Frank Dickmann

Relevante Metadatenstandards[1]

	Relevante Metadatenstandards
Autor(en)	Frank Dickmann
Editor(en)	Frank Dickmann
Datum	17.11.2011
Version des Dokuments	1.2.1

A: Status des Dokuments

Version 1.2.1

B: Bezug zum Projektplan

Deliverable D3.1 – Relevante Metadatenstandards

C: Abstract

Als Grundlage für das Projekt „Langzeitarchivierung biomedizinischer Forschungsdaten“ (LABIMI/F) wurden mit diesem Dokument relevante Metadatenstandards für zwei Use Cases untersucht: 1. biomedizinische Bilddaten in Magdeburg und 2. Genomdaten in Kiel. Ergänzend wurden in demselben Kontext Aspekte der Langzeitarchivierung bzw. des Forschungsdatenmanagements analysiert. Dazu wurde eine Literaturrecherche zu wissenschaftlicher Literatur ab 2009 durchgeführt. Die Ergebnisse zeigen, dass es einen konkreten Bedarf für eine Langzeitarchivierung auf Basis von Metadaten zur Nachnutzung von Forschungsdaten gibt. Metadaten beschreiben den Inhalt von (Forschungs-)Daten und sind damit praktisch das Rückgrat der Langzeitarchivierung. Mit inhaltsbezogenen Metadaten werden Forschungsdaten nicht nur über lange Zeiträume, sondern auch über Institutionsgrenzen hinweg verständlich.

[1]Dieses Dokument wurde im Rahmen des Projekts LABIMI/F erstellt, das unter dem Förderkennzeichen RI1000/2-1 von der Deutschen Forschungsgemeinschaft (DFG) gefördert wird.

D: Änderungen

Version	Datum	Name	Kurzbeschreibung
0.0.1	14.07.2011	F. Dickmann	Erste Dokumentversion
0.0.2	20.07.2011	F. Dickmann	Ergänzungen zu Metadatenstandards
0.0.3	28.07.2011	F. Dickmann	Ergänzungen zu Metadatenstandards
0.0.4	03.08.2011	F. Dickmann	Kriterien für die Literaturrecherche
0.0.5	10.08.2011	F. Dickmann	Weitere Literatur hinzugefügt
0.0.6	07.09.2011	F. Dickmann	Datenlebenszyklus eingefügt
0.1.0	26.09.2011	F. Dickmann	Ergänzungen
0.2.0	28.09.2011	F. Dickmann	Überarbeitungen, Ergänzungen
0.3.0	30.09.2011	F. Dickmann	Ergänzungen
0.4.0	08.10.2011	F. Dickmann	Ergänzungen
1.0.0	11.10.2011	F. Dickmann	Finalisierung
1.0.1	12.10.2011	F. Dickmann, A. Schneider	Lektorat
1.1.0	04.11.2011	F. Dickmann	Ergänzung zu Analyze
1.2.0	16.11.2011	F. Dickmann	Ergänzung zu ENCODE, RNA-Seq
1.2.1	17.11.2011	F. Dickmann	Ergänzungen zu Diskussion und zum Analyze-Format

E: Inhaltsverzeichnis

1 Einleitung

In der Antike dienten vor allem Bibliotheken als Forschungsumgebung sowie als Plattform des wissenschaftlichen Austauschs [1]. Bis zur Einführung der elektronischen Datenverarbeitung fand wissenschaftlicher Austausch in erster Linie über Publikationen und postalische Kommunikation statt. Obwohl es mithilfe der digitalen Infrastrukturen heute möglich wäre, wissenschaftliche Informationen ubiquitär zu nutzen, ist die medizinische Wissenschaftsgemeinschaft noch weit von einer übergreifenden Nutzung von Forschungsdaten entfernt. Hinzu kommt, dass die digitalen Informationen eine gänzlich andere Aufbewahrungsweise erfordern, als es für papiergebundene Daten üblich ist. An dieser Stelle setzt das Thema der Langzeitarchivierung digitaler Forschungsdaten (im Folgenden „Langzeitarchivierung“) an. Mit Modellen wie dem Open Archival Information System (OAIS) [2] werden mögliche Vorgehensweisen zum Erhalt von Forschungsdaten beschrieben. Grundsätzlich bezieht sich die Langzeitarchivierung damit auf das Forschungsdatenmanagement mit einem weit gesetzten Zeithorizont.

Im Kontext der Forschung zur Langzeitarchivierung nimmt das Thema Metadaten einen wichtigen Stellenwert ein. Generell sind Metadaten essenziell, damit Daten zu einem späteren Zeitpunkt problemlos von anderen Wissenschaftlern interpretiert werden können. Hinzu kommt, dass weltweit – vor allem in der Medizin – immer mehr digitale Daten produziert werden, die sinnvollerweise später wieder interpretiert und in weiterer Forschung genutzt werden sollen [3]. Ein zentrales Problem ist dabei, dass Daten vielfach nicht nach einheitlichen Standards gespeichert werden [4]. Dies trifft nach wie vor auch innerhalb von Instituten zu [5]. Aufgrund der Anforderung an die gute wissenschaftliche Praxis, dass Forschung nachvollziehbar sein soll, ist die Verwendung von Metadaten in der Forschung unerlässlich. Dementsprechend muss sich auch die medizinische Forschung mit dem Thema Metadaten auseinandersetzen [6].

Darüber hinaus wird „data sharing“ eine steigende Bedeutung beigemessen. Zum Beispiel fordern die US National Institutes of Health (NIH) die Veröffentlichung von Daten aus genomweiten Assoziationsstudien in einer dedizierten Datenbank [7]. „Data sharing“ kann für effizientere Forschungsprozesse durch die gemeinsame Nutzung von Daten sorgen, was allerdings nicht per se und trivial belegbar ist [8]. „Data sharing“ kann für die Versorgung weniger stark entwickelter Regionen mit einer besseren Forschungsdatenausstattung vorteilhaft sein, da diese Regionen viele Daten nicht selbst erzeugen können oder ihnen der Zugang fehlt. Mit einer besseren Forschungsdatenausstattung können auch diese Regionen die Ergebnisqualität ihrer Forschung steigern [9].

