITaxon_9606	Homo sapiens	
	EFO_0000887	liver	
	EFO_0000182	hepatocellular carcinoma	
properties	CCONT_0000081	adherent	
	CL_0000066	epithelial cell	
	CCONT_0000177	alpha-fetoprotein	
	CCONT_0000178	HBsAg	
	CCONT_0000102	cytogenetics	modal 60
	CCONT_0000100	Amelogenin	X
	CCONT_0000093	CSF1PO	8
	CCONT_0000090	D13S317	12,14
	CCONT_0000092	D16S539	10
	CCONT_0000089	D5S818	13
	CCONT_0000091	D7S820	8,1
	CCONT_0000086	THO1	6,7
	CCONT_0000096	TPOX	9
	CCONT_0000094	vWA	17
	CCONT_0000068	biosafety level 2	
	CCONT_0000179	EBV	negative
	CCONT_0000180	HBV	negative
	CCONT_0000181	HCV	negative
	CCONT_0000182	HIV	negative
	CCONT_0000183	HTLV-I-II	negative
	CCONT_0000184	SMRV	negative
	EFO_0000788	fungal component	negative
propagation	IEV_0000344	MEM(medium)	

Tabelle 22: (Fortsetzung)

GRUPPE	IDENTIFIKATION	KLASSENNAME	WERT
	CCONT_0000048	fetal bovine serum	10
	EFO_0001702	temperature	37.0
	EFO_0000273	atmosphere	95% air
	EFO_0000273	atmosphere	5% co2
	CCONT_0000077	confluence rate	3
	CCONT_0000079	seed density	0.5E6
	CCONT_0000078	split ratio	1:4
	CCONT_0000076	detachment aid	trypsin

LEBENSLAUF

PERSONALIEN

Name:	Matthias Ganzinger
Geburtsjahr:	1974
Geburtsort:	Schweinfurt
Familienstand:	verheiratet, zwei Kinder

SCHULISCHER WERDEGANG

1981–1985	Grundschule Heidenfeld
1985–1994	Walther-Rathenau-Gymnasium Schweinfurt
1. Juli 1994	Allgemeine Hochschulreife

UNIVERSITÄRER WERDEGANG

April 1996	Beginn des Studiums der Medizinischen Informatik an der Universität Heidelberg und Fachhochschule Heilbronn
18. Oktober 2001	Diplomprüfung im Studiengang Medizinische Informatik

BERUFLICHER WERDEGANG

2001–2013	IT-Architekt IBM Deutschland GmbH Heidelberg und Walldorf
seit 2010	Wissenschaftlicher Mitarbeiter Institut für Medizinische Biometrie und Informatik Sektion Medizinische Informatik

DANKSAGUNG

Frau Professorin Dr. Petra Knaup danke ich für die Gelegenheit, diese Arbeit anzufertigen und für ihr Engagement bei der Betreuung der Arbeit.

Weiterhin bedanke ich mich bei den Mitarbeitern des SFB/TRR77 für ihre Geduld und Mithilfe bei der Entwicklung der SOA-Plattform für den Forschungsverbund. Mein besonderer Dank gilt Herrn Priv.-Doz. Dr. Kai Breuhahn, Herrn Priv.-Doz. Dr. Jochen Heß sowie Herrn Priv.-Doz. Dr. Thomas Longerich für ihre ausgezeichneten Ideen zur Umsetzung von PELICAN.