Das DFG-Projekt LABIMI/F[2] [10] entwickelt daher eine Laborimplementierung zur Langzeitarchivierung von Forschungsdaten für zwei dedizierte biomedizinische Use Cases: 1.) Genomdaten und 2.) Bilddaten. Der vorliegende Bericht ist eine Übersicht zu Metadatenstandards auf Grundlage einer Literaturrecherche für die Langzeitarchivierung von Forschungsdaten in der Medizin mit Fokus Bilddaten und Genomdaten – entsprechend der beiden Use Cases im Projekt LABIMI/F.

[2] **L**angzeit**a**rchivierung **bi**o**m**ed**i**zinischer **F**orschungsdaten.

2 Material und Methoden

In der in der klinischen Forschung ist das Datenmanagement durch eine heterogene Softwarelandschaft geprägt [11]. Für die klinische Forschung ist es zudem notwendig, dass Standards für das GCP[3]-konforme Datenmanagement eingehalten werden [12]. Für die in LABIMI/F eingebundenen Use Cases ist GCP nicht notwendig, da es sich nicht um klinische Forschung handelt:

a.) Der Use Case der Bildverarbeitung dient in erster Linie den Neurowissenschaften und dort Studien mit freiwilligen Probanden.
b.) Der Use Case zu Genomdaten betrifft epidemiologische Studien. In den meisten Fällen sind dies Fall-Kontroll-Studien.

Dementsprechend wurde kein Fokus auf eine mögliche GCP-Konformität der betrachteten Metadatenstandards gelegt.

Zur Literaturrecherche wurden die Literatursuchmaschine Google Scholar [13] und das medizinische Literaturportal PubMed [14] eingesetzt (siehe Tabelle 1). Als Zeitraum für relevante Literatur wurde 01.01.2009 bis 30.06.2011 festgelegt. Damit wurde der aktuelle Stand untersucht. Aufgrund der geringen Anzahl von Suchergebnissen wurde keine weitere Detaillierung im Hinblick auf Genomdaten oder Bilddaten vorgenommen. Die gefundenen Quellen wurden in die Literaturverwaltung Zotero importiert.

1. Die Titel der Quellen wurden auf eine mögliche Eignung für eine weitere Analyse hin bewertet. Von den insgesamt 577 Quellen blieben in diesem Schritt 105 Quellen übrig.
2. Die Abstracts bzw. Buchinhalte wurden auf eine mögliche Eignung für eine vollständige Analyse hin bewertet. Von den 105 Quellen blieben sind 32 Quellen übrig geblieben.
3. Für eine strukturierte Analyse der Inhalte wurden konkret 21 wissenschaftliche Paper und Berichte analysiert. Die weiteren Quellen standen zum Zeitpunkt der Analyse nicht zur Verfügung oder waren nicht mehr direkt abrufbar.
4. Auf Basis von einheitlichen Parametern (siehe Tabelle 2) wurden die verbleibenden 21 Literaturquellen analysiert. Dabei erwiesen sich drei weitere Quellen als nicht relevant; weshalb ergänzend drei weitere Sekundärquellen in die Analyse aufgenommen wurden.

„Data sharing“ wurde vor dem aktuellen Hintergrund der Entwicklungen hin zu „Virtuellen Forschungsumgebungen“ / „Informationsinfrastrukturen für die Forschung“ berücksichtigt [15]. Die Parameter für die Literaturrecherche zu Metadaten umfassen grundsätzliche Bestandteile wie bibliographische Daten, aber auch weitergehende Elemente, die im Kontext des „data sharing“ stehen, z.B. die Interoperabilität von Datenformaten. Darüber hinaus ist ein eindeutiger Identifier (persistent identifier) für ein übergreifendes „data sharing“ notwendig.

[3] Good clinical practice.

Tabelle 1: Übersicht zur Literaturrecherche zu Metadaten in der biomedizinischen Forschung.

Quelle	Verwendete Suchbegriffe	Zeitraum	Trefferzahl
Google Scholar	Metadaten, Medizin, Forschungsdaten	01/2009 – 07/2011	7
Google Scholar	metadata, medicine, research data	01/2009 – 07/2011	568
PubMed	Metadaten, Medizin, Forschungsdaten	01/2009 – 07/2011	0
PubMed	metadata, medicine, research data	01/2009 – 07/2011	2
			577 gesamt

Tabelle 2: Angewendete Parameter der Literaturanalyse zu Metadaten in der biomedizinischen Forschung.

Parameter	Spezifikation
Autor; Titel; Jahr	Allgemeine Quellenparameter
Definition Metadaten	Allgemeine Definitionen zu Metadaten, bzw. was Metadaten sind
Nachhaltigkeitsprobleme; Finanzierung von Nachhaltigkeit	Adressierte Problemstellungen bezüglich Nachhaltigkeit im Rahmen von Langzeitarchivierung und langfristiger Datenerhaltung
Quellenkategorie	Kategorisieren der Quelle in Übersichtsarbeit, Bericht, Fachartikel, Wissenschaftliche Abschlussarbeit, Dissertation
Metadatenbezug	Einordnen in Generisch, Allgemein medizinisch, Bildverarbeitung, Genomforschung
Modell	Referenzmodell, Schema, Ontologie, Datenformat, Taxonomie, Framework, IT-Lösung, Standard, Organisation, Database
Anwendungsbereich	Lokal vs. übergreifend
Inhaltliche Ausrichtung	Grundlagenforschung, Klinische Forschung, Versorgung
Metadatenfokus	Technisch, Fachlich
Verwaltung großer Datenmengen	Petabytebereich und größer
Interoperabilität	Mit anderen Systemen / Formaten
„data sharing“	Gemeinsame (Nach-)Nutzung von Forschungsdaten
Sicherheit/Datenschutz	Unterstützung von Sicherheit / Zugriffsberechtigungen wird gefordert oder ist im Metadatenformat vorhanden
Provenienz	Nutzung von Metadaten für Provenienz / Nachvollziehbarkeit von Forschungsprozessen
Metadatenerzeugung	Verantwortlichkeit der Erzeugung fachlicher Metadaten
Benefit	Benefit für Forschung durch Metadaten / Langzeitarchivierung
Identifier	Eindeutige Identifizierung von Forschungsdaten (z.B. DOI[4])

Da Dublin Core [16] ein generisches Grundgerüst für Metadaten im bibliothekarischen Umfeld darstellt, wurden die Quellen bezüglich des Einsatzes von Dublin Core in der biomedizinischen Forschung hin untersucht. Dies ist sinnvoll, da Dublin Core nahezu beliebig erweiterbar ist und damit auch an die Use Cases in LABIMI/F angepasst werden kann.

[4] Digital object identifier.

Die umfangreich gesetzten Parameter wurden für die eigentliche Auswertung zur Analyse folgender Aspekte aggregiert:

Aspekt 1: Wie wird Langzeitarchivierung definiert?
Aspekt 2: Wie werden Metadaten definiert?
Aspekt 3: Wer ist verantwortlich für die Erzeugung von Metadaten?
Aspekt 4: Was sind relevante Langzeitarchivierungsstandards für die biomedizinische Forschung?
Aspekt 5: Was sind relevante Metadatenstandards für die biomedizinische Forschung?
Aspekt 6: Was sind relevante Datenformate für die biomedizinische Forschung?
Aspekt 7: Welche Vorteile und Bedenken stehen im Zusammenhang mit „data sharing“ für die biomedizinische Forschung?

Die strukturierte Literaturrecherche wurde für Versionen des Deliverables höher als 1.0.1 durch weitere Quellen ergänzt. Dies ist notwendig geworden, da das definierte Literaturspektrum von 2009 bis 2011 z.B. nicht alle relevanten Datenformate abgedeckt hat, die z.T. aus historischen Gründen auch in 2011 noch eine Rolle spielen.

3 Ergebnisse

Aspekt 1: Um ein einheitliches Verständnis von Langzeitarchivierung definieren zu können, wurden die Quellen auf folgende Eigenschaften und Aufgaben untersucht:

- Forschungsdaten können in nachfolgenden Projekten weiter genutzt werden, wodurch doppelte Arbeit in der Forschung vermieden werden kann [17]. Ohne die Nachnutzung von Forschungsdaten können wichtige Forschungsfragen nicht bearbeitet bzw. beantwortet werden [18].
- Aus Forschungsdaten generiertes Wissen als Publikation steht am Ende des Forschungsprozesses. Alljährlich werden große Geldsummen für das Erzeugen bereits existierender Forschungsdaten ausgegeben [19].
- Die Möglichkeit der Nachnutzung ist ein kritischer Bestandteil der Forschungsinfrastruktur [20].
- Auf lange Sicht sollten alle biomedizinischen Forschungsdaten ein großes Ganzes bilden, das maschinell ausgewertet werden kann. Darüber hinaus ist es wichtig, dass Forschungsdaten miteinander verknüpft werden können [21].
- Ein besseres und kosteneffektives Verständnis der Natur[5] wird durch das Dokumentieren und nachhaltige Bereitstellen von Forschungsdaten ermöglicht [22].
- Das Zusammentragen von Forschungsdaten kann sehr schwierig und teuer sein [23].

Insgesamt hat die Langzeitarchivierung biomedizinischer Forschungsdaten ein hohes Potenzial zur Vermeidung von Redundanz. Darüber hinaus wird die Nachnutzung von Forschungsdaten als notwendig für die biomedizinische Forschung eingestuft.

Aspekt 2: Um die Nachnutzung von Forschungsdaten realisieren zu können, sind Beschreibungen bzw. Annotationen notwendig, anhand derer Forschungsdaten wieder aufgefunden werden können. In diesem Zusammenhang werden Metadaten verwendet. Die untersuchten Quellen ordnen Metadaten folgende Eigenschaften zu:

- Die darstellenden Informationen dienen der Wiedergabe und dem Verständnis der Datenobjekte und können syntaktische und semantische Informationen enthalten [17].
- Metadaten beschreiben Forschungsdaten [24]. Durch eine durchgängige Dokumentation dieser Daten bezüglich ihrer Entstehung und Bearbeitung (Provenienz) können Daten übergreifend in weitere Forschungsleistungen eingebunden und mit anderen Daten verknüpft werden [17].
- Ontologien beschreiben formell die Semantik von Daten [4].
- Metadaten werden für jeden Datensatz verwendet, um diesen jeweils zu beschreiben [20].
- Themen im Zusammenhang mit Daten und Metadaten sind Datenbankarchitektur und -verwaltung, Datenmodellierung, Data Mining, Datenauswertung und -analyse, Datenqualität und Datensicherheit [25].

[5] Unter Natur wird in diesem Zusammenhang das naturwissenschaftliche Erkenntnisziel bei der Anwendung von MRT-Untersuchungen und Transkriptomanalyse verstanden.

- In Bezug auf radiologische Metadaten sind Beschreibungen von Bildern und Bildinhalten in einer verständlichen Sprache notwendig. Dies ist auf Basis von Ontologien realisierbar [26].

Aspekt 3: Metadaten können in verschiedenen Stadien des Forschungsprozesses auftreten bzw. müssen in verschiedenen Stadien erfasst werden. Als Grundlage für den Forschungsprozess bietet sich die Beschreibung von Ure et al. für den Lebenszyklus von Forschungsdaten an, da sie technische und fachliche Sichtweise integriert betrachten [4]. Dabei stellt sich heraus, dass Metadaten automatisiert durch Forschungsgeräte bei „Sampling“ und „Collecting“ beigefügt werden können. Diese Metadaten können einerseits Geräteparameter, andererseits aber auch weitere Experiment-spezifische Parameter enthalten. Um dies verlässlich realisieren zu können, muss ein Wissenschaftler vor einem Arbeitsschritt sicherstellen, dass Metadaten korrekt erzeugt werden.

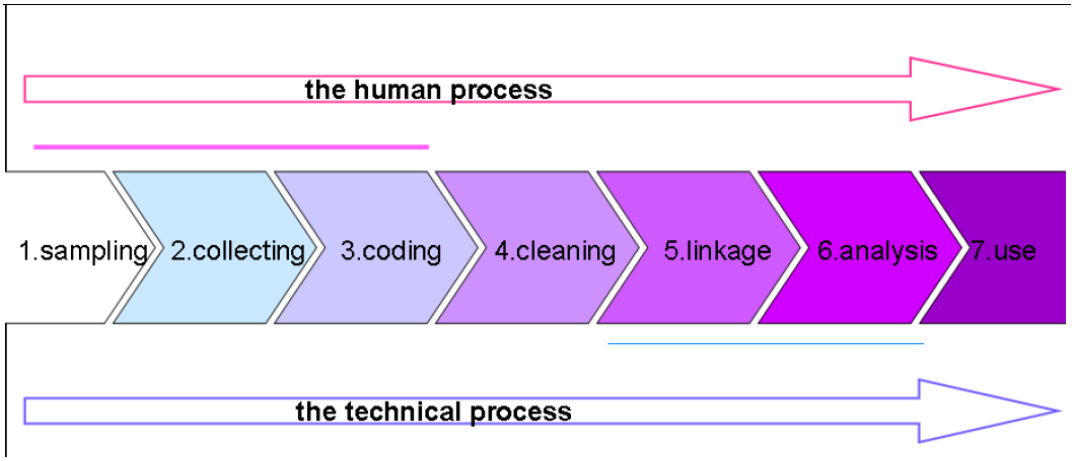

Abbildung 1: Technische und personelle Aufgaben im Rahmen des Datenlebenszyklus [4].

Aspekt 4: Bekannte relevante Langzeitarchivierungsstandards beziehen sich in erster Linie auf das Forschungsdatenmanagement:

- Das Open Archival Information System (OAIS) vertritt eine generische Vorreiterrolle (Referenzmodell) ohne fachspezifische Implementierungs-vorgaben [2,17]. Zentrale Aspekte sind strukturierte Workflows für das Einlagern von Forschungsdaten in ein Archiv, das Einbinden von Metadaten und die Möglichkeit des gezielten Wiederauffindens von und Zugriffs auf archivierte Forschungsdaten. Damit wird unterstrichen, dass mit OAIS kein „dark archive“ gemeint ist, welches nur Daten aufnimmt, auf die ein Zugriff aber nie erfolgt.
- In UK wurde mit dem Integrated Research Application System (IRAS) eine konkret auf geförderte Forschung bezogene Archivierung von Forschungsdaten implementiert. Der Fokus des IRAS-Portals liegt in erster Linie auf der nationalen Beantragung von Forschungsförderung [17].
- Konkrete Lösungen für eine längerfristige Bereitstellung und Nutzung biomedizinischer Bilddaten werden mit Picture Archive and Communications Systems (PACS) realisiert [26]. Implementierungen mit öffentlicher Bereitstellung von biomedizinischen Bilddaten aus den Neurowissenschaften sind Laboratory of Neuroimaging (LONI), Image Data Archive (IDA) und

Human Imaging Database (HID) in Biomedical Informatics Research Network (BIRN) [27].

- Konkrete Lösungen für eine längerfristige Bereitstellung und standortübergreifende Nutzung von Microarraydaten werden angeboten durch das European Molecular Biology Laboratory (EMBL) und ArrayExpress, wobei ArrayExpress konkret Transkriptomdaten adressiert [28]. ArrayExpress wird durch das European Bioinformatics Institute (EBI) angeboten, welches Gründungspartner im EMBL ist. Daneben existiert mit dem Gene Expression Omnibus (GEO) eine weitere internationale Lösung [25].

Aspekt 5: Metadatenstandards in der biomedizinischen Forschung umfassen:

- Mit Dublin Core (DC) existiert ein allgemeiner Standard aus dem Bibliothekswesen. Mit DC werden 15 Metadatenfelder definiert, die für die Erfassung und Nutzung wissenschaftlicher Literatur von allgemeiner Relevanz sind [16,25].
- Persistente und eineindeutige Identifier ermöglichen das widerspruchsfreie Wiederauffinden von Forschungsdaten. Ein wesentliches Metadatum ist daher ein persistenter Identifier (PID) wie z.B. Digital Object Identifier (DOI) [19,20,21,22] oder Life Science Identifier (LSID) [25,29]. Für Webservices und Webressourcen kommen Uniform Resource Identifier (URI) zum Einsatz [28].
- Die Metadata registries (MDR) als ISO/IEC 11179:2003 sind ein internationaler Standard für Aufbau, Struktur und Betrieb eines Metadatenrepositoriums [25,30,31].
- Metadatenstandards für die klinische Forschung sind Clinical Data Interchange Standards Consortium (CDISC), Logical Observations Identifiers, Names, Codes (LOINC) [17], sowie International Classification of Diseases (ICD) / Operationen- und Prozedurenschlüssel (OPS) [32]. Ebenfalls dazu zählt Medical Subject Headings (MeSH) [18,26].
- Für Microarraydaten wird mit Minimum Information About a Microarray Experiment (MIAME) ein Minimaldatensatz zur Annotation definiert; ein weiterer Minimaldatensatz ist der Minimum Information about a Genome Sequence (MIGS) [25]. Daneben hat das Projekt „The Encyclopedia of DNA Elements (ENCODE) Project“ [33] im Juni 2011 einen Standard für RNA-Sequenzierung herausgegeben: Standards, Guidelines and Best Practices for RNA-Seq [34].
- Der generische Standard für Ontologien ist die Web Ontology Language (OWL), die auch für biomedizinische Ontologien verwendet wird [17,28,29,31]. Alternativ zur OWL wird das Resource Description Framework (RDF) verwendet [28,29,35].
- Konkrete biomedizinische Ontologien verwenden den Standard Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE), der u.a. von HealthGrid-Projekten [17] und der OntoNeuroBase [4] verwendet wird. Ebenfalls auf Hirnbilddaten ausgerichtet ist die Ontologie der Reporting Terminology for Brain Arteriovenous Malformations (RadLex) [26]. Der Metathesaurus des US National Cancer Institute (NCI) beinhaltet eine Reihe biomedizinischer Vokabulare und Klassifikationen [31]:
    - CDISC CDASH Terminology, CDISC SDTM Terminology, CDISC SEND Terminology

  - Health Level Seven (HL7) Version 3
  - ICD-9-CM, ICD-10
  - Medical Dictionary for Regulatory Activities (MeDRA) - subset
  - NCI Thesaurus
  - Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT)
- Aufgrund der heterogenen Sprachlandschaft in der biomedizinischen Forschung wurde mit der Unified Medical Language System (UMLS) eine Zusammenführung aus verschiedenen biomedizinischen, kontrollierten Vokabularen und Klassifikationen entwickelt [4,18,35,36,37]. Dabei basiert die UMLS darauf, dass alle eingebundenen externen Standards über einen eindeutigen Identifier (UID) adressiert werden [35]. Zu den von UMLS eingebundenen Vokabularen und Klassifikationen zählen u.a. [18]:
  - ICD-10
  - LOINC
  - MeSH
  - SNOMED CT
  - …

Aspekt 6: Neben Ontologien und Minimaldatensätzen ist es wichtig, dass standardisierte Datenformate für Forschungsdaten etabliert werden [25]. Damit Forschungsdaten zu einem späteren Zeitpunkt sinnvoll verwendet werden können, müssen benutzbare Datenformate verwendet werden [22]. In der biomedizinischen Forschung gibt es folgende Datenformate / Dateiformate:

- Gegenwärtig am weitesten verbreitet sind Daten auf der Basis von XML, sofern es sich um nicht-proprietäre Datenformate handelt [27,28,29,36].
- Häufig werden in der klinischen Forschung Excel oder Comma separated values (CSV) verwendet [17].
- Im Zusammenhang mit Microarraydaten gewinnen einfache, Tabulator-getrennte Datenformate an Bedeutung; ein konkretes Format ist Microarray Gene Expression (MAGE-TAB) [25]. Das ENCODE Project hat zu Datenformaten eine Empfehlung in [34] spezifiziert: zur Dokumentation wird das General Feature Format (GFF) empfohlen [38]. De facto ist FASTQ, als Nachfolger bzw. Erweiterung von FASTA, das Standarddatenformat im Bereich der Genomforschung [39]. FASTQ speichert Sequenzdaten in ASCII-Kodierung als „ACTG“ mit zusätzlichen Informationen zur Qualität der jeweiligen Sequenz [40].
- Für biomedizinische Bilddaten werden in erster Linie die Datenstandards Digital Imaging and Communications in Medicine (DICOM) [18,26,27] und NIFTI [27] verwendet. DICOM ist seit 1993 der Nachfolger von ACR-NEMA[6] [41]. Beide Datenformate können Metadaten aufnehmen bzw. es sind konkrete Metadatenfelder vorgesehen. Unter anderem enthält DICOM einen Unique Identifier (UID) als Teil des Standards. Die beiden Bilddatenformate werden u.a. in der neurowissenschaftlichen Bildgebung (MRI[7], fMRI[8]) verwendet [27]. Ein historisches Datenformat für die neurowissenschaftliche

[6] American College of Radiology (ACR) - National Electrical Manufacturers Association (NEMA).
[7] Magnet resonance imaging.
[8] Functional magnetic resonance imaging.

Bildgebung ist Analyze [42,43], welches ebenfalls Metadaten aufnehmen kann [44].

Aspekt 7: Das gemeinsame Verwenden von Forschungsdaten ist einer der wesentlichen Gründe, warum die Dokumentation von Forschungsdaten mittels Metadaten als sehr wichtig angesehen wird [17]. Dabei muss jedoch geklärt werden, welche Vorteile und Bedenken das gemeinsame Verwenden von Forschungsdaten – das „data sharing“ – für die biomedizinische Forschung hat:

- Viele Forscher verbringen einen großen Teil ihrer Zeit damit, Informationen zu suchen und zu organisieren [23].
- Publizierte Ergebnisse sollten im Allgemeinen gemeinsam mit den zugrundeliegenden Forschungsdaten in öffentlichen Datenbanken bereitgestellt werden [21].
- Aufgrund des Einbindens weiterer Forschungsdaten aus anderen Quellen können publizierte Ergebnisse verifiziert, bessere Metaanalysen durchgeführt und neue Forschungsfragen abgeleitet werden [22]. Die Provenienz von Forschungsdaten kann dies sinnvoll unterstützen [17].
- Mit Grid-Technologie und Methoden zur Forschungsdatenkuration hat das Thema „data sharing“ schnelle Fortschritte erzielt [19].
- Da Forscher ihre Daten aus verschiedenen Gründen ungern teilen (Eigentumsverständnis von Daten [18], Mangel an Vertrauen in externe Dienstleister, andere Forscher könnten bessere Ergebnisse aus den bereitgestellten Forschungsdaten ableiten [45]) wurde von Albani et al. vorgeschlagen, dass Metadaten in erster Linie externen Wissenschaftlern bereitgestellt werden.
- Das Erheben von Metadaten wird von Forschern als zusätzliche Belastung angesehen und wirkt daher bislang einschränkend auf die Möglichkeiten zur gemeinsamen Nutzung von Forschungsdaten [4].
- Unterschiedliche institutionelle Regelungen wirken einer freien, gemeinsamen Nutzung von Forschungsdaten entgegen. Zudem ist es einfacher, Qualitätskriterien für einen eingeschränkten Institutionsrahmen zu spezifizieren als für eine ganze Community [29].

4 Diskussion

Zur Beantwortung der Fragestellungen im Kontext der einzelnen Aspekte wurde ausschließlich die selektierte Literatur verwendet. Daher stellen die Ergebnisse auf Basis aktueller Literatur einerseits einen sehr aktuellen Stand dar, andererseits können sie aufgrund des geringen Umfangs keinen Anspruch auf Vollständigkeit erheben.

Wie durch das RIN diskutiert, betrifft ein großer Teil der wissenschaftlichen Arbeit das Daten- bzw. Informationsmanagement [23]. Dementsprechend kann ein nachhaltiges und strukturiertes Datenmanagement entlastend für Forscher wirken. Unterstrichen wird dies durch die in Abschnitt 3 aufgeführten Aussagen, die unterstreichen, dass ein nachhaltiges Datenmanagement – inklusive „data sharing“ – wissenschaftliche Arbeitsprozesse effizienter gestalten kann.

Daneben sorgt ein verstärktes Erzeugen von Forschungsdaten nicht notwendigerweise für eine signifikante Steigerung von fachlich relevantem Wissen. Die erzeugten Forschungsdaten müssen sinnvoll aufbereitet und mit Metadaten annotiert werden, um sinnvoll in Zusammenhang gebracht werden zu können [21]. Bislang werden diese Metadaten i.d.R. in Laborbüchern durch Wissenschaftler erfasst. Lösungen für elektronische Laborbücher sind bislang in der akademischen Forschung noch wenig verbreitet. Daher entsteht ein Medienbruch, wenn Laborbücher nachträglich digitalisiert werden. Ein integriertes Forschungsdaten-management auf der Grundlage von Erkenntnissen aus der Langzeitarchivierung digitaler Daten wie OAIS kann helfen, Medienbrüche zu vermeiden. Als Konsequenz kann damit die Qualität von Forschungsdaten verbessert werden.

Eine wesentliche Schwierigkeit in der biomedizinischen Forschung ist es, allgemeingültige Metadatenstandards, Vokabulare und Ontologien zu etablieren [4]. Die „one size fits all“-Lösung nach Ure et al. [4] ist daher ein kaum zeitnah realisierbarer Ansatz, sofern nicht alle Beteiligten einer vorgeschlagenen Lösung zustimmen. Sinnvoll ist vielmehr ein generisches Basisprofil, welches durch lokale Anforderungen erweitert werden kann [4]. Hier bietet das generische Dublin Core-Schema eine entsprechende Grundlage für die beiden Use Cases im Projekt LABIMI/F.

PACS-Lösungen sowie HID, LONI und IDA adressieren den Use Case der biomedizinischen Bildverarbeitung. Ob diese Lösungen konkret für den Use Case zutrifft, muss im Weiteren genauer analysiert werden. Anhand der untersuchten Literatur wird jedoch deutlich, dass als Datenformate in erster Linie NIFTI/DICOM, oder Analyze in Frage kommen. Analyze, NIFTI und DICOM können Metadaten aufnehmen.

Für den Use Case zu Genomdaten besteht mit MIAME eine Grundlage für einen Minimaldatensatz an Metadaten. Mit ArrayExpress und GEO gibt es internationale Ansätze, die das „data sharing“ von Transkriptomdaten ermöglichen können. Bei der Veröffentlichung von Transkriptomdaten muss jedoch der Datenschutz gewahrt bleiben. In diesem Kontext muss eruiert werden, ob eine Patienteneinwilligung für Transkriptomdaten generell notwendig ist [8]. Als sinnvolles Datenformat bietet sich nach aktuellem Stand der Technik FASTQ an. Aufgrund der sich kontinuierlich

weiterentwickelnden Technologie im Bereich der Hochdurchsatzsequenzierung kann allerdings davon ausgegangen werden, dass in absehbarer Zeit weitere Formatvarianten entstehen werden. Bezüglich Metadaten sind die Standardvorgaben des ENCORE Projects eine sinnvolle Basis.

„Data sharing“ von Forschungsdaten ist nach wie vor wenig verbreitet. Erste Lösungen wie z.B. mit ArrayExpress und GEO zeigen jedoch, dass „data sharing“ in Zukunft verstärkt an Bedeutung gewinnen wird. Dabei wird vor allem die Güte von Metadaten einen entscheidenden Einfluss auf die Verwendbarkeit öffentlich bereitgestellter Forschungsdaten haben.

Trotz der Bedeutung eines methodisch gut strukturierten und nachhaltigen Forschungsdatenmanagements für nahezu alle Bereiche eines Forschungsprozesses wird Datenmanagement in der Biomedizin noch immer nicht konkret umgesetzt [23]. Ohne ein implementiertes Forschungsdatenmanagement wird die Qualität von Forschung in Zukunft stagnieren. Dies würde die Fortschritte durch den Einsatz von Informationstechnologie ad absurdum führen. Es ist daher notwendig, dass ein professionelles Datenmanagement durch wissenschaftliche Einrichtungen angeboten und von Wissenschaftlern in ihre Forschungsprozesse integriert wird.

5 Ausblick

Für die weitere Entwicklungsleistung im Projekt LABIMI/F wird es notwendig sein, eine ergänzte Version der vorliegenden Untersuchung zu relevanten Metadatenstandards zusammenzustellen. Die ergänzende Undersuchung wird im Rahmen von Workshops in den zwei Use Cases durchgeführt werden. Im Rahmen der beiden Workshops werden Experten aus dem Umfeld der biomedizinischen Bildverarbeitung und der Forschung mit Genomdaten teilnehmen.

Für die ergänzende Untersuchung wird es notwendig sein, die Relevanz der bislang dokumentierten Metadatenstandards deutlicher herauszustellen und mögliche weitere Standards in die Analyse aufzunehmen.

Die Bedeutung eines professionellen Forschungsdatenmanagements ist nicht nur für die beiden Use Cases in LABIMI/F von signifikanter Bedeutung; vielmehr sind alle wissenschaftlichen Bereiche betroffen. Daher ist es eine Aufgabe des Projekts, seine Ergebnisse der gesamten wissenschaftlichen Community zur Verfügung zu stellen.

6 Literaturverzeichnis

[1] Daum, J. (2009): Universitätsbibliotheken gestern und heute. In: Jahrbuch der Raabe-Gesellschaft (1971), Hoppe, K. und Oppermann, H. (Hrsg.), S. 103-117, Walter de Gruyter, Berlin ; New York.

[2] CCSDS (Consultative Committee for Space Data Systems) (2002): Reference Model for an Open Archival Information System (OAIS). Blue Book 1, Washington, DC, CCSDS Secretariat, 2008.08.25, URL: http://public.ccsds.org/publications/archive/650x0b1.pdf.

[3] Woodley, M.S. (2008): Crosswalks, Metadata Harvesting, Federated Searching, Metasearching: Using Metadata to Connect Users and Information. In: Introduction to Metadata, 2. Aufl., Baca, M. (Hrsg.), S. 38-62, The Getty Research Institute, Los Angeles, USA.

[4] Ure, J.; Procter, R.; Lin, Y.-w., et al. (2009): The Development of Data Infrastructures for eHealth: A Socio-Technical Perspective. Journal of the Association for Information Systems, 10 [5], S. 415-429.

[5] Science Staff (2011): Challenges and Opportunities. Science, 331 [6018], S. 692-693.

[6] Linkert, M.; Rueden, C.T.; Allan, C., et al. (2010): Metadata matters: access to image data in the real world. The Journal of Cell Biology, 189 [5], S. 777-782.

[7] National Institutes of Health (NIH) (2007): Policy for Sharing of Data Obtained in NIH Supported or Conducted Genome-Wide Association Studies (GWAS). National Institutes of Health (NIH), Letzter Zugriff: 2011.09.28, URL: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html.

[8] Krawczak, M.; Goebel, J.W. und Cooper, D.N. (2010): Is the NIH policy for sharing GWAS data running the risk of being counterproductive? Investigative Genetics, 1 [3].

[9] Lang, T. (2011): Advancing Global Health Research Through Digital Technology and Sharing Data. Science, 331 [6018], S. 714-717.

[10] Langzeitarchivierung biomedizinischer Forschungsdaten. Universitätsmedizin Göttingen, Göttingen, Letzter Zugriff: 2011.09.20, URL: http://www.labimi-f.de/.

[11] Kuchinke, W.; Ohmann, C.; Yang, Q., et al. (2010): Heterogeneity prevails: the state of clinical trial data management in Europe-results of a survey of ECRIN centres. Trials, 11 [1], S. 79.

[12] Ohmann, C.; Kuchinke, W.; Canham, S., et al. (2011): Standard requirements for GCP-compliant data management in multinational clinical trials. Trials, 12 [1], S. 85.

[13] Google Scholar. Google Inc., Letzter Zugriff: 2011.07.12, URL: http://scholar.google.de.

[14] PubMed. US National Library of Medicine National Institutes of Health, Letzter Zugriff: 2011.06.09, URL: http://www.ncbi.nlm.nih.gov/pubmed/.

[15] Wissenschaftsrat (2011): Übergreifende Empfehlungen zu Informationsinfrastrukturen. Berlin, Wissenschaftsrat, 10466-11, 64, 2011.01.28, Letzter Zugriff: 2011.02.03, URL: http://www.wissenschaftsrat.de/download/archiv/10466-11.pdf.

[16] American National Standards Institute (ANSI) und National Information Standards Organisation (NISO) (2007): The Dublin Core Metadata Element Set. Bethesda, MD, USA, NISO Press, Z39.85-2007, Letzter Zugriff: 2011.09.20, URL: http://www.niso.org/apps/group_public/download.php/6576/The%20Dublin%20Core%20Metadata%20Element%20Set.pdf.

[17] Irshad, T. und Ure, J. (2009): Clinical data from home to health centre: the Telehealth curation lifecycle, Digital Curation Centre SCARP Project Case Studies, University of Edinburgh, 3, 65, (DCC), D.C.C., https://www.era.lib.ed.ac.uk/handle/1842/3370.

[18] Albani, L.; Zg, J.M.V.; Giacomini, M., et al. (2011): Assessment of existing standards, Deliverable, D8.3.1, Cyclic and person-centric Health management (CHIRON), http://www.chiron-project.eu/wp-content/uploads/2011/05/CHIRON-D8_3_1-Assessment-of-existing-standards.pdf.

[19] Brase, J. (2009): DataCite-A global registration agency for research data. In: Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO 2009, Beijing, China.

[20] Thorisson, G.A. (2009): Accreditation and attribution in data sharing. Nature Biotechnology, 27 [11], S. 984-985.

[21] Attwood, T.K.; Kell, D.B.; McDermott, P., et al. (2009): Calling International Rescue: knowledge lost in literature and data landslide! Biochemical Journal, 424 [Pt 3], S. 317-317.

[22] Whitlock, M.C. (2010): Data archiving in ecology and evolution: best practices. Trends in Ecology & Evolution, 26 [2], S. 61-65.

[23] Research Information Network (RIN) und British Library (2009): Patterns of information use and exchange: case studies of researchers in the life sciences, Disciplinary Case Studies in the Life Science Project, London, UK, 56, (RIN), R.I.N., http://www.rin.ac.uk/system/files/attachments/Patterns_information_use-REPORT_Nov09.pdf.

[24] de Carvalho, E.C.A.; Batilana, A.P.; Simkins, J., et al. (2010): Application Description and Policy Model in Collaborative Environment for Sharing of Information on Epidemiological and Clinical Research Data Sets. PLoS ONE, 5 [2], S. e9314.

[25] Mayer, G. (2009): Data management in systems biology I - Overview and bibliography. arXiv.org [arXiv:0908.0411].

[26] Rubin, D.L. und Napel, S. (2010): Imaging Informatics: Toward Capturing and Processing Semantic Information in Radiology Images. In: IMIA Yearbook of Medical Informatics 2010, Kulikowski, C.A. und Geissbuhler, A. (Hrsg.), S. 34-42, Schattauer, Stuttgart.

[27] Adamson, C.L. und Wood, A.G. (2010): DFBIdb: A Software Package for Neuroimaging Data Management. Neuroinformatics, 8 [4], S. 273-284.

[28] Spjuth, O. (2009): Bioclipse: Integration of Data and Software in the Life Sciences. Uppsala University, Disciplinary Domain of Medicine and Pharmacy, Faculty of Pharmacy, Department of Pharmaceutical Biosciences, Uppsala University, Uppsala, Sweden, URL: http://uu.diva-portal.org/smash/record.jsf?pid=diva2:272465.

[29] Sreenivasaiah, P.K. und Kim, D.H. (2010): Current Trends and New Challenges of Databases and Web Applications for Systems Driven Biological Research. Frontiers in Physiology, 1, S. 147.

[30] ISO/IEC (2003): Information technology - Metadata registries (MDR) - Part 3: Registry metamodel and basic attributes. Geneva, Switzerland, International Organization for Standardization, ISO/IEC 11179-3:2003, ISO/IEC 11179-3:2003, 99, 2003.02.15, Letzter Zugriff: 2011.07.19.

[31] Jiang, G.; Solbrig, H.R.; Iberson-Hurst, D., et al. (2010): A Collaborative Framework for Representation and Harmonization of Clinical Study Data Elements Using Semantic MediaWiki. AMIA Summits on Translational Science Proceedings, 2010 [Mar 1], S. 11-15.

[32] Graubner, B. (2007): ICD und OPS - Historische Entwicklung und aktueller Stand. Bundesgesundheitsblatt, 50 [7], S. 932-943.

[33] The ENCODE Project Consortium (2007): Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature, 447 [7146], S. 799-816.

[34] The ENCODE Consortium (2011): Standards, Guidelines and Best Practices for RNA-Seq. Version 1.0, June 2011, Letzter Zugriff: 2011.11.16, URL: http://encodeproject.org/ENCODE/protocols/dataStandards/ENCODE_RNAseq_Standards_V1.0.pdf.

[35] Whetzel, P.L.; Noy, N.F.; Shah, N.H., et al. (2011): BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Research, 39 [suppl 2], S. W541-W541.

[36] Brooks, P. (2010): Standards and Interoperability in Healthcare Information Systems: Current Status, Problems, and Research Issues. In: 5th Midwest

Association for Information Systems Conference MWAIS 2010 Moorhead, MN, USA, URL: http://aisel.aisnet.org/mwais2010/18/.

[37] Sreenivasaiah, P.K. und Kim, D.H. (2010): Current Trends and New Challenges of Databases and Web Applications for Systems Driven Biological Research. Frontiers in Physiology, 1 [0].

[38] Durbin, R. und Haussler, D. (2000): GFF (General Feature Format) specifications document. Hinxton, UK, Wellcome Trust Sanger Institute, Genome Research Limited, Version 2000-9-29, Letzter Zugriff: 2011.11.16, URL: http://www.sanger.ac.uk/resources/software/gff/spec.html.

[39] Cock, P.J.A.; Fields, C.J.; Goto, N., et al. (2010): The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38 [6], S. 1767-1771.

[40] (2008): Maq: Mapping and Assembly with Qualities. Sourceforge, URL: http://maq.sourceforge.net/fastq.shtml.

[41] Graham, R.N.J.; Perriss, R.W. und Scarsbrook, A.F. (2005): DICOM demystified: A review of digital file formats and their use in radiological practice. Clinical Radiology, 60 [11], S. 1133-1140.

[42] Marcus, D.S.; Archie, K.A.; Olsen, T.R., et al. (2007): The Open-Source Neuroimaging Research Enterprise. Journal of Digital Imaging, 20 [1], S. 130-138.

[43] (2001): ANALYZE 7.5 File Format. Mayo Clinic, Letzter Zugriff: 2011.11.17, URL: http://www.mayo.edu/bir/PDF/ANALYZE75.pdf.

[44] Poldrack, R.A.; Mumford, J. und Nichols, T. (2011): Handbook of Functional MRI Data Analysis. Cambridge University Press, Cambridge.

[45] Nelson, B. (2009): Data sharing: Empty archives. Nature, 461 [7261], S. 160-163.